Robustness of classifiers to uniform ℓ_p and Gaussian noise

Jean-Yves Franceschi
Ecole Normale Supérieure de Lyon
LIP, UMR 5668

Alhussein Fawzi
UCLA Vision Lab[*]

Omar Fawzi
Ecole Normale Supérieure de Lyon
LIP, UMR 5668

Abstract

We study the robustness of classifiers to various kinds of random noise models. In particular, we consider noise drawn uniformly from the ℓ_p ball for $p \in [1, \infty]$ and Gaussian noise with an arbitrary covariance matrix. We characterize this robustness to random noise in terms of the distance to the decision boundary of the classifier. This analysis applies to linear classifiers as well as classifiers with locally approximately flat decision boundaries, a condition which is satisfied by state-of-the-art deep neural networks. The predicted robustness is verified experimentally.

1 INTRODUCTION

Image classification techniques have recently witnessed major advances leading to record performances on challenging datasets [He et al., 2016, Krizhevsky et al., 2012]. Besides reaching low classification error, it is equally important that classifiers deployed in real-world environments correctly classify perturbed and noisy samples. Specifically, when a sufficiently small perturbation alters a sample, it is desirable that the estimated label of the classifier remains unchanged. Altering perturbations can take various forms, such as additive perturbations, geometric transformations or occlusions for image data. The analysis of the robustness of classifiers under these perturbation regimes is crucial for unraveling their fundamental vulnerabilities. For example, state-of-the-art image classifiers have recently been empirically shown to be vulnerable to well-sought imperceptible additive perturbations [Biggio et al., 2013, Szegedy et al., 2014], and to more physically plausible nuisances in [Sharif et al., 2016]. The goal of this paper is to derive precise quantitative results on the robustness of general classifiers to *random* noise.

We specifically analyze two random noise models. Under our first perturbation model, we assume that noise is sampled uniformly at random from the ℓ_p ball, for $p \in [1, \infty]$. Different values of p allow us to model very different noise regimes; e.g., $p = 1$ corresponds to sparse noise, whereas $p = \infty$ models dense noise typically resulting from signal quantization. Under our second perturbation regime, the noise is modeled as Gaussian with arbitrary covariance matrix Σ. Our contributions are summarized as follows:

- For linear classifiers, we characterize up to constants the robustness to random noise, as a function of the distance to the decision boundary. We show in particular that, provided the weight vector of the linear classifier is randomly chosen, the robustness to random noise (uniform and Gaussian) scales as $\sqrt{d}$ times the distance to the decision boundary.
- We extend the results to nonlinear classifiers, and show that when the decision boundary is locally approximately flat (which is the case for state-of-the-art classifiers), the above result notably holds.
- Through experimental evidence on state-of-the-art image classifiers (deep nets), we show that the proposed bounds predict accurately the robustness of such classifiers. We finally show that our analysis predicts the high robustness of such classifiers to image quantization, which confirms previous empirical evidence.

Related work. The robustness properties of linear and kernel SVM classifiers have been studied in [Xu et al., 2009, Biggio et al., 2013], and robust optimization approaches for constructing robust classifiers have been proposed

Proceedings of the 21^{st} International Conference on Artificial Intelligence and Statistics (AISTATS) 2018, Lanzarote, Spain. PMLR: Volume 84. Copyright 2018 by the author(s).

[*]Now at DeepMind.

[Caramanis et al., 2012, Lanckriet et al., 2003]. More recently, the robustness properties of deep neural networks have been investigated. In particular, [Szegedy et al., 2014] shows that deep neural networks are not robust to worst-case, or adversarial, perturbations. Several works have followed and attempted to provide explanations to the vulnerability [Goodfellow et al., 2015, Tabacof and Valle, 2016, Tanay and Griffin, 2016, Sabour et al., 2016]. In particular, it was shown theoretically that the ratio of robustness to random noise and robustness to adversarial perturbations measured in the ℓ_2 norm scales as $\sqrt{d}$ for linear classifiers in [Fawzi et al., 2018] and more general classification functions in [Fawzi et al., 2016]. Therefore, when the data is sufficiently high dimensional, the robustness to adversarial perturbations is very small, which gives an explanation to the imperceptible nature of such perturbations. Our work generalizes [Fawzi et al., 2016] to broader noise regimes, such as sparse noise, quantization noise, or correlated Gaussian noise. Indeed, we follow a similar methodology to that of [Fawzi et al., 2016], where we first establish results for the linear case, and then extend the results to nonlinear classifiers satisfying a locally approximately flat decision boundary.

Outline. This paper is organized as follows. Section 2 introduces the framework of the robustness to random and adversarial perturbations. Section 3 presents theoretical estimates of such robustnesses for linear classifiers, which are generalized in Section 4 for classifiers with a locally approximately flat decision boundary. Section 5 then details experiments showing the validity of our bounds for state-of-the-art classifiers and exposing some applications of our results.

2 DEFINITIONS AND NOTATIONS

Let $f : \mathbb{R}^d \to \mathbb{R}^L$ be a L-class classifier. The estimated label of a datapoint $\boldsymbol{x} \in \mathbb{R}^d$ is set to $g(\boldsymbol{x}) = \operatorname{argmax}_k f_k(\boldsymbol{x})$, where $f_k(\boldsymbol{x})$ denotes the kth component of $f(\boldsymbol{x})$. Our goal in this paper is to analyze the robustness of f to random perturbations of the input. For that, we consider an arbitrary distribution ν on $\mathbb{R}^d$ that we interpret as giving the direction $\boldsymbol{v}$ of the noise, and we measure the length of the minimal scaling applied to $\boldsymbol{v}$ required to change the estimated label of f at $\boldsymbol{x}$ with probability at least ε. More precisely, let $\boldsymbol{v}$ be a random variable distributed according to ν; for a given $\varepsilon > 0$, we define $r_{\nu,\varepsilon}(\boldsymbol{x})$ as:

$$r_{\nu,\varepsilon}(\boldsymbol{x}) = \min_{\alpha} \left\{ |\alpha| \ \text{ s.t. } \ \mathbb{P}\left\{ g(\boldsymbol{x} + \alpha \boldsymbol{v}) \neq g(\boldsymbol{x}) \right\} \geq \varepsilon \right\}. \tag{1}$$

If the set is empty, we set $r_{\nu,\varepsilon}(\boldsymbol{x}) = +\infty$.[1] In this paper, we will focus on two families of choices for ν.

The first family is parameterized by a real number $p \in [1, \infty]$. The distribution ν is then the uniform distribution over the unit ball of ℓ_p^d, i.e., $\mathcal{B}_p = \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\|_p \leq 1\}$ where $\|\boldsymbol{x}\|_p = (\sum_{i=1}^d x_i^p)^{1/p}$. For this setting of distribution ν, we write $r_{p,\varepsilon}(\boldsymbol{x}) = r_{\nu,\varepsilon}(\boldsymbol{x})$. Observe that the Euclidean norm $\|.\|_2$ is invariant under an orthonormal basis change, but this is not the case for $\|.\|_p$ when $p \neq 2$, i.e., it depends on the basis that is chosen to write the signal $\boldsymbol{x}$; hence, this dependence also holds for $r_{p,\varepsilon}(\boldsymbol{x})$. Different choices of p allow us to span a range of realistic noise models. For example, choosing $p = 1$ leads to sparse noise vectors modeling salt and pepper noise, while $p = \infty$ leads to uniform noise vectors that allow us to model noise resulting from signal quantization [Bovik, 2005, Chapter 4.5]. An illustration of the different noise regimes can be found in Figure 1.

The second family is parameterized by an arbitrary positive definite matrix $\Sigma \in \mathbb{R}^{d \times d}$, that will generally be normalized with $\operatorname{Tr}(\Sigma) = 1$ to fix the scale. The distribution ν is then the multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix Σ. We use the notation $r_{\Sigma,\varepsilon}(\boldsymbol{x}) = r_{\nu,\varepsilon}(\boldsymbol{x})$ for this setting. A special case of this family is therefore the additive white Gaussian noise (where $\Sigma = \frac{I}{d}$); note however that this family is much broader and can model a noise that is correlated with the input $\boldsymbol{x}$, as no assumption is made on Σ.

In the remainder of this paper, our goal is to derive bounds on the robustness of classifiers f to random noise sampled from either of these two families. To do so, we first define a key quantity for our analysis, the robustness to worst-case perturbations:

$$\boldsymbol{r}_p^*(\boldsymbol{x}) = \operatorname*{argmin}_{\boldsymbol{r}} \left\{ \|\boldsymbol{r}\|_p \ \text{ s.t. } \ g(\boldsymbol{x} + \boldsymbol{r}) \neq g(\boldsymbol{x}) \right\}. \tag{2}$$

In other words, $\|\boldsymbol{r}_p^*(\boldsymbol{x})\|_p$ quantifies the length of the minimal perturbation required to change the estimated label of the classifier, or equivalently, the distance from the data point $\boldsymbol{x}$ to the decision boundary of the classifier. $\boldsymbol{r}_p^*(\boldsymbol{x})$ is often alternatively referred to as an *adversarial* perturbation, as it corresponds to the least noticeable perturbation an adversary would apply to fool a classifier. Note that, like $r_{p,\varepsilon}(\boldsymbol{x})$, it heavily depends on the choice of norm ℓ_p, and thus on the choice of orthonormal basis. Figure 2 illustrates the dependence on p of this perturbation. Such perturbations, which have been the subject of intense

[1] We should also technically consider the closure of the set to ensure the minimum is achieved, but we will avoid such technicalities throughout the paper as they are of no relevance for our study.

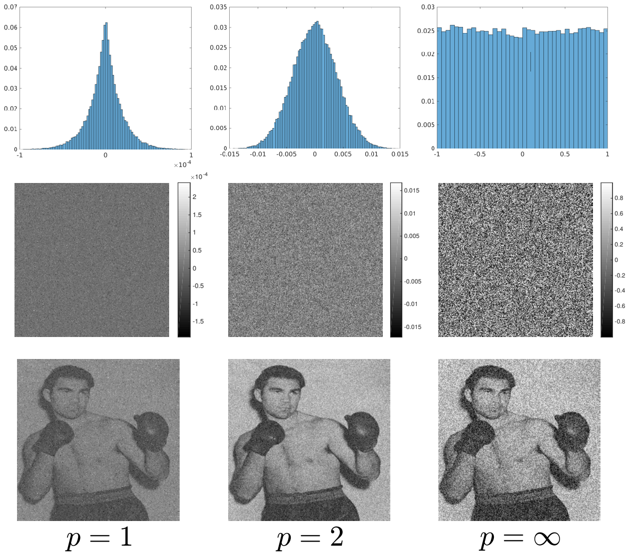

Figure 1: Illustration of noise with different values of p. First row: histogram of uniformly sampled noise from the unit ball of ℓ_p^d. Second row: Example of noise image. Third row: Example of noisy image. Note that different values of p result in perceptually different noise images.

studies, will be used to derive guarantees on the robustness to random noise.

In the next sections, we characterize the robustness of linear and nonlinear classifiers to random perturbations in terms of the robustness to worst-case perturbations. In the case of Gaussian random noise, we focus for $\boldsymbol{r}_p^*(\boldsymbol{x})$ on the norm $\|.\| = \|.\|_2$, even though all our results can be generalized to p-norms.

3 ROBUSTNESS OF LINEAR CLASSIFIERS

For simplicity of exposition, we state our results for binary classifiers, and we extend the results for multiclass classifiers in the supplementary material. The proofs may also be found in the supplementary material.

We consider in this section the particular case where f is a *linear* classifier, i.e., all the f_k's are linear functions. In particular, in the binary case, the setting can be simplified by considering a single linear function

$$f : \boldsymbol{x} \mapsto \boldsymbol{w}^T \boldsymbol{x} + \boldsymbol{b}. \tag{3}$$

In this case,[2] $g(\boldsymbol{x}) = 1$ if and only if $f(\boldsymbol{x}) > 0$.

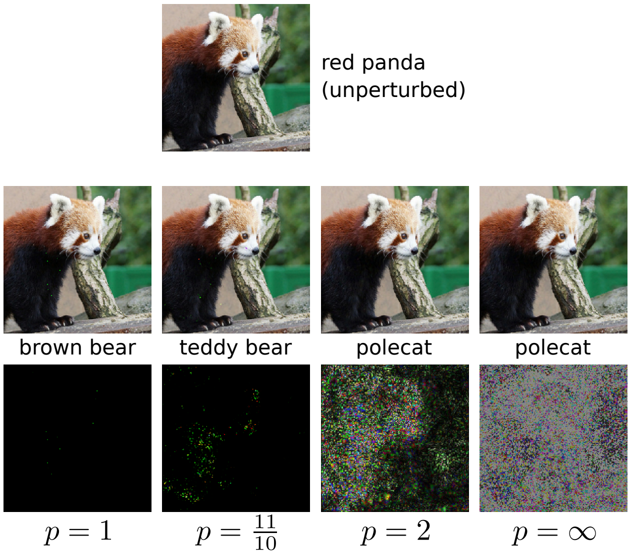

Figure 2: Illustration of adversarial perturbations with different values of p. First row: original image and its classification. Second row, for each column, from bottom to top: chosen p, adversarial perturbation, perturbed image with its classification. When $p \to \infty$, the perturbation tends to be distributed over all pixels; when $p \to 1$, it tends to be distributed over few pixels. Adversarial perturbations were estimated on the VGG-19 classifier [Simonyan and Zisserman, 2014] using the method presented in [Moosavi-Dezfooli et al., 2016].

3.1 Uniform ℓ_p noise

The following result bounds the robustness of a linear classifier to uniformly random noise with respect to its robustness to adversarial perturbations, for any norm ℓ_p.

Theorem 1. *Let $p \in [1, \infty]$. Let $p' \in [1, \infty]$ be such that $\frac{1}{p} + \frac{1}{p'} = 1$. There exist constants $\varepsilon_0, \zeta_1(\varepsilon), \zeta_2(\varepsilon)$ such that, for all $\varepsilon < \varepsilon_0$:*

$$\zeta_1(\varepsilon) d^{1/p} \frac{\|\boldsymbol{w}\|_{p'}}{\|\boldsymbol{w}\|_2} \leq \frac{r_{p,\varepsilon}(\boldsymbol{x})}{\|\boldsymbol{r}_p^*(\boldsymbol{x})\|_p} \leq \zeta_2(\varepsilon) d^{1/p} \frac{\|\boldsymbol{w}\|_{p'}}{\|\boldsymbol{w}\|_2}.$$

We can take $\zeta_1(\varepsilon) = C\sqrt{\varepsilon}$,[3] and $\zeta_2(\varepsilon) = \sqrt{\frac{1}{c - \sqrt{c'\varepsilon}}}$, for some constants C, c, c'.

More details on constants C, c, c' are available in the supplementary material. In words, our result demonstrates that $r_{p,\varepsilon}(\boldsymbol{x})$ is well estimated by $\|\boldsymbol{r}_p^*(\boldsymbol{x})\|_p$ times a multiplicative factor that is independent of $\boldsymbol{x}$ and is of the order $d^{1/p} \frac{\|\boldsymbol{w}\|_{p'}}{\|\boldsymbol{w}\|_2}$. The special case $p = 2$, for which this multiplicative factor becomes

[2] In the general multi-class setting, this corresponds to $f_0 = f$ and $f_1 = 0$.

[3] We show in the supplementary material that for $p > 1$, we can also choose $\zeta_1(\varepsilon) = \frac{C'}{\sqrt{\ln \frac{1}{\varepsilon}}}$ for some constant C'.

$\sqrt{d}$, was previously shown in [Fawzi et al., 2016] and [Fawzi et al., 2018]. For $p \neq 2$, this factor depends on the choice of the classifier through vector $\boldsymbol{w}$. Such a dependence was to be expected as the p-norm for $p \neq 2$ depends on the choice of basis. This dependence takes into account the relation between this choice of basis to write the signal and the direction $\boldsymbol{w}$ chosen by the classifier. For example, when $\boldsymbol{w} = (1\ 0\ \dots\ 0)^T$, we have a classifier that only uses the first component of the signal. So for $p = \infty$, the problem effectively becomes one-dimensional as only the first coordinate matters and we have $r_\infty(\boldsymbol{x}) = \|\boldsymbol{r}^*_\infty(\boldsymbol{x})\|_\infty$.

Nevertheless, for a typical choice of the vector $\boldsymbol{w}$ of the linear classifier, this factor stays of order $\sqrt{d}$ if $p > 1$.

Proposition 1. *For any $p \in (1, \infty]$, if $\boldsymbol{w}$ is a random direction uniformly distributed over the unit ℓ_2-sphere, then, as $d \to \infty$,*

$$\frac{d^{1/p}\frac{\|\boldsymbol{w}\|_{p'}}{\|\boldsymbol{w}\|_2}}{\sqrt{d}} \xrightarrow[a.s.]{} \sqrt{2}\left(\frac{\Gamma\left(\frac{2p-1}{2(p-1)}\right)}{\sqrt{\pi}}\right)^{1-\frac{1}{p}}.$$

Moreover, for $p = 1$,

$$\frac{d\frac{\|\boldsymbol{w}\|_\infty}{\|\boldsymbol{w}\|_2}}{\sqrt{2d\ln d}} \xrightarrow[a.s.]{} 1.$$

While this result is only asymptotic and valid for random decision hyperplanes, we experimentally show in Section 5 that its dependence in p allows us to propose an estimate providing a very good approximation of the robustness to random noise.

3.2 Gaussian noise

In the case where the uniformly random noise is replaced by a Gaussian noise with a given covariance matrix Σ, we can similarly characterize the ratio $\frac{r_{\Sigma,\varepsilon}(\boldsymbol{x})}{\|\boldsymbol{r}^*_2(\boldsymbol{x})\|_2}$ as a function of Σ and $\boldsymbol{w}$ as follows.

Theorem 2. *Let Σ be a $d \times d$ positive semidefinite matrix with $\mathrm{Tr}(\Sigma) = 1$.*[4] *There exist constants $\varepsilon'_0, \zeta'_1(\varepsilon), \zeta'_2(\varepsilon)$ such that, for all $\varepsilon < \varepsilon'_0$:*

$$\zeta'_1(\varepsilon)\frac{\|\boldsymbol{w}\|_2}{\|\sqrt{\Sigma}\boldsymbol{w}\|_2} \leq \frac{r_{\Sigma,\varepsilon}(\boldsymbol{x})}{\|\boldsymbol{r}^*_2(\boldsymbol{x})\|_2} \leq \zeta'_2(\varepsilon)\frac{\|\boldsymbol{w}\|_2}{\|\sqrt{\Sigma}\boldsymbol{w}\|_2}.$$

We can take $\varepsilon'_0 = \frac{1}{3}$, $\zeta'_1(\varepsilon) = \sqrt{\frac{1}{2\ln\left(\frac{1}{\varepsilon}\right)}}$ and $\zeta'_2(\varepsilon) = \sqrt{\frac{1}{1-\sqrt{3\varepsilon}}}$.

In this case, the multiplicative factor between robustnesses to random and adversarial perturbations is of the order $\frac{\|\boldsymbol{w}\|_2}{\|\sqrt{\Sigma}\boldsymbol{w}\|_2}$. Note that this factor lies in between $\lambda_{\max}(\sqrt{\Sigma})^{-1}$ and $\lambda_{\min}(\sqrt{\Sigma})^{-1}$. However, these values correspond to extremal cases, and for most choices of $\boldsymbol{w}$, this factor will be determined by a convex combination of eigenvalues of Σ. More precisely, if $\boldsymbol{u}_1, \dots, \boldsymbol{u}_d$ are the eigenvectors of Σ with eigenvalues $\lambda_1^2, \dots, \lambda_d^2$ and assuming $\|\boldsymbol{w}\|_2 = 1$ (without loss of generality), the factor is given by a weighted average of the eigenvalues $\lambda_1^2, \dots, \lambda_d^2$:

$$\frac{1}{\sqrt{\sum_{i=1}^d \lambda_i^2 |\boldsymbol{u}_i^T\boldsymbol{w}|^2}}.$$

In particular, if $\Sigma = \frac{1}{d}I_d$, then the factor is $\sqrt{d}$. Even more generally, for typical choices of $\boldsymbol{w}$ we expect $|\boldsymbol{u}_i^T\boldsymbol{w}|^2$ be of order $\frac{1}{d}$, in which case the factor will also be of order $\sqrt{d}$.

4 ROBUSTNESS OF NONLINEAR CLASSIFIERS

We now consider the general case where f is a nonlinear classifier. The goal of this section is to derive relations between $r_{p,\varepsilon}(\boldsymbol{x})$ and $\|\boldsymbol{r}^*_p(\boldsymbol{x})\|_p$ in this general case, under a reasonable hypothesis on the geometry of the decision boundary.

4.1 Locally Approximately Flat (LAF) Decision Boundary Model

Before giving the formal definition, let us describe the main idea behind the *Locally Approximately Flat* (LAF) decision boundary model. This model requires that the decision boundary can be locally sandwiched between two hyperplanes that are parallel to the tangent hyperplane. We do not ask for this to hold for every point on the decision boundary, but only to hold for the closest points on the decision boundary of our data points.

Definition 1 (LAF model)**.** *Let f be a binary classifier with smooth*[5] *decision boundary $\mathcal{S} = \{x \in \mathbb{R}^d : f(\boldsymbol{x}) = 0\}$. For $\boldsymbol{x}^* \in \mathcal{S}$, define $\mathcal{T}(\boldsymbol{x}^*)$ to be the hyperplane tangent to $\mathcal{S}$ at point $\boldsymbol{x}^*$. For $\boldsymbol{x} \in \mathbb{R}^d$ and $\boldsymbol{x}^* \in \mathcal{S}$, we define $\mathcal{H}^-_\gamma(\boldsymbol{x}, \boldsymbol{x}^*)$ to be the halfspace of points that are on the side of $\boldsymbol{x}$ of the hyperplane parallel to $\mathcal{T}(\boldsymbol{x}^*)$ that passes though the point $\gamma\boldsymbol{x} + (1-\gamma)\boldsymbol{x}^*$. Similarly, $\mathcal{H}^+_\gamma(\boldsymbol{x}, \boldsymbol{x}^*)$ is the halfspace of points that are not on the side of $\boldsymbol{x}$ of the hyperplane parallel to $\mathcal{T}(\boldsymbol{x}^*)$*

[4] Note that the condition $\mathrm{Tr}(\Sigma) = 1$ is not needed for the statement but its motivation is to fix the scale of $r_{\Sigma,\varepsilon}(\boldsymbol{x})$.

[5] This is a strong assumption that is only used here for the sake of exposition. Actually, the tangent $\mathcal{T}(\boldsymbol{x}^*)$ in the definition can be replaced by any hyperplane intersecting the decision boundary at $\boldsymbol{x}^*$. Then, in the results conditioned by the LAF model, the gradient of f at $\boldsymbol{x}^*$ can be replaced by a normal vector to this plane.

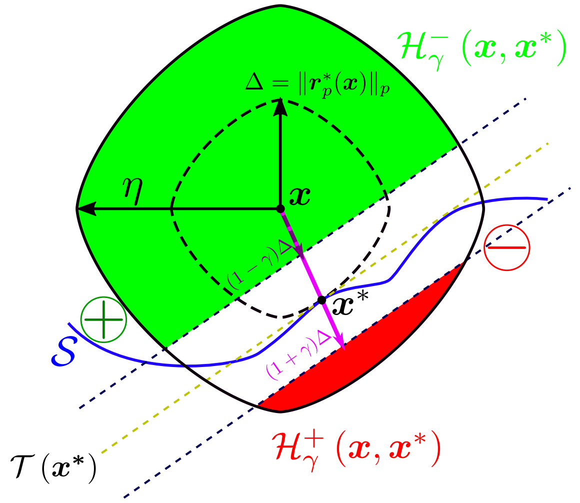
$\mathcal{H}_\gamma^-(\boldsymbol{x},\boldsymbol{x}^*)$
$\Delta = \|\boldsymbol{r}_p^*(\boldsymbol{x})\|_p$
η
$\boldsymbol{x}$
$(1-\gamma)\Delta$
$\boldsymbol{x}^*$
$(1+\gamma)\Delta$
⊕
⊖
$\mathcal{S}$
$\mathcal{T}(\boldsymbol{x}^*)$
$\mathcal{H}_\gamma^+(\boldsymbol{x},\boldsymbol{x}^*)$

Figure 3: Illustration of a (γ, η)-LAF classifier with $p = \frac{3}{2}$. $\mathcal{S}$ is the decision boundary of f, separating instances of the same predicted class as $\boldsymbol{x}$ (on the side of $\boldsymbol{x}$) from instances whose classification differs from $\boldsymbol{x}$ (on the other side). By definition, all inputs in $\mathcal{H}_\gamma^+(\boldsymbol{x},\boldsymbol{x}^*) \cap \mathcal{B}_p(\boldsymbol{x},\eta)$ (red area, other side of $\mathcal{S}$) are classified differently from $\boldsymbol{x}$, while the inputs in $\mathcal{H}_\gamma^-(\boldsymbol{x},\boldsymbol{x}^*) \cap \mathcal{B}_p(\boldsymbol{x},\eta)$ (green area, same side of $\mathcal{S}$ as $\boldsymbol{x}$) are classified as $\boldsymbol{x}$.

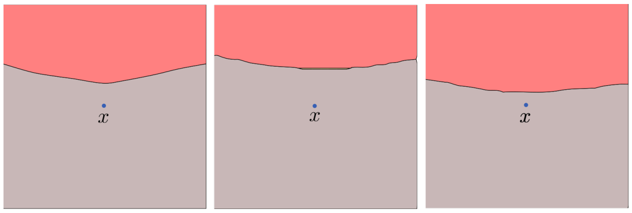

Figure 4: Two dimensional normal cross-sections of the decision boundary of a deep network classifier along random directions, in the vicinity of different natural images (denoted by $\boldsymbol{x}$ for each cross-section). The CaffeNet architecture [Jia et al., 2014] trained on ImageNet [Russakovsky et al., 2015] was used.

that passes though the point $\boldsymbol{x}^* + \gamma(\boldsymbol{x}^* - \boldsymbol{x})$ (see Figure 3).

We say that f is (γ, η)-Locally Approximately Flat at point $\boldsymbol{x}$ if for $\boldsymbol{x}^* \in \mathcal{S}$ minimizing $\|\boldsymbol{x} - \boldsymbol{x}^*\|_p$, the set $\mathcal{H}_\gamma^-(\boldsymbol{x},\boldsymbol{x}^*) \cap \mathcal{B}_p(\boldsymbol{x},\eta)$ is classified as $\boldsymbol{x}$ and $\mathcal{H}_\gamma^+(\boldsymbol{x},\boldsymbol{x}^*) \cap \mathcal{B}_p(\boldsymbol{x},\eta)$ is classified differently from $\boldsymbol{x}$. Here $\mathcal{B}_p(\boldsymbol{x},\eta)$ is the ℓ_p-ball centered at $\boldsymbol{x}$ with radius η.

The LAF model assumes that the decision boundary can be *locally* approximated by a hyperplane, in the vicinity of images $\boldsymbol{x}$ sampled from the data distribution. It should be noted that, in order to be able to define locality, we need a distance measure and thus the LAF property depends implicitly on the choice of norm. For $\gamma = 0$, the LAF model corresponds to locally exactly flat decision boundaries (no curvature). If in addition $\eta = \infty$, this corresponds to a linear decision boundary.

Prior empirical evidence has shown that state-of-the-art deep neural network classifiers have decision boundaries that are approximately flat along random directions [Warde-Farley et al., 2016, Fawzi et al., 2016]. Normal two-dimensional cross-sections (along random directions) of the decision boundary are illustrated in Figure 4. Note that such cross-sections have very low curvature, thereby providing evidence that the LAF assumption holds approximately (at least, with high probability) for complex classifiers, such as modern deep neural networks. It should further be noted that the LAF model is tightly related to the curvature condition of the decision boundary in [Fawzi et al., 2016]. The LAF model, however, does not assume any regularity condition on the decision surface, which is nonsmooth in many settings (e.g., deep neural networks due to piecewise linear activation functions). Finally, it should be noted that, for the sake of clarity, we assumed that the entire set $\mathcal{H}_\gamma^+(\boldsymbol{x},\boldsymbol{x}^*) \cap \mathcal{B}_p(\boldsymbol{x},\eta)$ (respectively, $\mathcal{H}_\gamma^-(\boldsymbol{x},\boldsymbol{x}^*) \cap \mathcal{B}_p(\boldsymbol{x},\eta)$) is classified differently from $\boldsymbol{x}$ (respectively, similarly to $\boldsymbol{x}$); however, the results in this section hold even if these conditions are only satisfied with high probability.

4.2 Robustness Results Under LAF Model

Our next result shows that, provided f is Locally Approximately Flat, a very similar result to Theorem 1 holds, with the normal vector $\boldsymbol{w}$ replaced by the gradient of f at the point $\boldsymbol{x}^*$ of the boundary that is closest to $\boldsymbol{x}$. It should be noted that for nonlinear classifiers, the gradient $\nabla f(\boldsymbol{x}^*)$ plays the same role as $\boldsymbol{w}$ for linear classifiers, as it is normal to the tangent to the decision boundary at $\boldsymbol{x}^*$.

Theorem 3. *Let $p \in [1, \infty]$. Let $p' \in [1, \infty]$ be such that $\frac{1}{p} + \frac{1}{p'} = 1$. Let $\varepsilon_0, \zeta_1(\varepsilon), \zeta_2(\varepsilon)$ be as in Theorem 1. Then, for all $\varepsilon < \varepsilon_0$, the following holds.*

Assume f is a classifier that is (γ, η)-LAF at point $\boldsymbol{x}$ and $\boldsymbol{x}^*$ be such that $\boldsymbol{r}_p^*(\boldsymbol{x}) = \boldsymbol{x}^* - \boldsymbol{x}$. Then:

$$(1-\gamma)\zeta_1(\varepsilon) d^{1/p} \frac{\|\nabla f(\boldsymbol{x}^*)\|_{p'}}{\|\nabla f(\boldsymbol{x}^*)\|_2} \le \frac{r_{p,\varepsilon}(\boldsymbol{x})}{\|\boldsymbol{r}_p^*(\boldsymbol{x})\|_p}$$

and

$$\frac{r_{p,\varepsilon}(\boldsymbol{x})}{\|\boldsymbol{r}_p^*(\boldsymbol{x})\|_p} \le (1+\gamma)\zeta_2(\varepsilon) d^{1/p} \frac{\|\nabla f(\boldsymbol{x}^*)\|_{p'}}{\|\nabla f(\boldsymbol{x}^*)\|_2},$$

provided

$$\eta \ge (1+\gamma)\zeta_2(\varepsilon) d^{1/p} \frac{\|\nabla f(\boldsymbol{x}^*)\|_{p'}}{\|\nabla f(\boldsymbol{x}^*)\|_2} \left\|\boldsymbol{r}_p^*(\boldsymbol{x})\right\|_p.$$

In the case where $\nabla f(\boldsymbol{x}^*)$ is uncorrelated with the basis used to write the signal (which we model by taking for $\nabla f(\boldsymbol{x}^*)$ a random direction in the ℓ_2 sphere), we obtain the same result as in Proposition 1 (i.e., by replacing $\boldsymbol{w}$ with $\nabla f(\boldsymbol{x}^*)$). This provides bounds on the robustness to random noise that only depend on $\|\boldsymbol{r}_2^*\|_2$ and d. We show that these asymptotic bounds provide accurate estimates of the empirical robustness in Section 5.

The result on Gaussian noise also holds for LAF classifiers.

Theorem 4. *Let Σ be a $d \times d$ positive semidefinite matrix with $\mathrm{Tr}(\Sigma) = 1$. Let $\varepsilon_0', \zeta_1'(\varepsilon), \zeta_2'(\varepsilon)$ as in Theorem 2. Then, for all $\varepsilon < \frac{1}{2}\varepsilon_0'$, the following holds.*

Assume f is a classifier that is (γ, η)-LAF at point $\boldsymbol{x}$ and $\boldsymbol{x}^*$ be such that $\boldsymbol{r}_2^*(\boldsymbol{x}) = \boldsymbol{x}^* - \boldsymbol{x}$. Then:

$$(1-\gamma)\zeta_1'\left(\frac{\varepsilon}{2}\right)\frac{\|\nabla f(\boldsymbol{x}^*)\|_2}{\|\sqrt{\Sigma}\nabla f(\boldsymbol{x}^*)\|_2} \leq \frac{r_{\Sigma,\varepsilon}(\boldsymbol{x})}{\|\boldsymbol{r}_2^*(\boldsymbol{x})\|_2}$$

and

$$\frac{r_{\Sigma,\varepsilon}(\boldsymbol{x})}{\|\boldsymbol{r}_2^*(\boldsymbol{x})\|_2} \leq (1+\gamma)\zeta_2'\left(\frac{3\varepsilon}{2}\right)\frac{\|\nabla f(\boldsymbol{x}^*)\|_2}{\|\sqrt{\Sigma}\nabla f(\boldsymbol{x}^*)\|_2},$$

provided, using $\psi(\varepsilon) = 8\,\mathrm{Tr}\left(\Sigma^2\right)\ln\frac{4}{\varepsilon}$,

$$\eta \geq (1+\gamma)(1+\psi(\varepsilon))\zeta_2'\left(\frac{3\varepsilon}{2}\right)\frac{\|\nabla f(\boldsymbol{x}^*)\|_2}{\|\sqrt{\Sigma}\nabla f(\boldsymbol{x}^*)\|_2}\|\boldsymbol{r}_2^*(\boldsymbol{x})\|_2.$$

5 EXPERIMENTS

Robustness of a binary linear classifier to uniform random noise. We now assess empirically our bounds for the robustness to random noise. We first consider the 10-class MNIST digit classification task [LeCun et al., 1998], and train a binary linear classifier separating digits 0 to 4 from digits 5 to 9, which achieves a performance of 84% in the test set. To assess our analytical results, we compare these to an empirical estimate of the robustness to uniform random noise for different values of p. It is based on the combination of the expressions found in Theorem 1 and Proposition 1; for a fixed ε, our estimate is:

$$\mathscr{E}_{r_p}(\boldsymbol{x}) = \zeta_0\sqrt{d}\left(\frac{\Gamma\left(\frac{2p-1}{2(p-1)}\right)}{\sqrt{\pi}}\right)^{1-\frac{1}{p}}\left\|\boldsymbol{r}_p^*(\boldsymbol{x})\right\|_p, \quad (4)$$

where ζ_0 is a constant. The empirical robustness of Eq. (1) is specifically computed through an exhaustive search of smallest radius of the ℓ_p ball leading to an ε fraction of misclassified samples. Note moreover that for this linear classifier, the worst-case robustness $\|\boldsymbol{r}_p^*\|_p$ is given by the distance to the hyperplane, and can therefore be computed in closed form (see the supplementary material). Figure 5 illustrates the empirical robustness, our theoretical bounds and our estimate (i.e., upper and lower bounds of Theorem 1, and estimate of Eq. (4)) with respect to p. In addition to providing accurate upper and lower bounds for all the range of tested p-norms, observe that our estimate provides a remarkably accurate approximation of the robustness to random noise, for all p. Our analytical results hence correctly predict the robustness behavior of this classifier through a wide variety of noise models, and can therefore be used to predict the robustness in these regimes.

Robustness of a multi-class deep neural network to uniform random noise. We now consider a more complex classification setting, where we evaluate the robustness of the VGG-19 deep neural network on the multi-class ImageNet dataset of natural images [Russakovsky et al., 2015]. Similarly to our experiment for the linear classifier, we compare the empirical value of the robustness for different values of p to our theoretical bounds from Theorem 3 and our estimate from Eq. (4). Note that unlike the previous case, the worst-case robustness $\|\boldsymbol{r}_p^*\|_p$ cannot be obtained in closed-form for deep networks; we therefore estimate it using the algorithm described in [Moosavi-Dezfooli et al., 2016]. The results are shown in Figure 6. Observe that, once again, our estimate predicts accurately the robustness of the deep neural network for different values of p. Hence, despite the high nonlinearity of the deep network as a function of the inputs, our bounds established under the LAF assumption hold accurately for all tested values of p.

Robustness of a deep neural network to quantization. We now leverage our analytical results to assess the robustness of a deep neural network classifier to image quantization. When a signal $\boldsymbol{x}$ is quantized into a discrete valued-signal $Q(\boldsymbol{x})$, the quantization noise $Q(\boldsymbol{x}) - \boldsymbol{x}$ is often modeled as a signal independent uniform random variable [Bovik, 2005, Chapter 4.5]. That is, under this assumption, $Q(\boldsymbol{x}) - \boldsymbol{x}$ is uniformly distributed over $\mathcal{B}_\infty(\boldsymbol{0}, \Delta/2)$, with Δ denoting the quantization step size. According to our analytical results in Section 4, the approximate step size Δ that the classifier can tolerate (without changing the estimated label of the quantized image) with probability $1 - \varepsilon$ is thus given by:

$$\Delta = \frac{2\zeta_0}{\sqrt{\pi}}\sqrt{d}\|\boldsymbol{r}_\infty^*(\boldsymbol{x})\|_\infty,$$

using the estimate of Eq. (4). Moreover, the number of quantization levels required to guarantee robustness

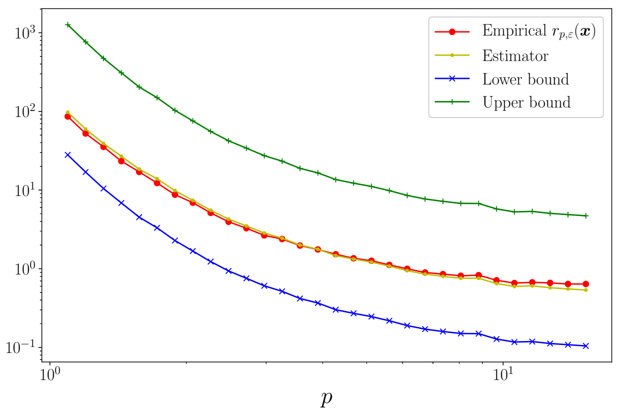

Figure 5: Empirical robustness to random uniform noise, derived upper and lower bounds (Theorem 1) and estimate from Eq. (4), as a function of p for a linear classifier trained on MNIST. For a given p, empirical robustness was computed through an exhaustive search of the smallest radius of the ball where an ε fraction of points sampled uniformly from the ball are misclassified. We choose $\varepsilon = 1.5\%$, empirically find $\zeta_0 \approx 0.72$, and run the experiments for each chosen p over 1,000 random images from the MNIST test set.

of the classifier is therefore estimated by

$$L_q = \frac{255}{\frac{2\zeta_0}{\sqrt{\pi}}\sqrt{d}\|\boldsymbol{r}^*_\infty(\boldsymbol{x})\|_\infty}. \qquad (5)$$

In other words, Eq. (5) predicts that images encoded with more than $\log_2\left(\frac{255}{\frac{2\zeta_0}{\sqrt{\pi}}\sqrt{d}\|\boldsymbol{r}^*_\infty(\boldsymbol{x})\|_\infty}\right)$ bits will have the same estimated label as the original image with high probability, despite quantization. Figure 7 shows that this prediction is a good approximation of the real quantization level computed for $8,000$ images from the ImageNet validation set for the VGG-19 classifier. In this experiment, we use a minimum variance quantization. Moreover, as commonly done, dithering is also applied to improve the perceptual quality of the quantized image. Interestingly, as predicted by our analysis, most images can be heavily quantized (with e.g., 3 bits) without changing the label of the classifier, despite the significant distortions to the images caused by heavy quantization (see Figure 8 for example images). Finally, note that our analytical results confirm and quantify earlier empirical observations that highlighted the high robustness of classifiers to compression mechanisms [Dodge and Karam, 2016, Paola and Schowengerdt, 1995].

Robustness to signal-dependent Gaussian noise. We now consider the case where some Gaussian noise that correlates with the input image is added to this image. That is, we consider a Gaussian noise $\mathcal{N}(0, \Sigma(\boldsymbol{x}))$, where $\Sigma(\boldsymbol{x})$ is a diagonal matrix such that $\Sigma(\boldsymbol{x})_{ii} = 1_{x_i \geq t} \cdot x_i$, where x_i denotes the value of pixel i, and t denotes a user-specified threshold.[6] $\Sigma(\boldsymbol{x})$ is further normalized to satisfy $\operatorname{Tr}(\Sigma(\boldsymbol{x})) = 1$. Under this noise model, noise is solely added to pixels that are "almost white" (i.e., pixels satisfying $x_i \geq t$), while all other pixels are left untouched. It should be noted that such signal-dependent noise models are commonly used to model physical deficiencies in acquisition, such as shot noise.

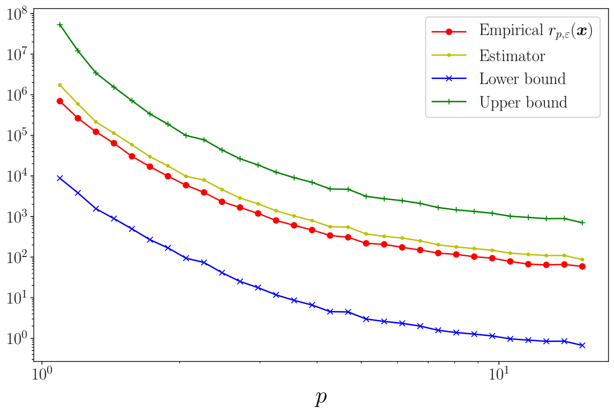

Figure 6: Empirical robustness to random uniform noise, derived upper and lower bounds (Theorem 3) and estimate from Eq. (4), as a function of p for the VGG-19 classifier trained on ImageNet. See the caption of Figure 5 for more details about the computation of $r_{p,\varepsilon}(\boldsymbol{x})$. We choose $\varepsilon = 1.5\%$, use $\zeta_0 \approx 0.72$ as in Figure 5, and run the experiments for each chosen p over 200 images from the ImageNet validation set.

Our analytical results for the Gaussian case predict that the robustness to such noise (provided the gradient directions are "typical") should be independent of the distribution of eigenvalues $\Sigma(\boldsymbol{x})$, and should moreover satisfy[7]

$$\frac{1}{2}\sqrt{d} \leq \frac{r_{\Sigma(\boldsymbol{x}),\varepsilon}(\boldsymbol{x})}{\|\boldsymbol{r}^*_2(\boldsymbol{x})\|_2} \leq 2\sqrt{d}, \qquad (6)$$

where $\varepsilon = 0.15$. To verify this hypothesis, we show in Figure 9 the ratio $\frac{r_{\Sigma(\boldsymbol{x}),\varepsilon}(\boldsymbol{x})}{\|\boldsymbol{r}^*_2(\boldsymbol{x})\|_2}$ over 30,000 images from the ImageNet validation set for the VGG-19 classifier, as a function of the "whiteness" of the image; i.e., $W(\boldsymbol{x}) = \sum_i 1_{x_i \geq t} x_i$. Similarly to previous experiments, $r_{\Sigma,\varepsilon}(\boldsymbol{x})$ is estimated using an exhaustive line

[6] We consider in practice color images; the quantity x_i refers in this case to $x_{i,r} + x_{i,g} + x_{i,b}$, where $x_{i,r}, x_{i,g}, x_{i,b}$ respectively denote the red, green and blue channels.

[7] We stress here that, due to the normalization $\operatorname{Tr}(\Sigma(\boldsymbol{x})) = 1$, the same amount of noise is added to all images. It is only the distribution of noise that differs: noise is concentrated on few pixels for images with few white pixels, and spread for white images.

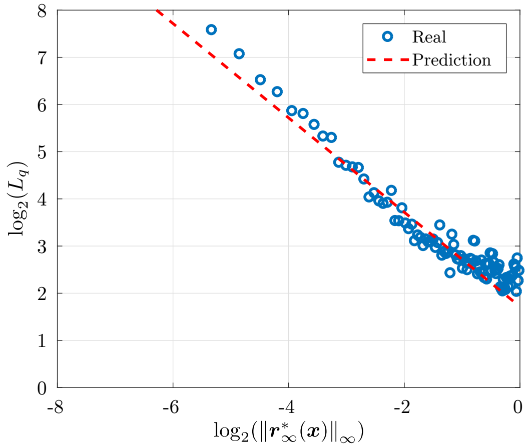

Figure 7: Minimum number of bits required to encode an image to guarantee similar estimated label as original image vs. $\log_2(\|\boldsymbol{r}^*_\infty(\boldsymbol{x})\|_\infty)$. *Real* points are computed through an exhaustive search of the required quantization level (with different images), and *Prediction* is computed using Eq. (5). We choose $\varepsilon = 1.5\%$ and $\zeta_0 = 0.72$ as in Figure 6.

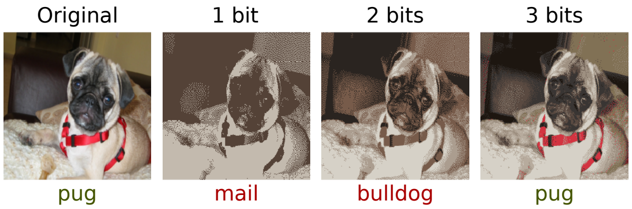

Figure 8: Example image, where a quantization (with dithering) using 3 bits leads to correct classification.

search. It can be seen that the ratio $\frac{r_{\Sigma(\boldsymbol{x}),\varepsilon}(\boldsymbol{x})}{\|\boldsymbol{r}^*_2(\boldsymbol{x})\|_2}$ approximately satisfies the bounds in Eq. (6), although the empirical ratio can surpass the upper bound, for images with significant white pixels. This is potentially due to our assumption on the randomness of the direction of the decision boundary, which can be violated in this case: in fact, white pixels (i.e., non-zero eigenvectors of Σ) appear often in the background of images, and are thus correlated with the decision boundaries of the classifier. Despite this assumption not being satisfied, our bounds allow us to predict fairly accurately the behavior of a complex deep network in presence of image-dependent Gaussian noise.

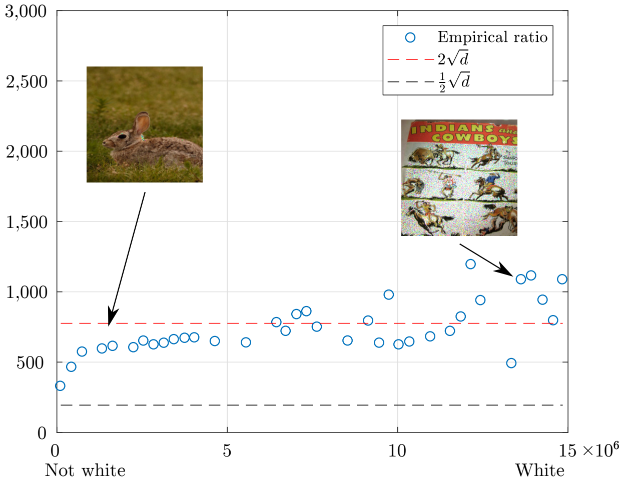

Figure 9: Fraction $\frac{r_{\Sigma(\boldsymbol{x}),\varepsilon}(\boldsymbol{x})}{\|\boldsymbol{r}^*_2(\boldsymbol{x})\|_2}$, where $\varepsilon = 15\%$, as a function of $W(\boldsymbol{x})$. $W(\boldsymbol{x})$ encodes how "white" the pixels of the image are. Under our noise model, images with small $W(\boldsymbol{x})$ will have the noise concentrated along a few pixels, while images with large $W(\boldsymbol{x})$ will have their noise spread across most pixels in the image. Each circle represents the average robustness ratio of images with the same $W(\boldsymbol{x})$.

6 CONCLUSION

We have derived precise bounds on the robustness of linear and nonlinear classifiers to random noise, under two noise distributions: uniform noise in the ℓ_p unit ball, and Gaussian noise. Our quantitative results show that state-of-the-art classifiers are orders of magnitude more robust to typical random noise than to worst-case perturbations, typically of order the square root of the input dimension. Such bounds are shown to hold in challenging settings, where a state-of-the-art deep network is used on a large scale multi-class dataset such as ImageNet. Our analysis can be leveraged to quantify the effect of many disturbances (e.g., image quantization) on classifiers, and provide robustness guarantees when such systems are deployed in real world environments. Moreover, our analysis allows us to draw links between different noise regimes, and show the effect of the robustness to adversarial perturbations (or equivalently, the distance to the decision boundary) on other noise regimes.

In this work, we have studied the robustness with respect to generic ℓ_p norms. For future work, we believe it would be very interesting to characterize the robustness of classifiers to random perturbations by using perceptual similarity metrics adapted to different modalities, such as images [Wang et al., 2004] and speech.

Acknowledgements

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

References

[Biggio et al., 2013] Biggio, B., Corona, I., Maiorca, D., Nelson, B., Šrndić, N., Laskov, P., Giacinto, G., and Roli, F. (2013). Evasion attacks against machine learning at test time. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 387–402. Springer Berlin Heidelberg.

[Bovik, 2005] Bovik, A. C. (2005). *Handbook of image and video processing*. Academic Press, 2nd edition.

[Caramanis et al., 2012] Caramanis, C., Mannor, S., and Xu, H. (2012). Robust optimization in machine learning. In Sra, S., Nowozin, S., and Wright, S. J., editors, *Optimization for machine learning*, chapter 14, pages 369–402. MIT Press.

[Dodge and Karam, 2016] Dodge, S. and Karam, L. (2016). Understanding how image quality affects deep neural networks. In *IEEE 8th International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6.

[Fawzi et al., 2018] Fawzi, A., Fawzi, O., and Frossard, P. (2018). Analysis of classifiers' robustness to adversarial perturbations. *Machine Learning*, 107(3):481–508.

[Fawzi et al., 2016] Fawzi, A., Moosavi-Dezfooli, S., and Frossard, P. (2016). Robustness of classifiers: from adversarial to random noise. In *Advances in Neural Information Processing Systems (NIPS) 29*, pages 1632–1640.

[Goodfellow et al., 2015] Goodfellow, I. J., Shlens, J., and Szegedy, C. (2015). Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*.

[He et al., 2016] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

[Jia et al., 2014] Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., and Darrell, T. (2014). Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22Nd ACM International Conference on Multimedia (MM)*, pages 675–678.

[Krizhevsky et al., 2012] Krizhevsky, A., Sutskever, I., and Hinton, G. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS) 25*, pages 1106–1114.

[Lanckriet et al., 2003] Lanckriet, G., Ghaoui, L., Bhattacharyya, C., and Jordan, M. (2003). A robust minimax approach to classification. *The Journal of Machine Learning Research*, 3:555–582.

[LeCun et al., 1998] LeCun, Y., Cortes, C., and Burges, C. J. (1998). The MNIST database of handwritten digits. *http://yann.lecun.com/exdb/mnist/*.

[Moosavi-Dezfooli et al., 2016] Moosavi-Dezfooli, S.-M., Fawzi, A., and Frossard, P. (2016). Deepfool: a simple and accurate method to fool deep neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2574–2582.

[Paola and Schowengerdt, 1995] Paola, J. D. and Schowengerdt, R. A. (1995). The effect of lossy image compression on image classification. In *Geoscience and Remote Sensing Symposium (IGARSS)*, volume 1, pages 118–120. IEEE.

[Russakovsky et al., 2015] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A., and Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252.

[Sabour et al., 2016] Sabour, S., Cao, Y., Faghri, F., and Fleet, D. J. (2016). Adversarial manipulation of deep representations. In *International Conference on Learning Representations (ICLR)*.

[Sharif et al., 2016] Sharif, M., Bhagavatula, S., Bauer, L., and Reiter, M. K. (2016). Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 1528–1540.

[Simonyan and Zisserman, 2014] Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*.

[Szegedy et al., 2014] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. (2014). Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*.

[Tabacof and Valle, 2016] Tabacof, P. and Valle, E. (2016). Exploring the space of adversarial images. In *2016 IEEE International Joint Conference on Neural Networks (IJCNN)*, pages 426–433.

[Tanay and Griffin, 2016] Tanay, T. and Griffin, L. (2016). A boundary tilting persepective on the phenomenon of adversarial examples. *arXiv preprint arXiv:1608.07690*.

[Wang et al., 2004] Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on image processing*, 13(4):600–612.

[Warde-Farley et al., 2016] Warde-Farley, D., Goodfellow, I., Hazan, T., Papandreou, G., and Tarlow, D. (2016). Adversarial perturbations of deep neural networks. In Hazan, T., Papandreou, G., and Tarlow, D., editors, *Perturbations, Optimization, and Statistics*, chapter 11. MIT Press.

[Xu et al., 2009] Xu, H., Caramanis, C., and Mannor, S. (2009). Robustness and regularization of support vector machines. *The Journal of Machine Learning Research*, 10:1485–1510.