Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Adversarial Robustness for Latent Models: Revisiting the Robust-Standard Accuracies Tradeoff

Adel Javanmard, Mohammad Mehrabi

To cite this article:
Adel Javanmard, Mohammad Mehrabi (2024) Adversarial Robustness for Latent Models: Revisiting the Robust-Standard Accuracies Tradeoff. Operations Research 72(3):1016-1030. https://doi.org/10.1287/opre.2022.0162

https://pubsonline.informs.org/journal/opre

OPERATIONS RESEARCH
Vol. 72, No. 3, May–June 2024, pp. 1016–1030
ISSN 0030-364X (print), ISSN 1526-5463 (online)

Crosscutting Areas

Adversarial Robustness for Latent Models: Revisiting the Robust-Standard Accuracies Tradeoff

Adel Javanmard,[a,*] **Mohammad Mehrabi**[a]

[a] Data Sciences and Operations Department, University of Southern California, Los Angeles, California 90089
*Corresponding author
Contact: ajavanma@usc.edu, https://orcid.org/0000-0003-1934-8747 (AJ); mehrabim@usc.edu (MM)

Received: March 31, 2022
Revised: February 27, 2023
Accepted: June 9, 2023
Published Online in Articles in Advance: August 22, 2023

Area of Review: Machine Learning and Data Science

https://doi.org/10.1287/opre.2022.0162

Abstract. Over the past few years, several adversarial training methods have been proposed to improve the robustness of machine learning models against adversarial perturbations in the input. Despite remarkable progress in this regard, adversarial training is often observed to drop the standard test accuracy. This phenomenon has intrigued the research community to investigate the potential tradeoff between standard accuracy (a.k.a generalization) and robust accuracy (a.k.a robust generalization) as two performance measures. In this paper, we revisit this tradeoff for latent models and argue that this tradeoff is mitigated when the data enjoys a low-dimensional structure. In particular, we consider binary classification under two data generative models, namely Gaussian mixture model and generalized linear model, where the features data lie on a low-dimensional manifold. We develop a theory to show that the low-dimensional manifold structure allows one to obtain models that are nearly optimal with respect to both, the standard accuracy and the robust accuracy measures. We further corroborate our theory with several numerical experiments, including Mixture of Factor Analyzers (MFA) model trained on the MNIST data set.

Funding: A. Javanmard was partially supported by the Sloan Research Fellowship in mathematics, an Adobe Data Science Faculty Research Award, the National Science Foundation Career Award DMS-1844481, and the National Science Foundation Award 2311024.
Supplemental Material: The e-companion is available at https://doi.org/10.1287/opre.2022.0162.

Keywords: adversarial training • robust machine learning • low-dimensional structures • classification

1. Introduction

We are witnessing an unparalleled growth of machine learning tools in various applications domain, where these tools are deployed to inform decisions that directly impact human's lives, from health interventions to credit decisions, sentencing and autonomous driving. Given the safety-critical nature of these applications, reliability and robustness of machine learning systems have become one of the paramount goals of today's AI.

Robust estimation has been one of the central topics in statistics, notably by the seminal work of Tukey (1960), Huber (1992), and Hampel (1968), among others. The majority of work in this area has focused on robustness with respect to outliers (a small fraction of predictors and/or response variables, which are contaminated by gross errors).

Another relevant notion that has spurred a surge of interest in recent years is that of *adversarial robustness*. While machine learning models, and deep learning in particular, have shown remarkable empirical performance, many of these models are known to be highly vulnerable to adversarially chosen perturbations to the input data at test time, known as *adversarial attacks*. Even more surprisingly, many of such adversarial attacks can be designed to be slight modifications of the input which are seemingly innocuous and imperceptible. For example, in image processing and video analysis there are several examples of adversarial attacks in form of indiscernible pixel-wise perturbations which can significantly degrade the performance of the state-of-the-art classifiers (Biggio et al. 2013, Szegedy et al. 2014). Other examples include well-designed malicious contents like malware which can pass the scanning classifiers and yet harm the system, or adversarial attacks on speech recognition systems, such as GoogleNow or Siri, which are incomprehensible or even completely inaudible to human and can still control the virtual assistant software (Vaidya et al. 2015, Carlini et al. 2016, Zhang et al. 2017).

In response to this fragility, a growing body of work in the past few years has sought to improve the robustness of machine learning systems against adversarial attacks. Despite remarkable progress in designing robust training algorithms and certifiable defenses, it is often observed that these methods compromise the statistical

accuracy on unperturbed test data (i.e., test data drawn from the same distribution as training data). Such observation had led prior work to speculate a tradeoff between the two fundamental notions of *robustness* and *generalization* (for a nonexhaustive list see, e.g., Madry et al. 2018, Raghunathan et al. 2019, Min et al. 2020, Mehrabi et al. 2021). For example, the highest obtained ℓ_∞-robust accuracy on CIFAR10 (without using additional data) with $\varepsilon_\infty = 8/255$ is 60%, with standard accuracy of 85% (which is 10% less than state-of-the-art standard accuracy for $\varepsilon_\infty = 0$).

Some of the promising adversarial training methods, such as TRADES (Zhang et al. 2019) acknowledge such tradeoff by including a regularization parameter which allows to tune between these two measures of performance. There has been also recent line of work (Javanmard et al. 2020, Javanmard and Soltanolkotabi 2022) which provides precise asymptotic theory for this tradeoff and how it is quantitively shaped by different components of the learning problem (e.g., adversary's power, geometry of perturbations set, overparameterization, noise level in training data, etc.) For the setting of linear regression and binary classification it is proved that there is an inherent tradeoff between robustness and standard accuracy (generalization) which holds at population level and for any (potentially computationally intensive) training algorithms (Dobriban et al. 2020, Javanmard et al. 2020, Mehrabi et al. 2021). Nonetheless, these work make strong assumptions on the distribution of data (e.g., Gaussian or Gaussian mixture models), which fail to capture various natural structures in data. This stimulates the following tantalizing question:

> *(*) Are there natural data generative models under which the tradeoff between robustness and the standard accuracy (generalization) vanishes, in the sense that one can find models which are performing well (or even optimal) with respect to both measures?*

As a step toward answering this question, Yang et al. (2020) show that when data are well separated, there is no inherent conflict between standard accuracy and robustness. It also provides numerical experiments on a few image datasets to argue that these data are indeed r-well separated for some value r larger than the perturbation radii used in adversarial attacks (i.e., data from different classes are at least r distance apart in the pixel domain). In Xing et al. (2021), adversarially robust estimators are studied for the setup of linear regression and a lower bound on their statistical minimax rate is derived. The minimax rate lower bound for sparse model is much smaller than the one for dense model, whereby Xing et al. (2021) argues the importance of incorporating sparsity structure in improving robustness.

The current work takes another perspective toward question (*) by considering the low-dimensional manifold structures in data. Many high-dimensional real-world data sets enjoy low-dimensional structures, and learning low-dimensional representations of raw data are a common task in information processing. In fact, the entire field of dimensionality reduction and manifold learning has been developed around this task. To give concrete examples, the MNIST database of handwritten digits consists of images of size 28×28 (i.e., ambient dimension of 784), while its intrinsic (manifold) dimension is estimated to be ≈ 14, based on local neighborhoods of data. Likewise, the CIFAR10 database consists of color images of size 32×32 (i.e., ambient dimension of 3,072), but its intrinsic dimension is estimated to be ≈ 35 (Costa and Hero 2004, Rozza et al. 2012, Spigler et al. 2020). The high-level message of the current work is that the low-dimensional structures in data can mitigate the tradeoff between standard accuracy and robustness, and potentially enable training models that perform gracefully (or even optimal) with respect to both measures.

1.1. Summary of Contributions

In this work we focus on two widely used models for binary classification, namely Gaussian-mixture model and the generalized linear model, where we also assume that the feature vectors lie on a k-dimensional manifold in a d-dimensional space ($k < d$). We consider adversarial setting with norm-bounded perturbation (in ℓ_p norm), for general $p \geq 2$.

We use the minimum nonzero singular value of the "lifting matrix" $W \in \mathbb{R}^{d\times k}$ (between the manifold and the ambient space) as a measure of low-dimensional structure of data; cf. (5) and (6). We assess the generalization property of a model through the notion of *standard risk*, and its robustness against adversarial perturbation through the notion of *adversarial risk* (See Section 2 for formal definition.) Our main contributions are summarized as follows:

- Under both data generative models, we derive the Bayes-optimal estimators, which provably attain the minimum standard risk. We prove that as long as $\sigma_{\min}(W)$ diverges as $d \to \infty$, with a growth rate that depends on the adversary's power ε_p and the perturbation norm ℓ_p, then the Bayes-optimal estimator asymptotically achieves the minimum adversarial risk as well. This implies that the tradeoff between robustness and generalization asymptotically disappears as data becomes more structured.
- While the gap between the optimal standard risk and optimal adversarial risk shrinks for data with low-dimensional structure, we show that these two risk measures (as functions of estimators) stay away from each other. Specifically, we come up with an estimator for which the standard and adversarial risks remain away from each other by a constant $c > 0$ independent of k, d.
- In Section 4.1, we consider an adversarial training method based on robust empirical loss minimization.

While this algorithm is structure-agnostic we provably show that it results in models that are robust and also generalize well, under low-dimensional data structure. Note that the data structures (distribution), even if not deployed by the training procedure, still comes into picture as the adversarial risk and standard risk are defined with respect to this data distribution.

- We corroborate our theoretical findings with several synthetic simulations. We also train Mixture of Factor Analyzers (MFA) models on the MNIST image data set. This results in low-rank models from which we can generate new images. Furthermore, the Bayes-optimal classifier can be precisely computed for the MFA model. We show empirically that as the ratio of ambient dimension to the rank diverges (data becomes more structured) the gap between standard risk and adversarial risk vanishes for the Bayes-optimal classifier. In other words, Bayes-optimal estimator becomes optimal with respect to both risks.

1.2. Related Work

There is a growing body of work on the tradeoff between robustness and generalization (see e.g., Madry et al. 2018, Tsipras et al. 2018, Raghunathan et al. 2019, Zhang et al. 2019, Min et al. 2020, Yang et al. 2020, Mehrabi et al. 2021). In particular, Dobriban et al. (2020) consider the isotropic Gaussian-mixture model with two and three classes, and derive Bayes-optimal robust classifiers for ℓ_2 and ℓ_∞ adversaries. This work proves a tradeoff between the standard and the robust risks, which grows when the classes are imbalanced.

The prior work (Jalal et al. 2017, Song et al. 2018, Stutz et al. 2019) proposed the concept of on-manifold attack, where the adversarial perturbations are done in the latent low-dimensional space. In Stutz et al. (2019), it is argued that on-manifold adversarial examples are acting as generalization error and adversarial training against such attacks improve the generalization of the model as well. In addition, a so-called on-manifold adversarial training (based on minimax formulation) has been proposed which is similar to the adversarial training method of Madry et al. (2018) but tailored to perturbations in the manifold space. The subsequent work (Lin et al. 2020) proposes dual manifold adversarial training (DMAT) method which considers adversarial perturbations in both the manifold and the image space to robustify models against a broader class of adversarial attacks. In this terminology, in our current work we consider out-of-manifold perturbations (in the ambient space). Also let us emphasize that Stutz et al. (2019) and Lin et al. (2020) are based on empirical studies on image databases and more on an algorithmic front. The current work contributes to this literature by developing a theory for the role of manifold structure of data in the interplay between robustness and generalization, under specific binary classification setups (viz. Gaussian-mixture model and generalized linear model).

2. Problem Formulation

In this section we discuss the problem setting and formulation of this paper in greater detail. After adopting some notations, we give a brief overview of adversarial setting and describe two data generative models, namely the Gaussian mixture models (GMMs) and generalized linear models (GLMs), which incorporate latent low-dimensional manifold structure. We then conclude this section by a short background on the Bayes-optimal binary classifiers.

2.1. Notations

For a matrix $W \in \mathbb{R}^{d\times k}$, let $\|W\|$ denote its operator norm, $W^\dagger$ stand for the Moore-Penrose inverse, and $\sigma_{\min}(W)$ denote its smallest "nonzero" singular value. For a vector $x \in \mathbb{R}^d$ and $p \geq 1$, we define the ℓ_p norm $\|x\|_{\ell_p} = \left(\sum_{i=1}^d x_i^p\right)^{1/p}$. In addition, let $B_\varepsilon(x)$ denote the ℓ_2-ball centered at x with radius ε. Throughout the paper, for two functions f, g from integers to positive real numbers, we say $f(d) = o_d(g(d))$, as d grows to infinity, if for every $\delta > 0$, we can find a positive integer ℓ such that for $d \geq \ell$, we have $f(d)/g(d) \leq \delta$. In addition, let $\mathsf{N}(\mu, \Sigma)$ denote the probability density of a multivariate normal distribution with mean μ and covariance Σ.

2.2. Adversarial Setting

In the binary classification problem, we are given a set of labeled data points $\{(x_i, y_i)\}_{i=1:n}$ which are drawn i.i.d. from a common law $\mathcal{P}$, where $x \in \mathcal{X} \subset \mathbb{R}^d$ is the feature vector and $y \in \{+1, -1\}$ is the label associated to the feature x. The goal is to predict the label of a new test data point with a feature vector drawn from the similar population. To this end, the learner tries to fit a binary classification model to the training set, which results in an estimated model $\hat{h} : \mathcal{X} \to \{-1, +1\}$. The conventional metric to measure the accuracy of a classifier h is its average error probability on an unseen data point $(x, y) \sim \mathcal{P}$. This is often referred to as the *standard risk* of the classifier, a.k.a. generalization error. Concretely, standard risk of a classifier h is defined as the following:

$$\mathsf{SR}(h) := \mathbb{P}(h(x)y \leq 0). \tag{1}$$

Despite the remarkable success in deriving classifiers with high accuracy (low standard risk) during the past decades, it has been observed that even the state-of-the-art classifiers are vulnerable to minute but adversarially chosen perturbations on test data points.

The adversarial setting is often formulated as a game between the learner and the adversary. Given access to unperturbed training data, the learner fits a model $h : \mathcal{X} \to \{-1, +1\}$. After observing the model h and each

test data point (x, y) generated from the distribution $\mathbb{P}$, the adversary perturbs the data point arbitrarily as far as its within its budget. A common and widely-used adversarial model is that of norm-bounded perturbations, where for each test data point (x,y) the adversary chooses an arbitrary perturbation δ from the ℓ_p ball of radius ε_p and replaces x by $x+\delta$. Here, ε_p is a parameter of the setting which quantifies adversary's power.[1] The *adversarial risk* of the classifier h is defined as the following:

$$\mathsf{AR}(h) = \mathbb{E}_{(x,y)\sim\mathcal{P}}\left(\sup_{\|\delta\|_{\ell_p}\le\varepsilon_p} \ell(h(x+\delta),y)\right), \tag{2}$$

for some loss function ℓ. For the 0-1 loss $\ell(s,t)=\mathbb{I}(st\le 0)$, this measure amounts to

$$\mathsf{AR}(h) = \mathbb{P}\left(\inf_{\|x'-x\|_{\ell_p}\le\varepsilon_p} h(x')y \le 0\right). \tag{3}$$

Remark 1. A couple points are worth noting regarding the adversarial setting:

- The adversary chooses perturbation "*after*" observing the test data point. The perturbation δ can in general depend on x, that is, different data points can be perturbed differently. Therefore, in the definition (2), the supremum is taken inside the expectation.
- In the above setting, the perturbations are added in the test time, while the learner is given access to unperturbed training data. Other adversarial setups are also studied in the literature; see for example, where an attacker can observe and modify all training data samples adversarially so as to maximize the estimation error caused by his attack.
- Another popular adversarial model is the so-called distribution shift. In this model, in contrast to norm bounded perturbations as discussed above, the adversary can shift the test data distribution. The adversary's power is measured in terms of the Wasserstein distance between the test and the training distributions; see Staib and Jegelka (2017), Dong et al. (2020), Pydi and Jog (2020), and Mehrabi et al. (2021) for a nonexhaustive list of references. That said, our focus in this paper is on the norm-bounded perturbations.

From the definition of standard risk and adversarial risk given by (1) and (3), it can be seen that the adversarial risk is always at least as large as the standard risk. We refer to the nonnegative difference of adversarial risk and standard risk as the *boundary risk*, formulated by

$$\begin{aligned}\mathsf{BR}(h) &:= \mathsf{AR}(h) - \mathsf{SR}(h)\\ &= \mathbb{P}\left(h(x)y\ge 0, \inf_{\|x'-x\|_{\ell_p}\le\varepsilon_p} h(x)h(x')\le 0\right).\end{aligned} \tag{4}$$

The boundary risk can be considered as the average vulnerability of the classifier with respect to small perturbations on successfully labeled data points. In other words, it measures the likelihood that the classifier correctly determines the label of a data point, but fails to label another test input very close to the primary data point. In the main result section, we study the boundary risk of optimal classifiers (having the lowest standard risk) in scenarios that features vectors lie on a low-dimensional manifold. We next discuss the data generative models.

2.3. Data Generative Model

2.3.1. Latent Low-Dimensional Manifold Models. We focus on the binary classification problem with high-dimensional features generated from a low-dimensional latent manifold. Specifically, we assume that for the features vector $x\in\mathbb{R}^d$, and the binary label $y\in\{+1,-1\}$, there exists an inherent low-dimensional link $z\in\mathbb{R}^k$ such that $x \perp\!\!\!\perp y \mid z$. This structure can be perceived as a transformed binary classification model, where low-dimensional features $z\in\mathbb{R}^k$ of a hidden classification problem with labels $y\in\{+1,-1\}$, are embedded in a high-dimensional space by a mapping $G:\mathbb{R}^k\to\mathbb{R}^d$. The learner observes the embedded high-dimensional features $x_i=G(z_i)$ and the primary binary labels y_i, while being oblivious to the low-dimensional latent vector z_i.

Throughout the paper, we consider a special case of this model, where $G(z)=\varphi(Wz)$ with $W\in\mathbb{R}^{d\times k}$ a tall full-rank weight matrix, and φ acting entry-wise on vector inputs with a derivative $d\varphi/dt\ge c$, for some positive constant $c>30$.

2.3.2. Classification Settings. The focus of this paper is on two widely used binary classification settings: (i) Gaussian mixture models (ii) generalized linear models which we briefly explain below.

- **Gaussian mixture models.** In the Gaussian mixture model, the binary response value y accepts the positive label with probability π, and the negative label with probability $1-\pi$. In this setting, labels are assigned independently from the feature vector z, while feature vectors are generated from a multivariate Gaussian distribution with the mean vector $y\mu$, and a certain covariance matrix. Concretely, the data generating law for the Gaussian mixture problem with features coming from a low-dimensional manifold can be written as the following:

$$y\sim\mathrm{Bern}(\pi,\{+1,-1\}),\ x=\varphi(Wz),\ z\sim\mathsf{N}(y\mu,I_k). \tag{5}$$

In this model, we consider low-dimensional isotropic Gaussian features. In other words, manifold features z are drawn from a Gaussian distribution with identity covariance matrix.

- **Generalized linear models.** In binary classification under a generalized linear model, there is an

increasing function $f:\mathbb{R}:\to[0,1]$, a.k.a. link function, along with a linear predictor $\beta\in\mathbb{R}^k$, where the score function $f(z^\mathsf{T}\beta)$ denotes the likelihood of feature vector z accepting the positive label. Formally, the data generating law for this classification problem under the low-dimensional manifold model can be formulated as the following:

$$y=\begin{cases}+1 & \text{w.p. } f(z^\mathsf{T}\beta),\\ -1 & \text{w.p. } 1-f(z^\mathsf{T}\beta).\end{cases},\quad x=\varphi(Wz),\quad z\sim\mathsf{N}(0,I_k). \tag{6}$$

Popular choices of the link function f are the logistic model $f(t)=1/(1+\exp(-t))$, and the probit model $f(t)=\Phi(t)$ with $\Phi(t)$ being the standard normal cumulative distribution function.

2.4. Background on Optimal Classifiers

For each classification setup described in the previous section, we want to identify the classifiers that are optimal with respect to the standard risk. To this end, we provide a summary of the Bayes-optimal classifiers. For a data point $(x,y)\sim\mathcal{P}$, consider the conditional distribution function $\eta(x):=\mathbb{P}(y=+1\mid X=x)$. This distribution function can be perceived as the likelihood of assigning the positive label to a data point with feature vector x. The Bayes-optimal classifier simply assigns label $y=+1$ to the feature vector x, if for this feature there is a higher likelihood to accept the label $+1$ than -1. In other words, $h_{\text{Bayes}}(x)=\text{sign}(\eta(x)-1/2)$. We formalize it in the next proposition, which is a standard result (see e.g., Devroye et al. 2013, theorem 2.1).

Proposition 2.1. (Devroye et al. 2013, theorem 2.1) *Among all the classifiers $h:\mathbb{R}^d\to\{+1,-1\}$, such that h is a Borel function, the Bayes-optimal classifier $h_{\text{Bayes}}(x)=\text{sign}(\eta(x)-1/2)$ has the lowest standard risk.*

The next corollary uses Proposition 2.1 to characterize the Bayes-optimal classifier under each of the binary classification settings described earlier in Section 2.3.

Corollary 1. *Under the Gaussian mixture model (5), the Bayes-optimal classifier can be formulated by*

$$h^*(x)=\text{sign}(\varphi^{-1}(x)^\mathsf{T}(WW^\mathsf{T})^\dagger W\mu-q/2),$$

with $q=\log\left(\frac{1-\pi}{\pi}\right)$. Moreover, under the generalized linear model (6), the Bayes-optimal classifier is given by

$$h^*(x)=\text{sign}(f(\beta^\mathsf{T}(W^\mathsf{T}W)^{-1}W^\mathsf{T}\varphi^{-1}(x))-1/2).$$

It is worth noting that in the described manifold latent model of Section 2.3, the weight matrix W is tall and full-rank, and φ is strictly increasing hence both $W^\mathsf{T}W$ and φ are invertible.

3. Main Results

We will focus on the described binary classification settings of Section 2.3. In each setting, we characterize the asymptotic behavior of the associated boundary risk of Bayes-optimal classifiers, when the ambient dimension d grows to infinity. We aim at studying the role of low-dimensional latent structure of data in obtaining a vanishing boundary risk for the Bayes-optimal classifiers. In this case, the Bayes-optimal classifiers are optimal with respect to both measures of standard accuracy and the robust accuracy.

3.1. Gaussian Mixture Model

Consider the Gaussian mixture model with features lying on a low-dimensional manifold, cf. (5). Recall that the learner only observes the ambient $d-$dimensional features x, and is oblivious to the original k-dimensional manifold features z. The next result states that under this setup, the boundary risk of the Bayes-optimal classifier will converge to zero, when the minimum nonzero singular value of the weight matrix W grows at a sufficient rate, which depends on adversary's power ε_p and the choice of perturbations norm ℓ_p.

Theorem 1. *Consider the binary classification problem under the Gaussian mixture model (5) in the presence of an adversary with ℓ_p-norm bounded adversary power ε_p, for $p\geq 2$. By letting the ambient dimension d grow to infinity, under the condition that the weight matrix W satisfies*

$$\frac{\varepsilon_p d^{\frac{1}{2}-\frac{1}{p}}}{\sigma_{\min}(W)}=o_d(1), \tag{7}$$

the boundary risk of the Bayes-optimal classifier converges to zero.

The proof of Theorem 1 is given in the appendix.

We proceed by discussing condition (7). As ε_p gets larger, the condition becomes more strict which is expected; larger value of ε_p indicates a stronger adversary which makes the boundary risk larger. In addition, $\sigma_{\min}(W)$ somewhat measures the extent of low-dimensional structure in data; small $\sigma_{\min}(W)$ indicates that there are directions in the low-dimensional space along which the energy of the signal is not scaled sufficiently large when transformed into the ambient space. Therefore, the adversary can perturb feature x along those dimensions as the existent signal is weak. In particular, if the means of mixture components are close to the space of small eigenvalues of W, then they are collapsed in the embedding from the latent space to the ambient space, making the Bayes-optimal classifier less robust. Finally, since $\|\delta\|_{\ell_p}\leq\|\delta\|_{\ell_2}$ for $p\geq 2$, an adversary with power ε in ℓ_p norm is stronger than an adversary with power ε in ℓ_2 norm. This is consistent with the fact that $d^{\frac{1}{2}-\frac{1}{p}}$ is increasing in p and so the condition becomes stronger for larger p.

Example. Consider the case of $\varphi(\cdot)$ being the identity function and $p=2$. We observe that for feature x with label y, $x_i \sim \mathsf{N}(y w_i^{\mathsf{T}}\mu, \|w_i\|_{\ell_2}^2)$. To be definite, we fix $\|w_i\|_{\ell_2}=1$, which implies in particular $\|W\|_F^2 = d$. To simplify further, we assume that all the nonzero singular values of W to be equal, whence $W^{\mathsf{T}}W = (d/k)I$. In this case, condition (7) reduces to $\varepsilon_2 = o(\sqrt{d/k})$. In particular, if $\varepsilon_2 = O(1)$ and the dimension ratio $d/k \to \infty$ the boundary risk converges to zero.

Figure 1 validates the result of Theorem 1 under the Gaussian mixture model (5) with $\pi = 1/2$, $\mu = \mathsf{N}(0, I_k/k)$, in the presence of an adversary with ℓ_2 bounded adversarial attacks of power ε_2. In this example, we fix the high-dimensional feature dimension $d=300$, and vary the dimensions ratio d/k from 1 to 300. Further, we consider the identical function $\varphi(t)=t$, and let the feature matrix W have independent Gaussian entries $\mathsf{N}(0,1/k)$. Figure 1(a) shows the effect of dimensions ratio d/k on the standard risk, adversarial risk, and the boundary risk of the Bayes-optimal classifier. For each fixed values (k, d), we generate $M=100$ independent realizations and compute the risks. The shaded area around each curve denotes one standard deviation (computed over M realizations) above and below the average curve. As it can be seen, the boundary risk will eventually converge to zero. Finally, in Figure 1(b), we consider several values for adversary's power, where it can be observed that for all adversary's power ε_2, the boundary risk decays to zero as the feature dimensions d/k grows.

The standard risk and the robust risk in Figure 1(a) are calculated using the following proposition.

Proposition 3.1. *Consider the mixture of Gaussian classification setup (5) with the identity mapping $\varphi(t)=t$ and balanced classes (i.e., $\pi = 1/2$). For a linear classifier $h(x) = \mathrm{sign}(x^{\mathsf{T}}a)$ under ℓ_p-bounded adversarial setup, adversarial and standard risks are given by*

$$\mathsf{AR}(h) = \Phi\left(\frac{\varepsilon_p\|a\|_q - a^{\mathsf{T}}W\mu}{\|W^{\mathsf{T}}a\|_2}\right), \quad \mathsf{SR}(h) = \Phi\left(\frac{-a^{\mathsf{T}}W\mu}{\|W^{\mathsf{T}}a\|_2}\right),$$

where $\|.\|_q$ is the dual norm of $\|.\|_p$, that is, $\frac{1}{p}+\frac{1}{q}=1$.

We refer to the appendix for the proof of Proposition 3.1.

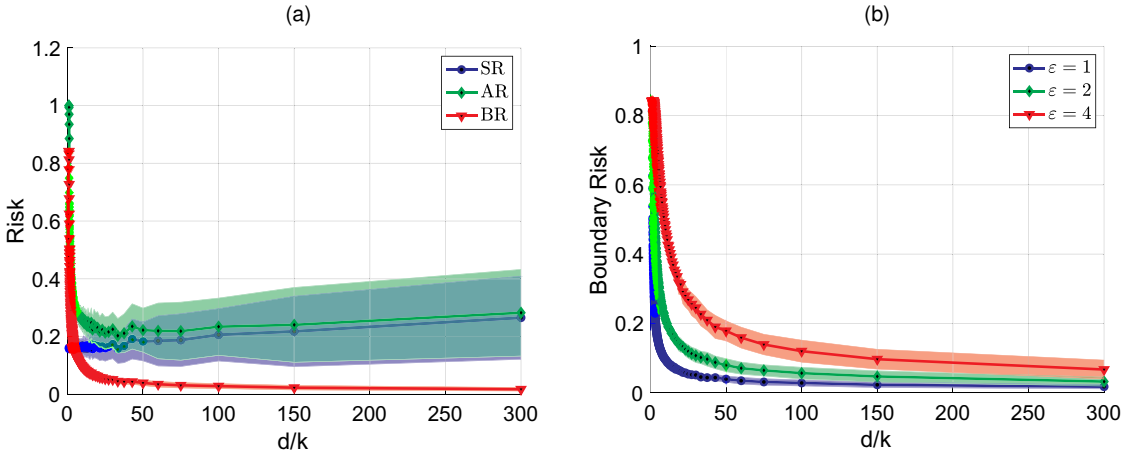
In Theorem 1, we showed that when features lie on a low-dimensional manifold, the Bayes-optimal classifier is also optimal with respect to the adversarial risk. In other words, the adversarial risk is always at least as large as the standard risk, for any classifier, and the gap between the "minimum" of these two risks converges to zero. This result raises the below natural question:

> *"Does the boundary risk of any classifier vanish under the low-dimensional latent structure?"*

In the next proposition, we provide a simple example to show that such behavior (vanishing boundary risk) does not necessarily happen for all classifiers.

Proposition 3.2. *Consider the Gaussian mixture model (5) with μ having i.i.d. $\mathsf{N}(0,1/k)$ entries with class probability $\pi = 1/2$ in the presence of an adversary with bounded ℓ_p perturbations of size ε_p. In addition, suppose that the rows of the feature matrix W are sampled from the k-dimensional unit sphere and consider φ being the identity*

Figure 1. (Color online) Effect of the Dimensions Ratio d/k on the Standard, Adversarial, and the Boundary Risk of the Bayes-Optimal Classifier with ℓ_2 Perturbations Under the Gaussian Mixture Model (5), Where Features Lie on a Low-Dimensional Manifold



Notes. Solid curves represent the average values, and the shaded area around each curve represents one standard deviation above and below the computed average curve over the $M = 100$ realizations. (a) Behavior of the standard and adversarial risks of the Bayes-optimal classifier for the adversary's power $\varepsilon_2 = 1$. (b) Behavior of the boundary risk of the Bayes-optimal classifier for the several values of adversary's power ε_2.

function. Then, the boundary risk of the classifier $h(x) = \text{sign}(e_1^T x)$ *with* $e_1 = (1,0,0,\dots,0)$ *is lower bounded by some constant* c_{ε_p}*, where* c_{ε_p} *depends only on* ε_p *(independent of dimensions k, d), and is strictly positive for positive values of* ε_p*.*

We refer to the appendix for proof of this proposition.

3.2. Binary Classification Under Generalized Linear Models

Consider a binary classification problem under a generalized linear model with features enjoying a low-dimensional latent structure, cf. (6). The next result states that under certain conditions on the weight matrix, the boundary risk of the Bayes-optimal classifier will converge to zero, as the ambient dimension grows to infinity.

Theorem 2. *Consider the binary classification problem under the generalized linear model* (6) *in the presence of an adversary with* ℓ_p*-bounded perturbations of power* ε_p *for some* $p \geq 2$*. Assume that as the ambient dimension d grows to infinity, the weight matrix W satisfies the following condition*:

$$\frac{\varepsilon_p d^{\frac{1}{2}-\frac{1}{p}}}{\sigma_{\min}(W)} = o_d(1). \tag{8}$$

Then the boundary risk of the Bayes-optimal classifier converges to zero.

The proof of Theorem 2 is given in the appendix. Note that the condition (8) is similar to condition (7) for the Gaussian mixture model.

Figure 2 validates the result of Theorem 2 for binary classification under the generalized linear model (6) with identity φ mapping and ℓ_2 perturbations. In this example, the ambient dimension d is fixed at 300, and the manifold dimension k varies from 1 to 300. In addition, the linear predictor β and the weight matrix W have i.i.d. $\mathsf{N}(0, 1/k)$ entries. For each fixed values (k,d), we generate $M = 100$ independent realizations, and we compute the average and the standard deviation of total M obtained values. In each figure, the shaded areas are obtained by moving the average values one standard deviation above and below. Figure 2(a) denotes the behavior of the standard risk, adversarial risk, and the boundary risk of the Bayes-optimal classifier, as the dimensions ratio d/k grows. Further, Figure 2(b) exhibits a similar behavior for several values of adversary's power ε, in which it can be observed that the boundary risk decays to zero.

The standard risk and the robust risk in Figure 2(a) are calculated using the following proposition.

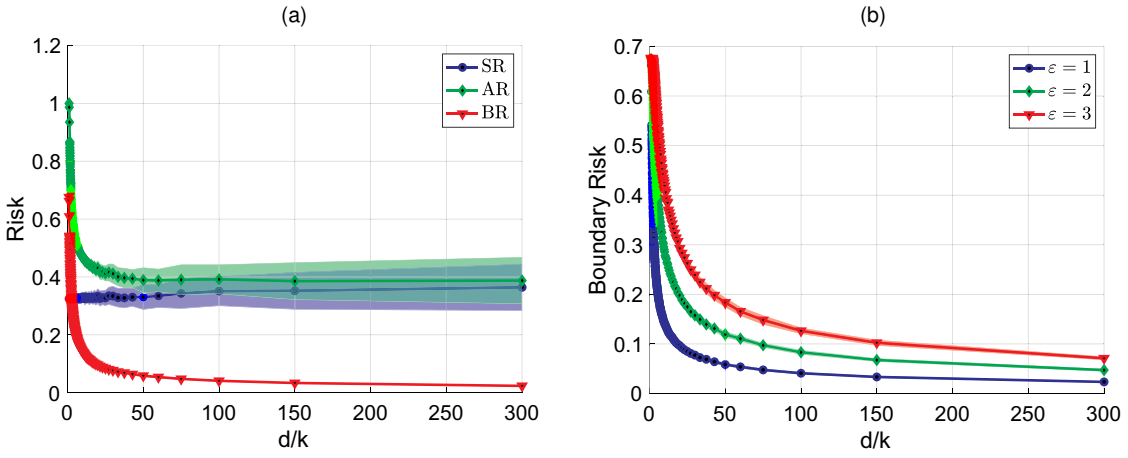
Proposition 3.3. *Consider the generalized linear model* (6) *with the identity mapping* $\varphi(t) = t$*. Under this setup, the adversarial and standard risks of linear classifier* $h(x) = \text{sign}(f(\theta^T x) - 1/2)$ *are given by*

$$\begin{aligned}\text{AR}(h) &= \mathbb{E}_{[u_1,u_2]\sim\mathsf{N}(0,\Sigma_u)}[f(u_1)\mathbb{I}(u_2 \leq c + \varepsilon_p\|\theta\|_q) \\ &\quad + (1 - f(u_1))\mathbb{I}(u_2 \geq c - \varepsilon_p\|\theta\|_q)], \\ \text{SR}(h) &= \mathbb{E}_{[u_1,u_2]\sim\mathsf{N}(0,\Sigma_u)}[f(u_1)\mathbb{I}(u_2 \leq c) \\ &\quad + (1 - f(u_1))\mathbb{I}(u_2 \geq c)],\end{aligned}$$

where the covariance matrix $\Sigma_u = \begin{bmatrix} \|\beta\|_2^2 & \beta^T W^T \theta \\ \beta^T W^T \theta & \|W^T\theta\|_2^2 \end{bmatrix}$, $c = f^{-1}(1/2)$*, and* $\|.\|_q$ *is the dual norm of* $\|.\|_p$*.*

In addition, for the Bayes-optimal classifier $h^*(x) = \text{sign}(f(x^T\theta^*) - 1/2)$ *with* $\theta^* = W(W^TW)^{-1}\beta$ *(see Corollary 1)*

Figure 2. (Color online) Effect of the Dimensions Ratio d/k on the Standard, Adversarial, and the Boundary Risk of the Bayes-Optimal Classifier of the Generalized Linear Model (6) with ℓ_2 Perturbations, in Which Features Are Coming From a Low-Dimensional Manifold



Notes. Solid curves represent the average values, and the shaded areas represent one standard deviation above and below the corresponding curves over $M = 100$ realizations. (a) Behavior of the standard and adversarial risks of the Bayes-optimal classifier for the adversary's power $\varepsilon_2 = 1$. (b) Behavior of the boundary risk of the Bayes-optimal classifier for the several values of adversary's power ε_2.

we have

$$\mathrm{AR}(h^*) = \mathbb{E}_{u\sim\mathsf{N}(0,\|\beta\|_2^2)}[f(u)\mathbb{I}(u \le c + \varepsilon_p\|\theta^*\|_q) + (1-f(u))\mathbb{I}(u \ge c - \varepsilon_p\|\theta^*\|_q)],$$

$$\mathrm{SR}(h^*) = \mathbb{E}_{u\sim\mathsf{N}(0,\|\beta\|_2^2)}[f(u)\mathbb{I}(u \le c) + (1-f(u))\mathbb{I}(u \ge c)].$$

We refer to the appendix for the proof of this proposition.

4. Discussion

4.1. Is it Necessary to Learn the Latent Structure to Obtain a Vanishing Boundary Risk? A Simple Case

In the previous sections, for the two binary classification settings, we showed that when the features inherently have a low-dimensional structure, the boundary risk of the Bayes-optimal classifiers will converge to zero, as the ambient dimension grows to infinity. A closer look at the Bayes-optimal classifier of each setting (can be seen in Corollary 1) reveals the fact that these classifiers directly use the knowledge of the nonlinear mapping from the low-dimensional manifold to the ambient space. In other words, the Bayes-optimal classifiers explicitly draw upon the generative components φ and W. In this section, we investigate the existence of classifiers that are agnostic to the mapping between the low-dimensional and the high-dimensional space, while they have asymptotically vanishing boundary risk. For this purpose, consider binary classification under the Gaussian mixture model (5). In addition, assume a training set $\{(x_i, y_i)\}_{i=1}^n$ sampled from (5). We focus on the class of linear classifiers $h_\theta(x) = \mathrm{sign}(x^\mathsf{T}\theta)$ with $\theta \in \mathbb{R}^d$ and ℓ_2 perturbation ($p = 2$).

We consider the logistic loss $\ell(t) = \log(1+\exp(-t))$, and assume that the adversary's power is bounded by ε. We consider the minimax approach of Madry et al. (2018) to adversarially train a model θ by solving the following robust empirical risk minimization (ERM):

$$\hat{\theta}^\varepsilon = \arg\min_{\theta\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n \max_{u\in B_\varepsilon(x_i)} \ell(y_i u^\mathsf{T}\theta).$$

This is a convex optimization problem, as it can be cast as a point-wise maximization of the convex functions $\ell(y_i u^\mathsf{T}\theta)$. Further, when perturbations are from the ℓ_2 ball, the inner maximization problem can be solved explicitly (see e.g., Javanmard and Soltanolkotabi 2022), which leads to the following equivalent problem:

$$\hat{\theta}^\varepsilon = \arg\min_{\theta\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n \ell(y_i x_i^\mathsf{T}\theta - \varepsilon\|\theta\|_{\ell_2}). \tag{9}$$

Figure 3 demonstrates the effect of the dimensions ratio d/k on the standard, adversarial, and the boundary risk of the classifier $h_{\hat{\theta}^\varepsilon}$ for four different choices of the feature mapping φ: (*i*) $\varphi_1(t) = t$, (*ii*) $\varphi_2(t) = t/4 + \mathrm{sign}(t)\,3t/4$, (*iii*) $\varphi_3(t) = t + \mathrm{sign}(t)t^2$, and (*iv*) $\varphi_4(t) = \tanh(t)$. In this example, we consider the ambient dimension $d = 100$, and number of samples $n = 300$. In addition, k varies from 1 to 100, and μ, W have i.i.d. entries $\mathsf{N}(0, 1/k)$. Further, we consider balanced classes (each label ± 1 occurs with probability $\pi = 1/2$). The plots in Figure 3 exhibit the behavior of the standard, adversarial, and the boundary risks of the classifier $h_{\hat{\theta}^\varepsilon}$, for each of these mappings and for the adversary's power $\varepsilon = 1$. For each fixed value of k, d, we consider $M = 20$ trials of the setup. The solid curve denote the average values over these M trials. The shaded areas are obtained by plotting one standard deviation above and below the main curves. The plots in Figure 4 showcase the boundary risk for different choices of ε. As we observe, the boundary risk decreases to zero, when the dimensions ratio d/k grows to infinity. Our next theorem proves this behavior for the special case of $\varphi(t) = t$.

Theorem 3. *Consider binary classification under the Gaussian mixture model (5) with identity mapping $\varphi(t) = t$ in the presence of an adversary with ℓ_p norm bounded perturbations of size ε_p for some $p \ge 2$. In addition, Let $h_\theta(x) = \mathrm{sign}(x^\mathsf{T}\theta)$ be a linear classifier with $\theta \in \mathbb{R}^d$ and assume that as the ambient dimension d grows to infinity, the following condition on the weight matrix W and the decision parameter θ hold:*

$$\frac{\varepsilon_p d^{\frac{1}{2}-\frac{1}{p}}}{\sigma_{\min}(W)}\left(1 - \|P_{\mathsf{Ker}(W^\mathsf{T})}(\theta)\|_{\ell_2}^2/\|\theta\|_{\ell_2}^2\right)^{-1/2} = o_d(1), \tag{10}$$

where $P_{\mathsf{Ker}(W)}(\theta)$ stands for the ℓ_2-projection of vector θ onto the kernel of the matrix W. Then the boundary risk of the classifier h_θ will converge to zero.

In particular, assume that $\hat{\theta}^\varepsilon$ is the solution of the following adversarial empirical risk minimization (ERM) problem:

$$\hat{\theta}^\varepsilon = \arg\min_{\theta\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n \sup_{u\in B_\varepsilon(x_i)} \ell(y_i u^\mathsf{T}\theta),$$

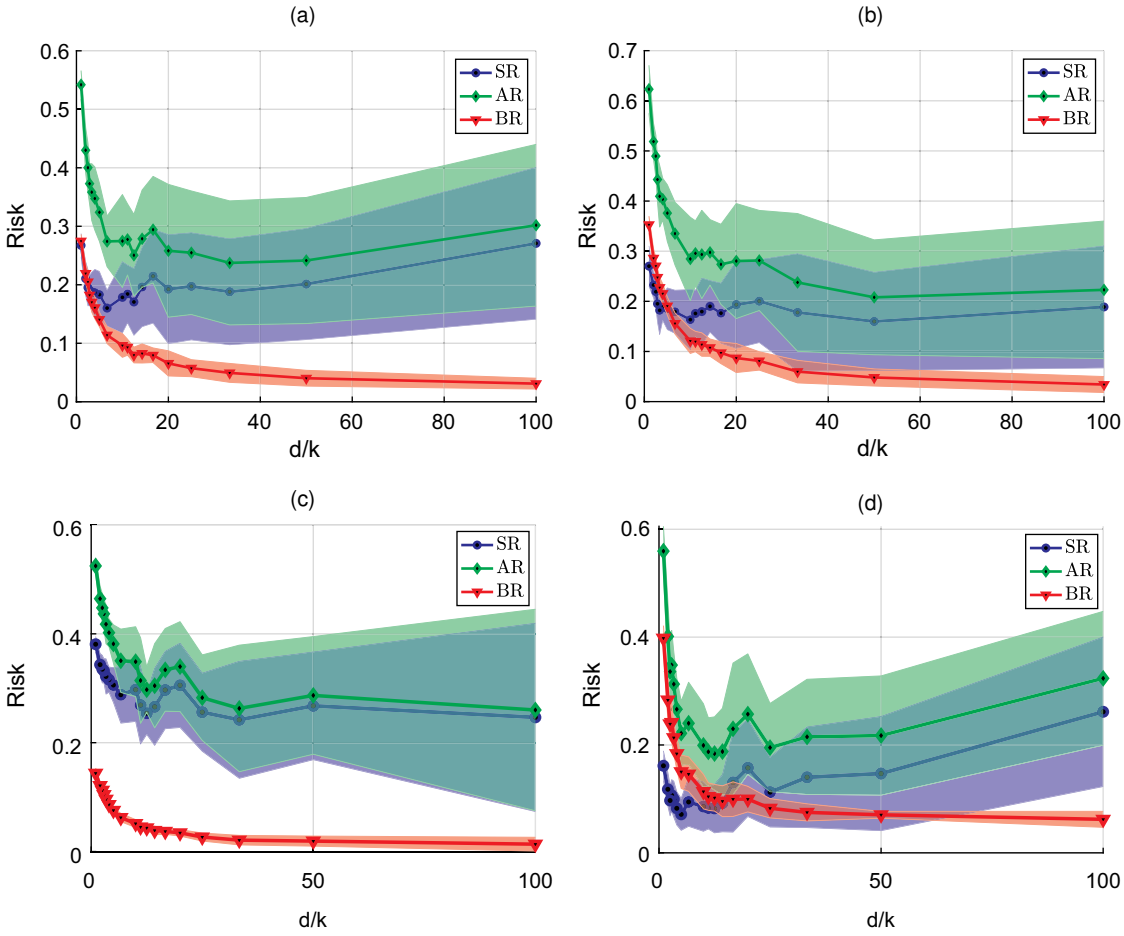
with $\ell : \mathbb{R} \to \mathbb{R}^{\ge 0}$ being a strictly decreasing loss function. In this case, with the weight matrix W satisfying $\frac{\varepsilon_p d^{\frac{1}{2}-\frac{1}{p}}}{\sigma_{\min}(W)} = o_d(1)$, the boundary risk of $h_{\hat{\theta}^\varepsilon}$ converges to zero.

We refer to the appendix for the proof of Theorem 3.

4.2. Effect of Perturbation in the Lifting Matrix *W*

The Bayes-optimal classifiers in previous sections were computed using the true lifting matrix W. In practice, this matrix (or in general the embedding from the latent to ambient space) should be learned and therefore deviates from the true W. Our goal in this section is to show through numerical experiments that our result about the boundary risk being decreasing as d/k grows, remains

Figure 3. (Color online) Effect of Dimensions Ratio d/k on the Standard, Adversarial, and Boundary Risks of the Linear Classifier $h_\theta(x) = \text{sign}(x^\mathsf{T}\theta)$ with θ Being the Robust Empirical Risk Minimizer (9)



Notes. Samples are generated from the Gaussian mixture model (5) with balanced classes ($\pi = 1/2$), and with four choices of feature mapping φ: (a) $\varphi(t) = t$, (b) $\varphi(t) = 3t/4 + \text{sign}(t)t/4$, (c) $\varphi(t) = t + \text{sign}(t)t^2$ and (d) $\varphi(t) = \tanh(t)$. In these experiments, the ambient dimension d is fixed at 100, and the manifold dimension k varies from 1 to 100. The sample size is $n = 300$ the classes average μ and the weight matrix W have i.i.d. entries from $\mathsf{N}(0, 1/k)$. The adversary's power is fixed at $\varepsilon = 1$. For each fixed values of k and d, we consider $M = 20$ trials of the setup. Solid curves represent the average results across these trials, and the shaded areas represent one standard deviation above and below the corresponding curves. (a) Feature mapping $\varphi(t) = t$ and adversary's power $\varepsilon = 1$. (b) Feature mapping $\varphi(t) = t/4 + \text{sign}(t)3t/4$ and adversary's power $\varepsilon = 1$. (c) Feature mapping $\varphi(t) = t + \text{sign}(t)t^2$ and adversary's power $\varepsilon = 1$. (d) Feature mapping $\varphi(t) = \tanh(t)$ and adversary's power $\varepsilon = 1$.

valid even if a perturbed lifting matrix $\hat{W}$ is used in the Bayes-optimal classifier.

Consider the perturbed matrix $\hat{W} = W + N$ where N has i.i.d Gaussian entries $\mathsf{N}(0, \sigma_N^2)$. We start by the Gaussian mixture mode. Using Corollary 1 with mapping $\varphi(t) = t$ and balanced class probabilities ($\pi = 1/2$), the Bayes-optimal classifier reads

$$\hat{h}(x) = \text{sign}\left(x^\mathsf{T}(\hat{W}\hat{W}^\mathsf{T})^\dagger \hat{W}\mu\right). \tag{11}$$

We first characterize the adversarial and standard risk of classifier $\hat{h}$ using Proposition 3.1 and then compute the boundary risk $\mathsf{BR}(\hat{h}) = \mathsf{AR}(\hat{h}) - \mathsf{SR}(\hat{h})$.

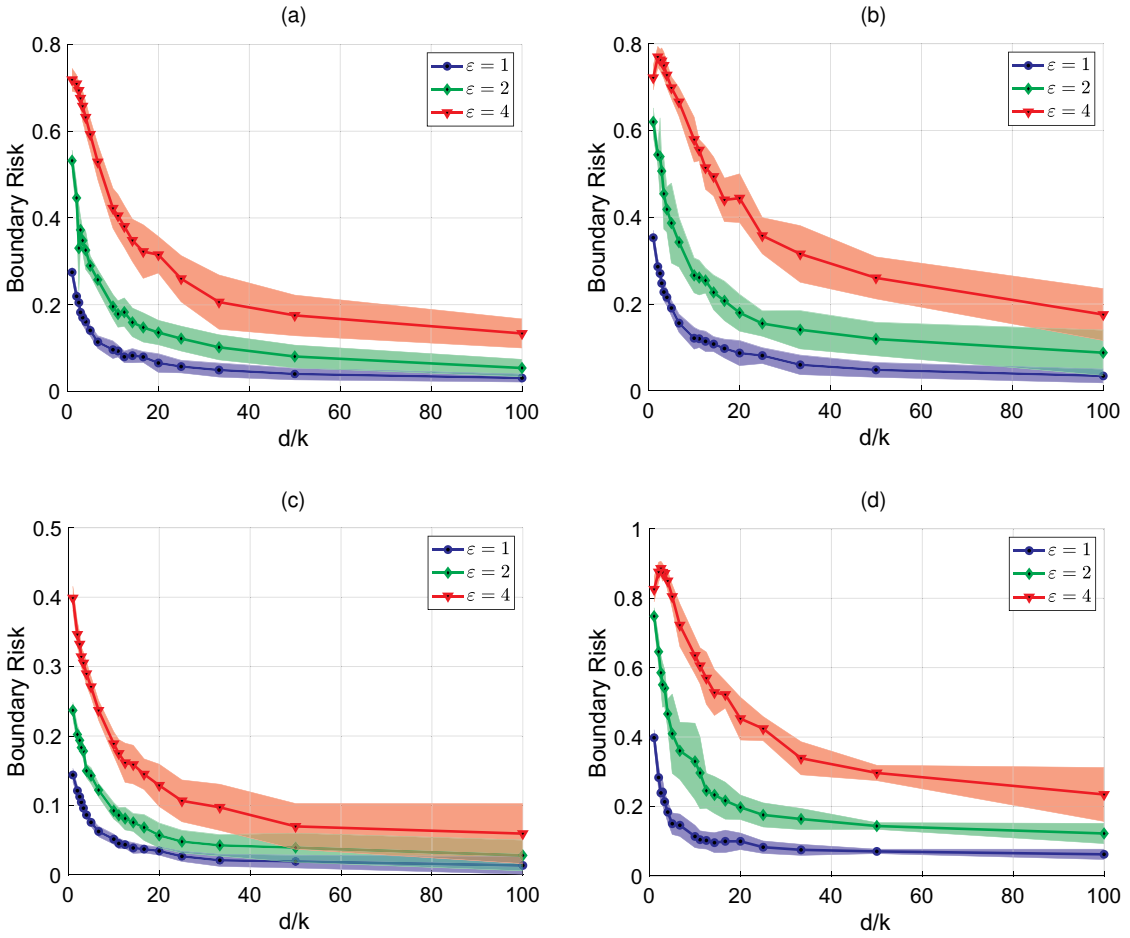
We generate W and μ similar to the previous experiments setup (results in Figure 1) with $d = 300$. In Figure 5(a), we fix $\sigma_N = 1$ and plot the adversarial, standard and boundary risks, averaged over 200 independent instances of the problem. The shaded areas around each curve represent the one standard deviation (computed over 200 experiments) above and below the mean curve. As can be observed, the boundary risk vanishes as the dimension ratio d/k grows to infinity. In Figure 5(b), we plot the boundary risk for $\sigma_N \in \{1, 2, 3, 4, 5\}$. As can be seen, even for large perturbation levels σ_N, the boundary risk converges to zero, albeit at a slower rate.

We next conduct similar experiments for the generalized linear model. In this setting, the Bayes-optimal classifier is given by

$$\hat{h}(x) = \text{sign}\left(f\left(\beta^\mathsf{T}(\hat{W}^\mathsf{T}\hat{W})^{-1}\hat{W}^\mathsf{T}x\right) - 1/2\right). \tag{12}$$

By an application of Proposition 3.3 for $\hat{\theta} = \hat{W}(\hat{W}^\mathsf{T}\hat{W})^{-1}\beta$ we compute the adversarial and the standard risks of $\hat{h}$. For our numerical experiment, we consider the logistic model $f(t) = \frac{1}{1+\exp(-t)}$. In Figure 6(a), we fix $\varepsilon_2 = 1$ (under ℓ_2 adversarial setup), $\sigma_N = 1$ and $d = 300$,

Figure 4. (Color online) Effect of Dimensions Ratio d/k on the Boundary Risk of the Linear Classifier $h_\theta(x) = \text{sign}(x^\mathsf{T}\theta)$ with θ Being the Robust Empirical Risk Minimizer (9)



Notes. Samples are generated from the Gaussian mixture model (5) with balances classes ($p = 1/2$), and with four choices of feature mapping φ: (a) $\varphi(t) = t$, (b) $\varphi(t) = 3t/4 + \text{sign}(t)t/4$, (c) $\varphi(t) = t + \text{sign}(t)t^2$ and (d) $\varphi(t) = \tanh(t)$. In these experiments, the ambient dimension d is fixed at 100, and the manifold dimension k varies from 1 to 100. The sample size is $n = 300$ the classes average μ and the weight matrix W have i.i.d. entries from $\mathsf{N}(0, 1/k)$. We consider different levels of the adversary's power $\varepsilon \in \{1, 2, 4\}$. For each fixed values of k and d, we consider $M = 20$ trials of the setup. Solid curves represent the average results across these trials, and the shaded areas represent one standard deviation above and below the corresponding curves. (a) Boundary risk with the feature mapping $\varphi(t) = t$ and for multiple values of adversary's power ε. (b) Boundary risk with the feature mapping $\varphi(t) = t/4 + \text{sign}(t)3t/4$ and for multiple values of adversary's power ε. (c) Boundary risk with the feature mapping $\varphi(t) = t + \text{sign}(t)t^2$ and for multiple values of adversary's power ε. (d) Boundary risk with the feature mapping $\varphi(t) = \tanh(t)$ and for multiple values of adversary's power ε.

and vary k. The reported results are averaged over 50 independent experiments with shaded curves showing one standard deviation above and below the averaged values (computed over 50 realizations). As can be seen, even though the Bayes-optimal classifier is computed using the perturbed matrix $\hat{W}$, the boundary risk converges to zero as d/k grows to infinity. In Figure 6(b), we repeat similar experiments for different values of $\sigma_N \in \{1, 2, 3, 4, 5\}$. As we see, even for large perturbation levels σ_N, the boundary risk converges to zero, albeit at a slower rate.

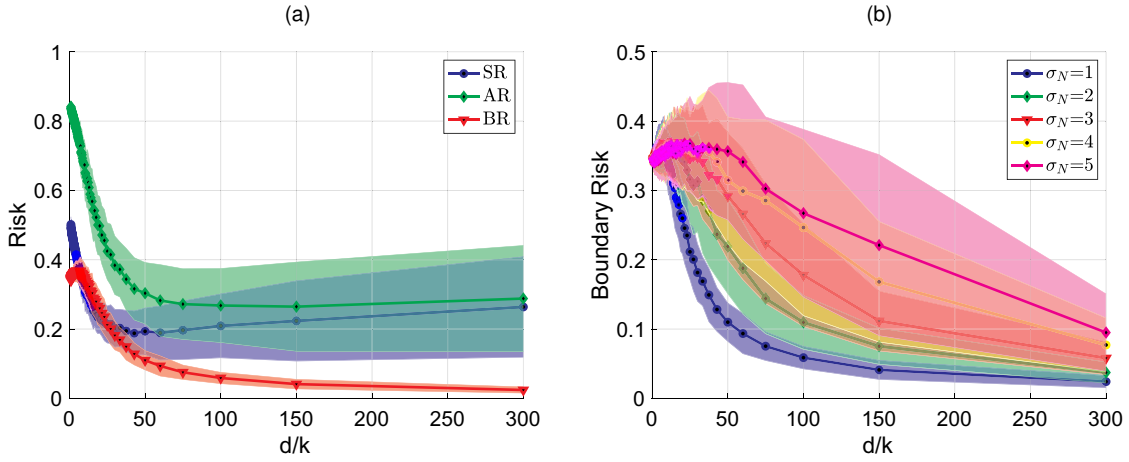
The above experiments demonstrate that for the Bayes-optimal classifier, constructed using a perturbed lifting matrix $\hat{W}$, the robust-standard risks tradeoff is attenuated as the low-dimensional structure of data becomes stronger.

5. Boundary Risk of Bayes-Optimal Image Classifiers

We next provide several numerical experiments on the MNIST image data to corroborate our findings regarding the role of low-dimensional structure of data on the boundary risk of Bayes-optimal classifiers. Of course, the evaluation of this finding on image data are challenging since learning particular structure of the underlying image distribution is notoriously a difficult problem. There have been a few well-established techniques for this task that we briefly discuss below.

Generative Adversarial Net (GAN) (Goodfellow et al. 2014) is among the most successful methods in modeling the statistical structure of image data. Despite the remarkable success of GANs in generating realistic high

Figure 5. (Color online) Effect of the Dimensions Ratio d/k on the Standard, Adversarial, and the Boundary Risk of the Bayes-Optimal Classifier Obtained From a Noisy Weight Matrix $\hat{W}$ with ℓ_2 Perturbations Under the Gaussian Mixture Model (5), Where Features Lie on a Low-Dimensional Manifold

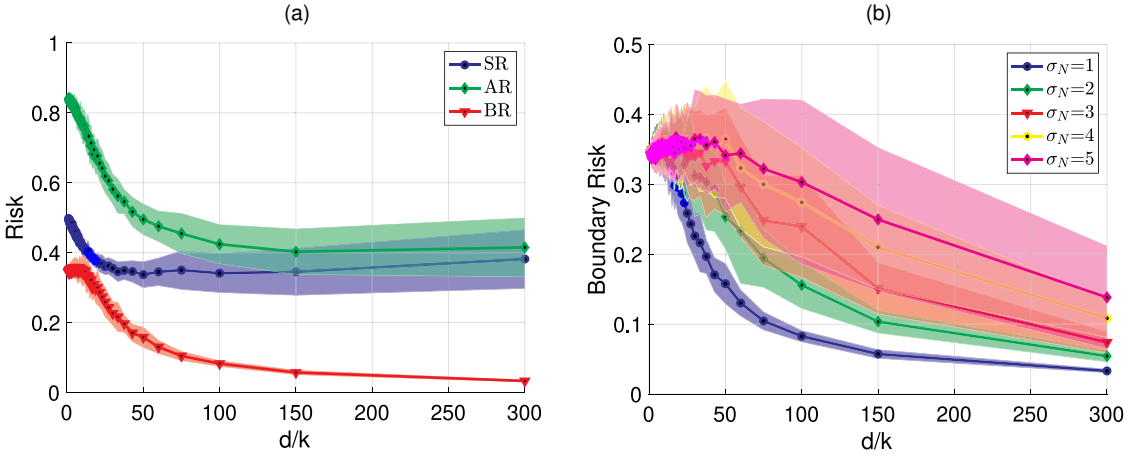


Notes. Solid curves represent the average values, and the shaded area around each curve represents one standard deviation above and below the computed average curve over the $M = 200$ realizations. (a) Behavior of the standard and adversarial risks of the Bayes-optimal classifier (11) computed with noisy weight matrix $\hat{W} = W + N$ with N having i.i.d entries $\mathsf{N}(0,1)$ for the adversary's power $\varepsilon_2 = 1$. (b) Behavior of the boundary risk of Bayes-optimal classifier (11) computed with noisy weight matrix $\hat{W} = W + N$ with N having i.i.d entries $\mathsf{N}(0,\sigma_N^2)$ for the adversary's power $\varepsilon_2 = 1$ and $\sigma_N \in \{1,2,3,4,5\}$.

resolution images, it has been observed that they may fail in capturing the full data distribution, which is referred to as *model collapse*. In addition, computing the likelihood of image data with GANs requires to perform complex computations. As a direct implication of these observations, it is not statistically accurate and efficient to deploy GANs to formulate the Bayes optimal classifiers (Richardson and Weiss 2018, 2021).

Fitting elementary statistical models can mitigate the statistical inaccuracies of GANs. (Richardson and Weiss 2018) learns the statistical structure of image data by using the class of *Gaussian Mixture Models* (GMM). This choice is motivated by the statistical power of GMMs that they are universal approximators of probability densities (Goodfellow et al. 2016). On the other hand, working with general Gaussian covariance matrices can

Figure 6. (Color online) Effect of the Dimensions Ratio d/k on the Standard, Adversarial, and the Boundary Risk of the Bayes-Optimal Classifier Obtained From a Noisy Weight Matrix $\hat{W}$ with ℓ_2-Perturbations Under the Generalized Linear Model (6)



Notes. Solid curves represent the average values, and the shaded area around each curve represents one standard deviation above and below the computed average curve over the $M = 50$ realizations. (a) Behavior of the standard and adversarial risks of the Bayes-optimal classifier (12) under generalized linear model (6) computed with noisy weight matrix $\hat{W} = W + N$ with N having i.i.d entries $\mathsf{N}(0,1)$ for the adversary's power $\varepsilon_2 = 1$. (b) Behavior of the boundary risk of Bayes-optimal classifier (12) under generalized linear model (6) computed with noisy weight matrix $\hat{W} = W + N$ with N having i.i.d entries $\mathsf{N}(0,\sigma_N^2)$ for the adversary's power $\varepsilon_2 = 1$ and $\sigma_N \in \{1,2,3,4,5\}$.

make the estimation problem both in terms of computational cost and memory storage extremely prohibitive. (Richardson and Weiss 2018) deployed *Mixture of Factor Analyzers (MFA)* (Ghahramani and Hinton 1996) to avoid storage and computation with such high-dimensional matrices. This deployment is aligned with the former intuition that the space of meaningful images is indeed a small portion of the entire high-dimensional space. In addition, they show that with moderate number of components in the GMM one can produce adequately realistic images, and further reduce the computational burden.

We will adopt the MFA procedure introduced in (Richardson and Weiss 2018) to generate realistic image data for our numerical experiments. The main reasons for this adoption are: (i) this framework is flexible for generating realistic images from a low-dimensional subspace, and (ii) it enables us to accurately and efficiently calculate the log-likelihood of images, which can be used later to formulate the Bayes-optimal classifier. It is worth noting that, using a class of less complex models, in this case GMMs rather than GANs, will output images with lower resolution, which is not a major concern for the main purpose of this numerical study. In the next sections, we first provide a brief overview of the GMM estimation steps and then review some of the standard frameworks to produce adversarially crafted examples.

5.1. Learning Image Data with Gaussian Mixture Models (GMM)

A general setup for fitting a GMM to image data $\{x_i\}_{1:n} \in \mathbb{R}^d$ is based on the following model

$$x \sim \sum_{k=1}^{K} \alpha_k \mathsf{N}(\mu_k, \Sigma_k), \quad \Sigma_k \in \mathbb{R}^{d\times d}, \ \ \mu_k \in \mathbb{R}^d,$$

where K denotes the number of components in GMM and α_i are mixing weights. This problem, without imposing any extra structure on image data, involves learning $O(Kd^2)$ parameters which can be extremely difficult for high-dimensional images. (Richardson and Weiss 2018) deployed a mixture of factor analyzers (MFA), where they use tall matrices $A_{d\times\ell}$ to embed a low-dimensional subspace in the full data space. In this case, the following model is considered

$$x \sim \sum_{k=1}^{K} \alpha_k \mathsf{N}(\mu_k, A_k A_k^{\mathsf{T}} + D_k),$$
$$A_k \in \mathbb{R}^{d\times\ell}, \ \ D_k \in \mathbb{R}^{d\times d}, \ \ \mu_k \in \mathbb{R}^d,$$

where D_k is a diagonal matrix showing the variance on each single pixel. This model ameliorates the previous high storage and computational cost, as in this case $Kd(\ell+2)$ learning parameters exist, which scales linearly with image dimension d. This model is intimately related to the specific case of the low-dimensional manifold models on features in (5) and (6) with the identity mapping $\varphi(t) = t$. The only difference is in the entrywise independent Gaussian noise coming from diagonal matrix D_k, however in the numerical experiments we observed that indeed the estimated values of D_k entries are extremely small, which makes this difference negligible. We use the maximum likelihood estimator to compute the model parameters. For this purpose, we need to compute the log-probabilities. For a single component of GMM we first compute the likelihood $p(x \mid A, D, \mu)$ given by

$$p(x \mid A, D, \mu) = -\frac{1}{2}[d\log(2\pi) + \log\det(AA^{\mathsf{T}} + D) + (x-\mu)^{\mathsf{T}}(D + AA^{\mathsf{T}})^{-1}(x-\mu)]. \tag{13}$$

Richardson and Weiss (2018) used the following algebraic identities with $u = x - \mu$, $L = A^{\mathsf{T}}DA \in \mathbb{R}^{\ell\times\ell}$ to avoid large matrix storage and multiplications:

$$u^{\mathsf{T}}\Sigma^{-1}u = u^{\mathsf{T}}[D^{-1}u - D^{-1}AL^{-1}(A^{\mathsf{T}}D^{-1}u)],$$
$$\log\det(AA^{\mathsf{T}} + D) = \log\det L + \sum_{j=1}^{d}\log(d_j),$$

where d_i denotes the i-th entry on the diagonal of matrix D. In addition, Richardson and Weiss (2018) employed *differentiable programming* framework to efficiently solve the corresponding Maximum-likelihood optimization problem on GPU. We use their publicly available code at https://github.com/eitanrich/gans-n-gmms to fit GMMs on full image data sets. As a simple example, we consider models with $K = 10$ components and the manifold dimensions $\ell = 1, 10, 100$. We fit these three models to the training samples of the MNIST data set (Deng 2012) that are labeled "6". Figure 7 exhibits sample images generated from the learned GMM models.

5.2. PGD and FGM Adversarial Attacks

Recall that adversarial examples are meant to be close enough to original samples, yet be able to degrade the classifier performance. For a loss function $\mathcal{L}$ consider the adversarial optimization problem

$$\max_{\|x-x'\|_{\ell_2}\le\varepsilon} \mathcal{L}(h(x'), y).$$

Using the 0-1 loss $\mathcal{L}(h(x'), y) = \mathbb{I}(y \neq h(x'))$ yields the inner optimization problem (3). A large body of proposed methods to produce adversarial examples consider the first-order linear approximation of the loss function around the original sample. More precisely, x' is written as $x + \delta$, and then single/multi steps of gradient descent (GD) of the negative loss function is considered. In this framework, first a powerful predictive model, for example, a neural network, is fit to the training samples, which will be used as a surrogate for the learner's model h. For ℓ_2-bounded adversary's budget ε,

Figure 7. Sample Generated Images From a GMM Model Fit on MNIST Training Set Images with Label Six

Note. Three GMM models with number of components $K = 10$, and manifold dimensions $\ell = 1, 10, 100$ (from left to right) are considered.

the *Fast Gradient Method* (FGM) performs a single step of normalized GD which yields

$$x' = x + \varepsilon \frac{\nabla_x \mathcal{L}(h(x), y)}{\|\nabla_x \mathcal{L}(h(x), y)\|_{\ell_2}}.$$

This method is first introduced in Goodfellow et al. (2015) for ℓ_∞-bounded attacks under the name *Fast Gradient Sign Method (FGSM)*. Other variants for other ℓ_p-bounded adversarial attacks are introduced in Tramèr et al. (2017), generally called the FGM (removing "sign" from FGSM). A more general scheme to produce adversarial examples is via a multistep implementation of the above procedure with the projected gradient descent (PGD). This attack is introduced in Madry et al. (2018), with iterative updates given by

$$x^{t+1} = \Pi_{B_\varepsilon(x)}\left(x^t + \varepsilon \frac{\nabla_x \mathcal{L}(h(x^t), y)}{\|\nabla_x \mathcal{L}(h(x^t), y)\|_{\ell_2}}\right),$$

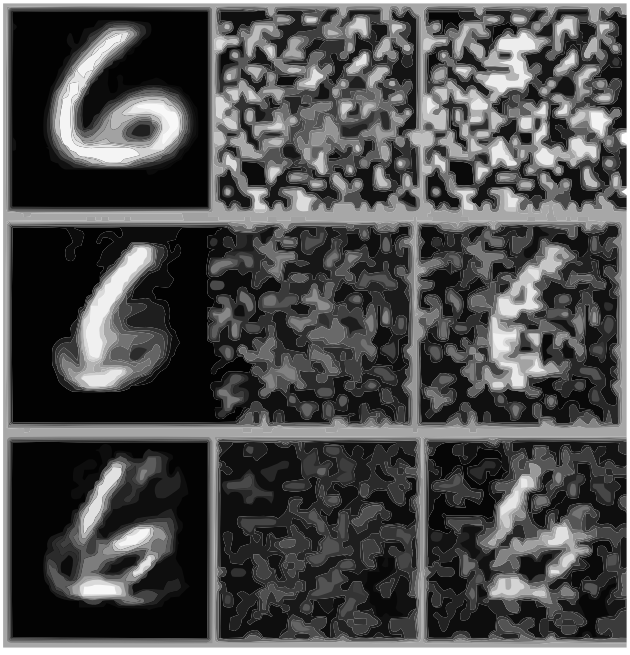
where $\Pi_{B_\varepsilon(x)}$ stands for the projection operator to the ℓ_2-ball centered at x with radius ε. For our image classification numerical experiments, we will use FGM and PGD attacks to produce adversarial examples. We follow the same implementation of PGD and FGM adversarial attacks provided in CleverHans library v4.0.0 (Papernot et al. 2016). The code for this implementation can be accessed at https://github.com/cleverhans-lab/cleverhans. In our implementation, the original image values are normalized to be in the interval $[0, 1]$, and we clip perturbed pixel values to be in the same interval. We next present key findings from our numerical experiments.

5.3. Main Experiments and Key Findings

In this section, we connect the previously described parts. Put all together, these are the three main steps of our experiments:

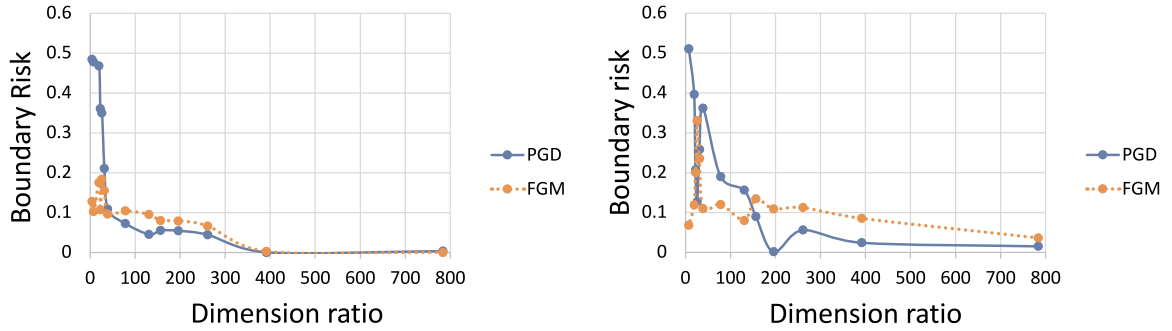
1. For several choices of K (number of components) and latent dimensions ℓ, we first fit two GMM models to zeros and sixes of the training set of MNIST data set. By deploying the learned models, we generate 5,000 new images with uniform probability on labels 6 or 0, that is, at the beginning of generating an image, with equal chance we decide to use either of the models. In addition, for the defined binary classification problem (0 versus 6), we deploy (13) to obtain two likelihood

Figure 8. Attacks and Perturbed Images for Three Different Latent Dimensions $\ell = 1, 10, 100$, Respectively From Top to Bottom



Notes. In each row, from left to right, the original sample, the adversarially crafted perturbation, and the perturbed image is exhibited. The original images are generated from a GMM with the number of components $K = 10$. In all images under the PGD adversarial perturbation, we start with the ℓ_2-bounded adversarial power $\varepsilon = 0$, and incrementally increase it to the point that the Bayes-optimal classifier fails to infer the correct label. In this experiment, the Bayes-optimal misclassification on sampled images with $\ell = 1, 10, 100$ occurs at ℓ_2-bounded adversarial power $\varepsilon = 22.6, 11, 7$, respectively. It can be seen that samples coming from smaller latent dimension ℓ require stronger perturbation to get misclassified by the Bayes-optimal classifier.

Figure 9. (Color online) Boundary Risk of the Bayes-Optimal Classifier on 1,000 Test Images Generated From GMM Models with Number of Components $K = 1$ and $K = 10$ (From Left to Right)



Notes. In each experiment, the adversary's ℓ_2-bounded perturbation power is fixed at $\varepsilon = 12$, and both adversarial attacks PGD and FGM are considered. It can be observed that the boundary risk of the Bayes-optimal classifier will converge to zero as the dimension ratio d/ℓ grows.

models $p_0(x), p_1(x)$. This can be used to formulate the Bayes-optimal classifier $\mathbb{I}(p_1(x) > p_0(x))$.

2. In this step, we adversarially attack the generated images. To this end, the data set is split into 80%–20% training-test samples. The training set is used to train a neural network for PGD and FGM attacks. The obtained model will be used later to craft adversarial examples for the 1,000 test images.

3. Finally, the performance of the Bayes-optimal classifier on adversarially perturbed test images (size 1,000) is evaluated.

In our first experiment, we consider a fixed number of components $K = 10$ along with three different latent dimensions $\ell = 1, 10,$ and 100. For each pair of (K, ℓ), we randomly select one sample with label 6 among the original 1000 test images. For PGD adversarial attacks, we start from $\varepsilon = 0$, and incrementally increase the adversary's ℓ_2-bounded power ε until to the point that the Bayes-optimal classifier fails to correctly label the sample. Figure 8 displays the original, adversarial attack, and the perturbed images for each ℓ value. The Bayes-optimal classifier fails at adversarial power $\varepsilon = 22.6, 11, 7$ for $\ell = 1, 10, 100$ respectively. The result conforms to the fact that samples coming from a higher value of dimension ratio d/ℓ (here $d = 784$) indeed require stronger adversarial attacks (larger ε).

In the second experiment, we consider the two choices of $K = 1$ and $K = 10$ and vary the latent dimension ℓ. For each pair of (K, ℓ), we repeat the above three-step procedure for adversary's ℓ_2-bounded power $\varepsilon = 12$ and compute the boundary risk of the Bayes-optimal classifier on the PGD and FGM perturbed images. The plots are included in Figure 9. As observed, by increasing the dimension ratio d/ℓ, the boundary risk of the Bayes-optimal classifier decreases to zero.

6. Conclusion

In this paper, we studied the role of data distribution (in particular latent low-dimensional manifold structures of data) on the tradeoff between robustness (against adversarial perturbations in the input, at test time) and generalization (performance on test data drawn from the same distribution as training data). We developed a theory for two widely used classification setups (Gaussian-mixture model and generalized linear model), showing that as the ratio of the ambient dimension to the manifold dimension grows, one can obtain models which both are robust and generalize well. This highlights the role of exploiting underlying data structures in improving robustness and also in mitigating the tradeoff between generalization and robustness. Through numerical experiments, we demonstrate that low-dimensional manifold structure of data, even if not exploited by the training method, can still weaken the robustness-generalization tradeoff.

Endnote

[1] We will drop the index p and write ε for the adversary's power when it is clear from the context.

References

Biggio B, Corona I, Maiorca D, Nelson B, Srndić N, Laskov P, Giacinto G, Roli F (2013) Evasion attacks against machine learning at test time. *Machine Learn. Knowledge Discovery Databases: Eur. Conf., ECML PKDD 2013 Proc., Part III 13* (Springer, Berlin), 387–402.

Carlini N, Mishra P, Vaidya T, Zhang Y, Sherr M, Shields C, Wagner D, Zhou W (2016) Hidden voice commands. *25th USENIX Security Sympos. (USENIX Security 16)* (USENIX Association, Berkeley, CA), 513–530.

Costa JA, Hero AO (2004) Learning intrinsic dimension and intrinsic entropy of high-dimensional datasets. *2004 12th Eur. Signal Processing Conf.* (IEEE, New York), 369–372.

Deng L (2012) The MNIST database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Process. Mag.* 29(6):141–142.

Devroye L, Györfi L, Lugosi G (2013) *A Probabilistic Theory of Pattern Recognition*, vol. 31 (Springer Science & Business Media, New York).

Dobriban E, Hassani H, Hong D, Robey A (2020) Provable tradeoffs in adversarially robust classification. Preprint, submitted June 9, https://arxiv.org/abs/2006.05161.

Dong Y, Deng Z, Pang T, Su H, Zhu J (2020) Adversarial distributional training for robust deep learning. Preprint, submitted February 14, https://arxiv.org/abs/2002.05999.

Ghahramani Z, Hinton GE (1996) The EM algorithm for mixtures of factor analyzers. Technical report, Technical Report CRG-TR-96-1, University of Toronto.

Goodfellow I, Bengio Y, Courville A (2016) *Deep Learning* (MIT Press, Cambridge, MA).

Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. *Adv. Neural Inform. Process. Systems* 27:1–9.

Goodfellow IJ, Shlens J, Szegedy C (2015) Explaining and harnessing adversarial examples. Preprint, submitted December 20, 2014, https://arxiv.org/abs/1412.6572.

Hampel FR (1968) *Contributions to the Theory of Robust Estimation* (University of California, Berkeley).

Huber PJ (1992) Robust estimation of a location parameter. *Breakthroughs in Statistics* (Springer, New York), 492–518.

Jalal A, Ilyas A, Daskalakis C, Dimakis AG (2017) The robust manifold defense: Adversarial training using generative models. Preprint, submitted December 26, https://arxiv.org/abs/1712.09196.

Javanmard A, Soltanolkotabi M (2022) Precise statistical analysis of classification accuracies for adversarial training. *Ann. Statist.* 50(4):2127–2156.

Javanmard A, Soltanolkotabi M, Hassani H (2020) Precise tradeoffs in adversarial training for linear regression. Proc. Machine Learn. Res., Conf. *Learn. Theory (COLT)*, vol. 125 (PMLR), 2034–2078.

Lin W-A, Lau CP, Levine A, Chellappa R, Feizi S (2020) Dual manifold adversarial robustness: Defense against Lp and non-Lp adversarial attacks. Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H, eds. *Advances in Neural Information Processing Systems*, vol. 33 (Curran Associates, Inc., New York), 3487–3498.

Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A (2018) Toward deep learning models resistant to adversarial attacks. *6th Internat. Conf. Learn. Representations, ICLR 2018* (OpenReview.net, Vancouver).

Mehrabi M, Javanmard A, Rossi RA, Rao A, Mai T (2021) Fundamental tradeoffs in distributionally adversarial training. *Proc. 38th Internat. Conf. Machine Learn.*, vol. 139 (PMLR), 7544–7554.

Min Y, Chen L, Karbasi A (2020) The curious case of adversarially robust models: More data can help, double descend, or hurt generalization. Preprint, submitted February 25, https://arxiv.org/abs/2002.11080.

Papernot N, Faghri F, Carlini N, Goodfellow I, Feinman R, Kurakin A, Xie C, et al (2016) Technical report on the Cleverhans v2.1.0 adversarial examples library. Preprint, submitted October 3, https://arxiv.org/abs/1610.00768.

Pydi MS, Jog V (2020) Adversarial risk via optimal transport and optimal couplings. *Internat. Conf. Machine Learn.* (PMLR), 7814–7823.

Raghunathan A, Xie SM, Yang F, Duchi JC, Liang P (2019) Adversarial training can hurt generalization. Preprint, submitted June 14, https://arxiv.org/abs/1906.06032.

Richardson E, Weiss Y (2018) On GANs and GMMs. *Adv. Neural Inform. Process. Systems.* 31:1–12.

Richardson E, Weiss Y (2021) A bayes-optimal view on adversarial examples. *J. Machine Learn. Res.* 22(221):1–28.

Rozza A, Lombardi G, Ceruti C, Casiraghi E, Campadelli P (2012) Novel high intrinsic dimensionality estimators. *Machine Learn.* 89(1):37–65.

Song Y, Shu R, Kushman N, Ermon S (2018) Constructing unrestricted adversarial examples with generative models. Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R, eds. *Advances in Neural Information Processing Systems*, vol. 31 (Curran Associates, Inc., New York), 1–12.

Spigler S, Geiger M, Wyart M (2020) Asymptotic learning curves of kernel methods: Empirical data vs. teacher–student paradigm. *J. Statist. Mech. Theory Experiment* 2020(12):124001.

Staib M, Jegelka S (2017) Distributionally robust deep learning as a generalization of adversarial training. *NIPS Workshop on Machine Learning and Computer Security.*

Stutz D, Hein M, Schiele B (2019) Disentangling adversarial robustness and generalization. *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognition* (IEEE Computer Society, Washington, DC), 6976–6987.

Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, Fergus R (2014) Intriguing properties of neural networks. *Internat. Conf. Learn. Representations (ICLR)* (OpenReview.net, Banff, Canada).

Tramèr F, Papernot N, Goodfellow I, Boneh D, McDaniel P (2017) The space of transferable adversarial examples. Preprint, submitted April 11, https://arxiv.org/abs/1704.03453.

Tsipras D, Santurkar S, Engstrom L, Turner A, Madry A (2018) Robustness may be at odds with accuracy. Preprint, submitted May 30, https://arxiv.org/abs/1805.12152.

Tukey JW (1960) A survey of sampling from contaminated distributions. *Contributions to Probability and Statistics* (Princeton University, Princeton, NJ), 448–485.

Vaidya T, Zhang Y, Sherr M, Shields C (2015) Cocaine noodles: exploiting the gap between human and machine speech recognition. *9th USENIX Workshop on Offensive Technologies WOOT 15)* (USENIX Association, Berkeley, CA).

Xing Y, Zhang R, Cheng G (2021) Adversarially robust estimate and risk analysis in linear regression. *Internat. Conf. Artificial Intelligence Statist.* (PMLR), 514–522.

Yang Y-Y, Rashtchian C, Zhang H, Salakhutdinov RR, Chaudhuri K (2020) A closer look at accuracy vs. robustness. Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H, eds. *Advances in Neural Information Processing Systems*, vol. 33 (Curran Associates, Inc., New York), 8588–8601.

Zhang G, Yan C, Ji X, Zhang T, Zhang T, Xu W (2017) DolphinAttack: Inaudible voice commands. *Proc. 2017 ACM SIGSAC Conf. Comput. Comm. Security* (Association for Computing Machinery, New York), 103–117.

Zhang H, Yu Y, Jiao J, Xing EP, Ghaoui LE, Jordan MI (2019) Theoretically principled trade-off between robustness and accuracy. *Proc. 36th Internat. Conf. Machine Learn.* (PMLR, New York), 7472–7482.

Adel Javanmard is an associate professor in the Department of Data Sciences and Operation at the Marshall School of Business, the University of Southern California, where he also holds a courtesy appointment with the computer science department. Prior to joining USC in 2015, he was a NSF postdoctoral research fellow at the Center for Science of Information, with worksite at UC Berkeley and Stanford University. He completed his PhD at Stanford University in 2014. His research interests are broadly in the area of high-dimensional statistics, machine learning, optimization, and personalized decision-making. Adel is the recipient of several awards and fellowships, including 2021 Alfred P. Sloan Research Fellow in Mathematics, the IMS Tweedie Researcher award, the NSF CAREER award, Adobe Faculty Research awards (2020 and 2022), Google Faculty Research award, the Thomas Cover dissertation award from the IEEE Society, Douglas Basil Award for Junior Business Faculty, the Zumberge Faculty Research and Innovation award, and the NSF CSoI Postdoctoral Fellowship.

Mohammad Mehrabi is a PhD candidate in the Department of Data Sciences and Operations at the Marshall School of Business, University of Southern California. He is interested in high-dimensional statistics, machine learning, and uncertainty-aware decision making. Mohammad is the recipient of USC Marshall Outstanding Researcher Award, and USC Marshall Graduate Fellowship. He is also awarded the USC CAMS Prize Honorable Mention.