Robustness of classifiers to uniform $\ell_p$ and Gaussian noise

Jean-Yves Franceschi, Alhussein Fawzi, Omar Fawzi

To cite this version:
Jean-Yves Franceschi, Alhussein Fawzi, Omar Fawzi. Robustness of classifiers to uniform $\ell_p$ and Gaussian noise. Twenty-first International Conference on Artificial Intelligence and Statistics, Apr 2018, Playa Blanca, Spain. pp.1280-1288. hal-01715012

HAL Id: hal-01715012
https://hal.archives-ouvertes.fr/hal-01715012

Submitted on 22 Feb 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License
Robustness of classifiers to uniform $\ell_p$ 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 $\ell_p$ 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 $\ell_p$ 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 $\Sigma$. 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.

*Now at DeepMind.
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 [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 \( \ell_2 \) 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 \( f : \mathbb{R}^d \rightarrow \mathbb{R}^L \) be a \( L \)-class classifier. The estimated label of a datapoint \( x \in \mathbb{R}^d \) is set to \( g(x) = \arg\max_k f_k(x) \), where \( f_k(x) \) denotes the \( k \)th component of \( f(x) \). 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 \( \nu \) on \( \mathbb{R}^d \) that we interpret as giving the direction \( v \) of the noise, and we measure the length of the minimal scaling applied to \( v \) required to change the estimated label \( f \) at \( x \) with probability at least \( \varepsilon \). More precisely, let \( v \) be a random variable distributed according to \( \nu \); for a given \( \varepsilon > 0 \), we define \( r_{\nu,\varepsilon}(x) \) as:

\[
r_{\nu,\varepsilon}(x) = \min_{\alpha} \{ |\alpha| \; \text{s.t.} \; \mathbb{P}\{g(x + \alpha v) \neq g(x)\} \geq \varepsilon \}.
\]

If the set is empty, we set \( r_{\nu,\varepsilon}(x) = +\infty \)

In this paper, we will focus on two families of choices for \( \nu \).

The first family is parameterized by a real number \( p \in [1, \infty] \). The distribution \( \nu \) is then the uniform distribution over the unit ball of \( \ell_p^d \), i.e., \( B_p = \{ x \in \mathbb{R}^d : \|x\|_p \leq 1 \} \) where \( \|x\|_p = (\sum_{i=1}^d x_i^p)^{1/p} \). For this setting of distribution \( \nu \), we write \( r_{p,\varepsilon}(x) = r_{\nu,\varepsilon}(x) \). Observe that the Euclidean norm \( \|\cdot\|_2 \) is invariant under an orthonormal basis change, but this is not

\[^1\]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.
the case for $\| \cdot \|_p$ when $p \neq 2$, i.e., it depends on the basis that is chosen to write the signal $x$; hence, this dependence also holds for $r_{p,\epsilon}(x)$. 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 $\text{Tr}(\Sigma) = 1$ to fix the scale. The distribution $\nu$ is then the multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\Sigma$. We use the notation $r_{\Sigma,\epsilon}(x) = r_{\nu,\epsilon}(x)$ for this setting. A special case of this family is therefore the additive white Gaussian noise (where $\Sigma = \frac{1}{d} I$); note however that this family is much broader and can model a noise that is correlated with the input $x$, as no assumption is made on $\Sigma$.

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:

$$
\begin{align*}
    r^*_p(x) &= \arg\min_{r} \{ \| r \|_p \text{ s.t. } g(x + r) \neq g(x) \}. 
\end{align*}
$$

In other words, $\| r^*_p(x) \|_p$ quantifies the length of the minimal perturbation required to change the estimated label of the classifier, or equivalently, the distance from the data point $x$ to the decision boundary of the classifier. $r^*_p(x)$ 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 $r_{p,\epsilon}(x)$, it heavily depends on the choice of norm $\ell_p$, 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 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 $r^*_p(x)$ on the norm $\| \cdot \|_2$, 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 multi-class 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

$$
    f : x \mapsto w^T x + b.
$$

In this case\footnote{In the general multi-class setting, this corresponds to $f_0 = f$ and $f_1 = 0$.} $g(x) = 1$ if and only if $f(x) > 0$.

3.1 Uniform $\ell_p$ 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 $\ell_p$.

\begin{figure}
\centering
\includegraphics[width=0.8\textwidth]{figure1.png}
\caption{Illustration of the different noise regimes.}
\end{figure}
Figure 1: Illustration of noise with different values of $p$. First row: histogram of uniformly sampled noise from the unit ball of $\ell_p^d$. Second row: Example of noise image. Third row: Example of noisy image. Note that different values of $p$ result in perceptually different noise images.

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].
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 \( \varepsilon < \varepsilon_0 \):

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

We can take \( \zeta_1(\varepsilon) = C\sqrt{\varepsilon} \) and \( \zeta_2(\varepsilon) = \sqrt{\frac{1}{c-\sqrt{\varepsilon}}} \), 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 \( r_{p,\varepsilon}(x) \) is well estimated by \( \|r_{p,\varepsilon}(x)\|_p \) times a multiplicative factor that is independent of \( x \) and is of the order \( d^{1/p} \frac{\|w\|_{p'}}{\|w\|_2} \). The special case \( p = 2 \), for which this multiplicative factor becomes \( \sqrt{d} \), 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 \( w \). 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 \( w \) chosen by the classifier. For example, when \( w = (1 0 \ldots 0)^T \), 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 \( r_{\infty}(x) = \|r^*_{\infty}(x)\|_{\infty} \).

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

Proposition 1. For any \( p \in (1, \infty) \), if \( w \) is a random direction uniformly distributed over the unit \( \ell_2 \)-sphere, then, as \( d \to \infty \),

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

Moreover, for \( p = 1 \),

\[
\frac{\|w\|_{\infty}}{\sqrt{2d \ln d}} \overset{a.s.}{\longrightarrow} 1.
\]

While this result is only asymptotic and valid for random decision hyperplanes, we experimentally show in Section \( 7 \) 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 \( \Sigma \), we can similarly characterize the ratio \( \frac{r_{\Sigma,\varepsilon}(x)}{\|r^*_{\Sigma,\varepsilon}(x)\|_2} \) as a function of \( \Sigma \) and \( w \) as follows.

Theorem 2. Let \( \Sigma \) be a \( d \times d \) positive semidefinite matrix with \( \text{Tr}(\Sigma) = 1 \). There exist constants \( \varepsilon_0, \zeta_1(\varepsilon), \zeta_2(\varepsilon) \) such that, for all \( \varepsilon < \varepsilon_0 \):

\[
\zeta_1(\varepsilon) \frac{\|w\|_2}{\sqrt{\Sigma w}} \leq \frac{r_{\Sigma,\varepsilon}(x)}{\|r_{\Sigma,\varepsilon}(x)\|_2} \leq \zeta_2(\varepsilon) \frac{\|w\|_2}{\sqrt{\Sigma w}}.
\]

We can take \( \varepsilon_0 = 1/3 \), \( \zeta_1(\varepsilon) = \sqrt{\frac{1}{2 \ln(\frac{1}{\varepsilon})}} \) and \( \zeta_2(\varepsilon) = \sqrt{\frac{1}{1-\sqrt{3}\varepsilon}} \).

\(^3\)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 \( C' \).

\(^4\)Note that the condition \( \text{Tr}(\Sigma) = 1 \) is not needed for the statement but its motivation is to fix the scale of \( r_{\Sigma,\varepsilon}(x) \).
In this case, the multiplicative factor between robustnesses to random and adversarial perturbations is of the order $\frac{\|w\|_2}{\sqrt{\Sigma w w}}$. 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 $w$, this factor will be determined by a convex combination of eigenvalues of $\Sigma$. More precisely, if $\mathbf{u}_1, \ldots, \mathbf{u}_d$ are the eigenvectors of $\Sigma$ with eigenvalues $\lambda_1, \ldots, \lambda_d^2$ and assuming $\|w\|_2 = 1$ (without loss of generality), the factor is given by a weighted average of the eigenvalues $\lambda_1, \ldots, \lambda_d^2$:

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

In particular, if $\Sigma = \frac{1}{d} I_d$, then the factor is $\sqrt{d}$. Even more generally, for typical choices of $w$ we expect $|\mathbf{u}_i^T w|^2$ to 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 $r_{\eta,p}(\mathbf{x})$ and $\|r_{\eta,p}(\mathbf{x})\|_p$ 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 $S = \{x \in \mathbb{R}^d : f(x) = 0\}$. For $x^* \in S$, define $T(x^*)$ to be the hyperplane tangent to $S$ at point $x^*$. For $x \in \mathbb{R}^d$ and $x^* \in S$, we define $\mathcal{H}^-_x(x, x^*)$ to be the halfspace of points that are on the side of $x$ of the hyperplane parallel to $T(x^*)$ that passes though the point $x + (1 - \gamma)x^*$. Similarly, $\mathcal{H}^+_x(x, x^*)$ is the halfspace of points that are not on the side of $x$ of the hyperplane parallel to $T(x^*)$ that passes though the point $x^* + \gamma(x^* - x)$ (see Figure 3).

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

The LAF model assumes that the decision boundary can be locally approximated by a hyperplane, in the vicinity of images $x$ 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
Figure 3: Illustration of a \((\gamma, \eta)\)-LAF classifier with \(p = \frac{3}{2}\). \(S\) is the decision boundary of \(f\), separating instances of the same predicted class as \(x\) (on the side of \(x\)) from instances whose classification differs from \(x\) (on the other side). By definition, all inputs in \(H_\gamma^+ (x, x^*) \cap B_p (x, \eta)\) (red area, other side of \(S\)) are classified differently from \(x\), while the inputs in \(H_\gamma^- (x, x^*) \cap B_p (x, \eta)\) (green area, same side of \(S\) as \(x\)) are classified as \(x\).

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 \(H_\gamma^+ (x, x^*) \cap B_p (x, \eta)\) (respectively, \(H_\gamma^- (x, x^*) \cap B_p (x, \eta)\)) is classified differently from \(x\) (respectively, similarly to \(x\)); 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 \(w\) replaced by the gradient of \(f\) at the point \(x^*\) of the boundary that is closest to \(x\). It should be noted that for nonlinear classifiers, the gradient \(\nabla f (x^*)\) plays the same role as \(w\) for linear classifiers, as it is normal to the tangent to the decision boundary at \(x^*\).

**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 \((\gamma, \eta)\)-LAF at point \(x\) and \(x^*\) be such that \(r_p^*(x) = x^* - x\). Then:

\[
(1 - \gamma)\zeta_1 (\varepsilon) d^{1/p} \frac{\|\nabla f (x^*)\|_{p'}}{\|\nabla f (x^*)\|_2} \leq \frac{r_{p, \varepsilon} (x)}{r_p^*(x)} \leq (1 + \gamma)\zeta_2 (\varepsilon) d^{1/p} \frac{\|\nabla f (x^*)\|_{p'}}{\|\nabla f (x^*)\|_2},
\]
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 \( \mathbf{x} \) for each cross-section). The CaffeNet architecture [Jia et al., 2014] trained on ImageNet [Russakovsky et al., 2015] was used.

provided

\[ \eta \geq (1 + \gamma) \zeta_2(\varepsilon) d^{1/p} \frac{\|\nabla f(\mathbf{x}^*)\|_p'}{\|\nabla f(\mathbf{x}^*)\|_2} \left\| \mathbf{r}^*_p(\mathbf{x}) \right\|_p'. \]

In the case where \( \nabla f(\mathbf{x}^*) \) is uncorrelated with the basis used to write the signal (which we model by taking for \( \nabla f(\mathbf{x}^*) \) a random direction in the \( \ell_2 \) sphere), we obtain the same result as in Proposition 1 (i.e., by replacing \( \mathbf{w} \) with \( \nabla f(\mathbf{x}^*) \)). This provides bounds on the robustness to random noise that only depend on \( \|\mathbf{r}^*_p\|_2 \) 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 \( \Sigma \) be a \( d \times d \) positive semidefinite matrix with \( \text{Tr}(\Sigma) = 1 \). 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 \((\gamma, \eta)\)-LAF at point \( \mathbf{x} \) and \( \mathbf{x}^* \) be such that \( \mathbf{r}^*_2(\mathbf{x}) = \mathbf{x}^* - \mathbf{x} \). Then:

\[
(1 - \gamma) \zeta_1'\left(\frac{\varepsilon}{2}\right) \frac{\|\nabla f(\mathbf{x}^*)\|_2}{\|\nabla f(\mathbf{x}^*)\|_2} \leq \frac{r_{\Sigma, 2}(\mathbf{x})}{\|r_2^*(\mathbf{x})\|_2} \leq (1 + \gamma) \zeta_2'\left(\frac{3\varepsilon}{2}\right) \frac{\|\nabla f(\mathbf{x}^*)\|_2}{\|\nabla f(\mathbf{x}^*)\|_2},
\]

provided, using \( \psi(\varepsilon) = 8 \text{Tr}(\Sigma^2) \ln \frac{4}{\varepsilon} \),

\[ \eta \geq (1 + \gamma)(1 + \psi(\varepsilon)) \zeta_2'\left(\frac{3\varepsilon}{2}\right) \frac{\|\nabla f(\mathbf{x}^*)\|_2}{\|\nabla f(\mathbf{x}^*)\|_2} \left\| \mathbf{r}^*_p(\mathbf{x}) \right\|_2. \]

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 \( \varepsilon \), our estimate is:

\[
\delta_{p, \varepsilon}(\mathbf{x}) = \zeta_0 \sqrt{d} \left( \frac{\Gamma\left(\frac{2p-1}{2(p-1)}\right)}{\sqrt{\pi}} \right)^{1 - \frac{1}{p}} \left\| \mathbf{r}^*_p(\mathbf{x}) \right\|_p,
\]

\[ (4) \]
where $\zeta_0$ is a constant. The empirical robustness of Eq. (1) is specifically computed through an exhaustive search of smallest radius of the $\ell_p$ ball leading to an $\varepsilon$ fraction of misclassified samples. Note moreover that for this linear classifier, the worst-case robustness $\|r^*_p\|_p$ 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 $\|r^*_p\|_p$ 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 $p$. 

![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 $\varepsilon$ 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.](image-url)
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 $r_{p,\varepsilon}(x)$. 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.

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

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

using the estimate of Eq. (4). Moreover, the number of quantization levels required to guarantee robustness of the classifier is therefore estimated by

$$L_q = \frac{255}{\frac{2\zeta_0}{\sqrt{\pi}} \sqrt{\|r_{\infty}^*(x)\|_{\infty}}}. \quad (5)$$

In other words, Eq. (5) predicts that images encoded with more than $\log_2 \left( \frac{255}{\frac{2\zeta_0}{\sqrt{\pi}} \sqrt{\|r_{\infty}^*(x)\|_{\infty}}} \right)$ 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 $\mathcal{N}(0, \Sigma(x))$, where $\Sigma(x)$ is a diagonal matrix such that $\Sigma(x)_{ii} = 1_{x_i \geq t} \cdot x_i$, where $x_i$ denotes the value of pixel $i$, and $t$ denotes a user-specified threshold. $\Sigma(x)$ is further normalized to satisfy $\text{Tr}(\Sigma(x)) = 1$. 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.

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 $\Sigma(x)$, and should moreover satisfy:

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

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.

We stress here that, due to the normalization $\text{Tr}(\Sigma(x)) = 1$, 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.
where $\varepsilon = 0.15$. To verify this hypothesis, we show in Figure 9 the ratio $r_{\Sigma,\varepsilon}(x) \|r(x)\|_2$ 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., $W(x) = \sum_i 1_{x_i\geq t} x_i$. Similarly to previous experiments, $r_{\Sigma,\varepsilon}(x)$ is estimated using an exhaustive line search. It can be seen that the ratio $r_{\Sigma,\varepsilon}(x) \|r(x)\|_2$ 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 $\Sigma$) 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.

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

References


Proof. We distinguish between the three cases.

• if \( p = \infty \):
  \[
  \|r^{*}_{\infty}(x)\|_{\infty} = \frac{|f(x)|}{\|w\|_1};
  \]

• if \( p = 1 \):
  \[
  \|r^{*}_{1}(x)\|_{1} = \frac{|f(x)|}{\|w\|_{\infty}};
  \]

• if \( p \in (1, \infty) \):
  \[
  \|r^{*}_{p}(x)\|_{p} = \frac{|f(x)|}{\|w\|_{p}}.
  \]

Overall, for all \( p \in [1, \infty] \), the \( \ell_{p} \)-distance from any point \( x \) to the decision hyperplane \( H \) defined by \( f(z) = 0 \) is:

\[
\|r^{*}_{p}(x)\|_{p} = \frac{|f(x)|}{\|w\|_{p}}.
\]

Proof. We distinguish between the three cases.

• Suppose \( p = \infty \). The distance from \( x \) to \( H \) is equal to the minimum radius \( \alpha \) of a ball (i.e., for \( \ell_{\infty} \), a hypercube) centered at \( x \) that intersects \( H \). This intersection with minimum radius necessarily contains a vertex of the hypercube. To determine which one, it suffices to determine which vector \( x + \alpha \epsilon \), with \( \epsilon \in \{-1, 1\}^{d} \), first intersects \( H \) when \( \alpha \) increases starting at 0. Such an intersection arises when \( w^{T}(x + \alpha \epsilon) + b = 0 \), so \( \alpha = \frac{f(x)}{w^{T} \epsilon} \), and since \( \alpha \) must be non-negative:
  \[
  r^{*}_{\infty}(x) = \min_{f(x) \leq 0} \frac{f(x)}{w^{T} \epsilon} = \frac{|f(x)|}{\|w\|_1},
  \]
  because \( \epsilon \in \{-1, 1\}^{d} \) (simply choose \( \epsilon_i = \text{sign}(-f(x)w_i) \)).

• Suppose \( p = 1 \). In this case, the proof is symmetric to the one for \( p = \infty \), with \( \epsilon \in \{-1, 1\}^{d} \) having exactly one non-zero coordinate.

• Suppose \( p \in (1, \infty) \). The distance from \( x \) to \( H \) is equal to the minimum radius \( \alpha \) of an \( \ell_{p} \) ball \( B_{p} \) centered at \( x \) that intersects \( H \). This ball is described by the following equation (where \( z_i \) is the variable):
  \[
  \sum_{i=1}^{d} (z_i - x_i)^p \leq \alpha^p.
  \]
  For such a minimum radius, the plane described by \( f(z) = 0 \) is tangent to \( B_{p} \) at some point \( x + n \). Let us assume without loss of generality that every coordinate of \( n \) is non-negative. We also know that this hyperplane is described by the following equations (where \( z_i \) is the variable):
  \[
  \nabla_{z+n} \left( \sum_{i=1}^{d} (z_i - x_i)^p - \alpha^p \right)^T (z - (x + n)) = 0 \iff \sum_{i=1}^{d} n_i^{p-1} (z_i - x_i - n_i) = 0
  \]
  \[
  \iff \sum_{i=1}^{d} n_i^{p-1} (z_i - x_i) = \alpha^p,
  \]
Lemma 2. There exists a universal constant

Proof. Let us first express conveniently

B uniformly at random in

Theorem 1. Let

\[ \ell \]

\[ B \]

Robustness of Linear Classifiers to \( \ell_p \) Noise

B.1 Main Theorem

**Theorem 1.** Let \( p \in [1, \infty] \). Let \( p' \in [1, \infty] \) be such that \( \frac{1}{2} + \frac{1}{p'} = 1 \). Then there exist universal constants \( C, c, c' > 0 \) such that, for all \( \varepsilon < \frac{2}{\sqrt{d}} \):

\[
\zeta_1(\varepsilon) d^{1/p} \frac{\|\mathbf{w}\|_{p'}}{\|\mathbf{w}\|_2} \leq \frac{r_{p,\varepsilon}(\mathbf{x})}{\|r_{p,\varepsilon}^*(\mathbf{x})\|_p} \leq \zeta_2(\varepsilon) d^{1/p} \frac{\|\mathbf{w}\|_{p'}}{\|\mathbf{w}\|_2},
\]

where \( \zeta_1(\varepsilon) = C\sqrt{\varepsilon} \) and \( \zeta_2(\varepsilon) = \frac{1}{\sqrt{c-\sqrt{\varepsilon}}} \).

Theorem 1 is proved by the following lemmas.

**Lemma 2.** There exists a universal constant \( C > 0 \) such that

\[
\frac{r_{p,\varepsilon}(\mathbf{x})}{\|r_{p,\varepsilon}^*(\mathbf{x})\|_p} \geq \zeta_1(\varepsilon) d^{1/p} \frac{\|\mathbf{w}\|_{p'}}{\|\mathbf{w}\|_2},
\]

where \( \zeta_1(\varepsilon) = C\sqrt{\varepsilon} \).

**Proof.** Let us first express conveniently \( \mathbb{P}_{\mathbf{v} \sim B_p} \{ g(\mathbf{x}) \neq g(\mathbf{x} + \alpha \mathbf{v}) \} \), where \( \mathbf{v} \sim B_p \) means that \( \mathbf{v} \) is chosen uniformly at random in \( B_p \):

\[
\begin{align*}
\mathbb{P}_{\mathbf{v} \sim B_p} \{ g(\mathbf{x}) \neq g(\mathbf{x} + \alpha \mathbf{v}) \} &= \mathbb{P}_{\mathbf{v} \sim B_p} \{ f(\mathbf{x}) f(\mathbf{x} + \alpha \mathbf{v}) \leq 0 \} \\
&= \mathbb{P}_{\mathbf{v} \sim B_p} \left\{ \text{sign}(f(\mathbf{x})) (\mathbf{w}^T \mathbf{x} + \mathbf{b}) \leq -\text{sign}(f(\mathbf{x})) \alpha \mathbf{w}^T \mathbf{v} \right\} \\
&= \mathbb{P}_{\mathbf{v} \sim B_p} \left\{ \|\mathbf{w}\|_{p'} \|r_{p,\varepsilon}^*(\mathbf{x})\|_p \leq -\text{sign}(f(\mathbf{x})) \alpha \mathbf{w}^T \mathbf{v} \right\} \\
&= \mathbb{P}_{\mathbf{v} \sim B_p} \left\{ \|\mathbf{w}\|_{p'} \|r_{p,\varepsilon}^*(\mathbf{x})\|_p \leq \frac{\mathbf{w}^T \mathbf{v}}{\|\mathbf{w}\|_{p'}} \right\} \\
&= \frac{1}{2} \mathbb{P}_{\mathbf{v} \sim B_p} \left\{ \|\mathbf{w}\|_{p'} \|r_{p,\varepsilon}^*(\mathbf{x})\|_p \leq \frac{\mathbf{w}^T \mathbf{v}}{\|\mathbf{w}\|_{p'}} \right\}.
\end{align*}
\]

where Eq. (7) is given by Lemma 1 and Eq. (8) and (9) follow from \( \mathbf{v} \sim B_p \Rightarrow -\mathbf{v} \sim B_p \).
Markov’s inequality gives, from Eq. (2):

\[ \mathbb{P}_{v \sim B_p} \{ g(x) \neq g(x + \alpha v) \} \leq \frac{1}{2} \frac{\mathbb{E}_{v \sim B_p} \left[ \left( \sum_{i=1}^{d} w_i v_i \right)^2 \right]}{\left( \|w\|_p \right)^2} . \]

In [Barthe et al., 2005] Theorem 7, it is proved that there is a constant \( C_0 > 0 \) such that:

\[ \mathbb{E}_{v \sim B_p} \left[ \left( \sum_{i=1}^{d} w_i v_i \right)^2 \right] \leq \left( \frac{2C_0}{d^2} \|w\|_2^2 \right)^2 , \]

Therefore:

\[ \mathbb{P}_{v \sim B_p} \{ g(x) \neq g(x + \alpha v) \} \leq \frac{1}{2} \left( \frac{2C_0}{d^2} \|w\|_2^2 \right)^2 . \]

So, if \( |\alpha| < \sqrt{\frac{d^2}{2C_0}} \|w\|_2^2 \|r^*_p(x)\|_p \), then \( \mathbb{P}_{v \sim B_p} \{ g(x) \neq g(x + \alpha v) \} < \epsilon \). Thus, there is a universal constant \( C = \frac{1}{\sqrt{2C_0}} > 0 \) such that:

\[ \frac{r_p(x)}{r^*_p(x)} \geq \zeta_1(\epsilon) d^{1/p} \|w\|_p \|w\|_2 . \]

\[ \zeta_1(\epsilon) = \frac{1}{\sqrt{\epsilon - \sqrt{\epsilon}}} . \]

\[ \zeta_2(\epsilon) = \frac{1}{\sqrt{\epsilon - \sqrt{\epsilon}}} . \]

**Lemma 3.** There exist universal constants \( c, c' > 0 \) such that, for all \( \epsilon < \frac{c'}{\sqrt{\epsilon}} \):

\[ \frac{r_p(x)}{r^*_p(x)} \leq \zeta_2(\epsilon) d^{1/p} \|w\|_2 . \]

where \( \zeta_2(\epsilon) = \frac{1}{\sqrt{\epsilon - \sqrt{\epsilon}}} . \)

**Proof.** We first transform the expression of \( \mathbb{P}_{v \sim B_p} \{ g(x) \neq g(x + \alpha v) \} \):

\[ \mathbb{P}_{v \sim B_p} \{ g(x) \neq g(x + \alpha v) \} = \mathbb{P}_{v \sim B_p} \left\{ \|w\|_p \left\| \frac{r^*_p(x)}{|\alpha|} \right\|_p \leq w^Tv \right\} \]

\[ = \frac{1}{2} \mathbb{P}_{v \sim B_p} \left\{ \|w\|_p \left\| \frac{r^*_p(x)}{|\alpha|} \right\|_p \leq w^Tv \right\} \]

\[ = \frac{1}{2} \mathbb{P}_{v \sim B_p} \left\{ \frac{1}{\text{Var}(w^Tv)} \left( \|w\|_p \left\| \frac{r^*_p(x)}{|\alpha|} \right\|_p \right)^2 \leq \frac{(w^Tv)^2}{\text{Var}(w^Tv)} \right\} . \]

Paley-Zygmund’s inequality states that, if \( X \) is a random variable with finite variance and \( t \in [0,1] \), then:

\[ \mathbb{P} \left\{ X > t \mathbb{E}[X] \geq (1-t)^2 \frac{\mathbb{E}[X]^2}{\mathbb{E}[X^2]} \right\} . \]

Note that \( \mathbb{E}_{v \sim B_p} \left( \frac{(w^Tv)^2}{\text{Var}(w^Tv)} \right) = 1 \), because \( \mathbb{E}_{v \sim B_p} (w^Tv) = 0 \). So, by using Paley-Zygmund’s inequality with \( X = \frac{(w^Tv)^2}{\text{Var}(w^Tv)} \) and \( t = \frac{1}{\text{Var}(w^Tv)} \left( \|w\|_p \left\| \frac{r^*_p(x)}{|\alpha|} \right\|_p \right)^2 \), when \( |\alpha| \geq \frac{\|w\|_p \sqrt{\text{Var}(w^Tv)}}{\sqrt{\text{Var}(w^Tv)}} \|r^*_p(x)\|_p \):

\[ \mathbb{P}_{v \sim B_p} \{ g(x) \neq g(x + \alpha v) \} \geq \frac{1 - \frac{1}{\text{Var}(w^Tv)} \left( \|w\|_p \left\| \frac{r^*_p(x)}{|\alpha|} \right\|_p \right)^2}{2\mathbb{E}_{v \sim B_p} \left( \frac{(w^Tv)^4}{\text{Var}(w^Tv)^2} \right)^{1/2}} . \]

So, if \( |\alpha| > \sqrt{\frac{1}{\text{Var}(w^Tv)-(2\epsilon \mathbb{E}(w^Tv))}\|w\|_p \left\| r^*_p(x) \right\|_p} \), then \( \mathbb{P}_{v \sim B_p} \{ g(x) \neq g(x + \alpha v) \} > \epsilon \). According to [Barthe et al., 2005] Theorem 7, there is a universal constant \( c_0 > 0 \) such that:

\[ \]
• for \( \text{Var} (w^T v) \):
\[
\text{Var} (w^T v) \geq \left( \frac{c_0}{d^{rac{1}{p}}} \|w\|_2 \right)^2 ;
\]

• for \( \mathbb{E} \left[ (w^T v)^4 \right] \):
\[
\mathbb{E} \left[ (w^T v)^4 \right] \leq \left( \frac{4C_0}{d^{rac{1}{p}}} \|w\|_2 \right)^4 .
\]

So there are universal constants \( c = c_0^2, c' = 512C_0^4 > 0 \) such that:
\[
\frac{r_{p,\varepsilon}(x)}{\|r_p^*(x)\|_p} \leq \zeta_2(\varepsilon)d^{1/p} \frac{\|w\|_{\rho'}}{\|w\|_2} .
\]

\[\square\]

### B.2 Alternative Lower Bound

Actually, the lower bound of Theorem 1 may be improved for most \( p \)-norms by the following result.

**Lemma 4.** There exists a universal constant \( C' > 0 \) such that
\[
\frac{r_{p,\varepsilon}(x)}{\|r_p^*(x)\|_p} \geq \zeta_1(\varepsilon)d^{1/p} \frac{\|w\|_{\rho'}}{\|w\|_2} ,
\]
where \( \zeta_1(\varepsilon) = \frac{C'}{\sqrt{\log 2}} \left( 1 - \frac{1}{\min(p,2)} \right) \).

**Proof.** Let \( p_2 = \min(p,2) \). We have:
\[
P_{v \sim B_p} \{ g(x) \neq g(x + \alpha v) \} = P_{v \sim B_p} \left\{ \|w\| \frac{|r_p^*(x)|_p}{\alpha} \leq w^T v \right\} = P_{v \sim B_p} \{ e^{\theta t} \leq \text{exp} (\theta w^T v) \} ,
\]
where \( t = \|w\|_{\rho'} \frac{|r_p^*(x)|_p}{|\alpha|} \), for any \( \theta > 0 \). Markov’s inequality gives:
\[
P_{v \sim B_p} \{ g(x) \neq g(x + \alpha v) \} \leq \frac{1}{e^{\theta t}} \mathbb{E}_{v \sim B_p} \left[ \text{exp} (\theta w^T v) \right] = \frac{1}{e^{\theta t}} \sum_{k=0}^{\infty} \frac{1}{k!} \mathbb{E}_{v \sim B_p} \left[ (\theta w^T v)^k \right] ,
\]
since \( w^T v \) is symmetric. In [Barthe et al., 2005, Theorem 7], it is proved that:

- if \( k \leq d \) and \( p \leq 2 \):
\[
\mathbb{E}_{v \sim B_p} \left[ \sum_{i=1}^{d} w_i v_i \right]^k \leq \left( \frac{C_0 k^\frac{1}{p}}{d^\frac{1}{p}} \|w\|_2 \right)^k ;
\]

- if \( k \leq d \) and \( p > 2 \):
\[
\mathbb{E}_{v \sim B_p} \left[ \sum_{i=1}^{d} w_i v_i \right]^k \leq \left( \frac{C_0 k^\frac{1}{p}}{d^\frac{1}{p}} \|w\|_2 \right)^k ;
\]

- if \( k > d \) and \( p \leq 2 \):
\[
\mathbb{E}_{v \sim B_p} \left[ \sum_{i=1}^{d} w_i v_i \right]^k \leq (C_0 \|w\|_2)^k \leq \left( \frac{C_0 k^\frac{1}{p}}{d^\frac{1}{p}} \|w\|_2 \right)^k ;
\]

18
if $k > d$ and $p > 2$:

$$E_{v \sim B_p} \left[ \left\| \sum_{i=1}^{d} w_i v_i \right\|^k \right] \leq \left( \frac{C_0 d^2}{d^2} \right)^k \left\| w \right\|_2^k \leq \left( \frac{C_0 k^2}{d^2} \right)^k \left\| w \right\|_2^k,$$

where $C_0$ is a universal constant (the same as in the proof of Lemma 2). So, overall:

$$E_{v \sim B_p} \left[ \left\| \sum_{i=1}^{d} w_i v_i \right\|^k \right] \leq \left( \frac{C_0 k^2}{d^2} \right)^k \left\| w \right\|_2^k.$$

Thus:

$$P_{v \sim B_p} \{ g(x) \neq g(x + \alpha v) \} \leq \frac{1}{e^{\alpha t}} \sum_{k=0}^{\infty} \left( \frac{C_0 (2k)^{\frac{1}{2}}}{d^2} \right)^k \left\| w \right\|_2^k.$$

We can bound the following power series using Stirling-like bounds [Robbins, 1955] in (10) and (11):

$$\sum_{k=0}^{\infty} \frac{(2k)^{\frac{1}{2}}}{(2\pi)^k} x^k \leq 1 + \frac{1}{\sqrt{2\pi}} \sum_{k=1}^{\infty} \frac{(2k)^{\frac{1}{2}}}{(2\pi)^k} \left( e^2 x \right)^k \leq 1 + \frac{1}{\sqrt{2\pi}} \sum_{k=1}^{\infty} \left( \frac{1}{2} \left( 1 - \frac{1}{p^2} \right) \right)^k \left( e^2 x \right)^k \leq 1 + \frac{1}{\sqrt{2\pi}} \sum_{k=1}^{\infty} \exp \left( - \left( \frac{2}{1 - \frac{1}{p^2}} \right)^k \right) \left( e^2 x \right)^k \leq 1 + \frac{e^2}{\sqrt{2\pi}} \sum_{k=1}^{\infty} \frac{\exp \left( -2 \left( 1 - \frac{1}{p^2} \right) \right)}{k!} \left( e^2 x \right)^k.$$

Therefore:

$$P_{v \sim B_p} \{ g(x) \neq g(x + \alpha v) \} \leq 3e^{-\theta t} \exp \left( \theta^2 \frac{p_2}{2} \frac{e^2 C_0^2}{d^2} \left\| w \right\|_2^2 \right).$$

By choosing $\theta = \frac{1}{2} t \left( \frac{p_2}{p_2 - 1} \frac{e^2 C_0^2}{d^2} \right)^{-1}$:

$$P_{v \sim B_p} \{ g(x) \neq g(x + \alpha v) \} \leq 3 \exp \left( -t^2 \left( 1 - \frac{1}{p_2} \right) \frac{d^2}{2e^2 C_0^2 \left\| w \right\|_2^2} \right) = 3 \exp \left( - \left( \frac{\left\| r_p^*(x) \right\|_p}{\left\| \alpha \right\|} \right)^2 \left( 1 - \frac{1}{p_2} \right) \frac{d^2}{2e^2 C_0^2 \left\| w \right\|_2^2} \right).$$

So, if $|\alpha| < \frac{C}{\sqrt{\ln 2}} \left( 1 - \frac{1}{p_2} \right) d^2 \left\| w \right\|_2 \left\| r_p^*(x) \right\|_p$, then $P_{v \sim B_p} \{ g(x) \neq g(x + \alpha v) \} < \varepsilon$, where $C = \frac{1}{2e^2 C_0^2} > 0$ is a universal constant, and:

$$\frac{r_p^*(x)}{\left\| r_p^*(x) \right\|_p} \geq \frac{\zeta_1(\varepsilon)d^{1/p} \left\| w \right\|_2}{\left\| w \right\|_2}.$$
B.3 Typical Value of the Multiplicative Factor

**Proposition 1.** For any \( p \in (1, \infty] \), if \( w \) is a random direction uniformly distributed over the unit \( \ell_2 \)-sphere, then, as \( d \to \infty \):

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

Moreover, for \( p = 1 \),

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

**Proof.** \( w \) can be written as \( \frac{g}{\|g\|_2} \), where \( g = (g_1, \ldots, g_d) \) are i.i.d. with normal distribution \( \mu = 0, \sigma^2 = \frac{1}{2} \).

The law of large numbers gives that, for \( p' \neq \infty \):

\[
\frac{1}{d} \sum_{i=1}^d |g_i|^{p'} \xrightarrow{a.s.} \mathbb{E} \left( |g_1|^{p'} \right) = \frac{\Gamma \left( \frac{1+p'}{2} \right)}{\sqrt{\pi}}.
\]

Thus:

\[
\frac{1}{d^{1/p}} \|g\|_{p'} \xrightarrow{a.s.} \left( \frac{\Gamma \left( \frac{1+p'}{2} \right)}{\sqrt{\pi}} \right)^{\frac{1}{p}},
\]

and, for \( p \in (1, \infty] \):

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

because \( \frac{\|g\|_2}{\sqrt{d}} \xrightarrow{a.s.} \frac{1}{\sqrt{2}} \).

For \( p = 1 \) we use a result proved in [Galambos, 1987] Example 4.4.1] directly implying that

\[
\frac{\|g\|_{\infty}}{\sqrt{\ln d}} \xrightarrow{a.s.} 1.
\]

Using the previous computations for \( p = 2 \), we find:

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

\[ \square \]

C Robustness of Linear Classifiers to Gaussian Noise

C.1 Main Theorem

**Theorem 2.** For \( \varepsilon < \frac{1}{2}, \zeta'_1(\varepsilon) = \sqrt{\frac{1}{2 \ln \left( \frac{1}{\varepsilon} \right)}} \) and \( \zeta'_2(\varepsilon) = \sqrt{\frac{1}{1 - \sqrt{\varepsilon}}} \):

\[
\zeta'_1(\varepsilon) \|w\|_2 \leq \frac{r_{\Sigma, \varepsilon}(x)}{\|\sqrt{\Sigma}w\|_2} \leq \zeta'_2(\varepsilon) \frac{\|w\|_2}{\|\sqrt{\Sigma}w\|_2}.
\]

**Theorem 2** is proved by the following lemmas.

**Lemma 5.** For \( \zeta'_1(\varepsilon) = \sqrt{\frac{1}{2 \ln \left( \frac{1}{\varepsilon} \right)}} \),

\[
\frac{r_{\Sigma, \varepsilon}(x)}{\|r_\Sigma^2(x)\|_2} \geq \zeta'_1(\varepsilon) \frac{\|w\|_2}{\|\sqrt{\Sigma}w\|_2}.
\]
Proof. As in the proof of Lemma 2

\[ P_{v \sim \mathcal{N}(0, \Sigma)} \{ g(x + \alpha v) \neq g(x) \} = P_{v \sim \mathcal{N}(0, \Sigma)} \{ \| w \|_2 \| r_2^*(x) \|_2 \leq |\alpha| w^T v \}. \]

Since \( v \sim \mathcal{N}(0, \Sigma) \) follows a multivariate normal distribution with a positive definite covariance matrix \( \Sigma \), if \( \sqrt{\Sigma} \) is the (symmetric) square root of \( \Sigma \), then \( v = \sqrt{\Sigma} v' \) with \( v' \sim \mathcal{N}(0, I_d) \). So:

\[
P_{v \sim \mathcal{N}(0, \Sigma)} \{ g(x + \alpha v) \neq g(x) \} = \mathbb{P}_{v \sim \mathcal{N}(0, I_d)} \left\{ \| w \|_2 \| r_2^*(x) \|_2 \leq |\alpha| w^T \sqrt{\Sigma} v \right\} \]

\[
= \mathbb{P}_{v \sim \mathcal{N}(0, I_d)} \left\{ \| w \|_2 \| r_2^*(x) \|_2 \leq \left( |\alpha| \sqrt{\Sigma} w \right)^T v \right\}.
\]

If \( v \sim \mathcal{N}(0, I_d) \), then \( \left( |\alpha| \sqrt{\Sigma} w \right)^T v \sim \mathcal{N}(0, \alpha^2 \| \sqrt{\Sigma} w \|_2^2) \). Therefore:

\[
P_{v \sim \mathcal{N}(0, \Sigma)} \{ g(x + \alpha v) \neq g(x) \} \leq \exp \left( -\frac{1}{2} \left( \frac{\| w \|_2 \| r_2^*(x) \|_2}{|\alpha| \sqrt{\Sigma} w} \right)^2 \right).
\]

So, if \( |\alpha| < \sqrt{\frac{1}{2 \ln 2} \| \sqrt{\Sigma} w \|_2^2} \| r_2^*(x) \|_2 \), then \( P_{v \sim \mathcal{N}(0, \Sigma)} \{ g(x + \alpha v) \neq g(x) \} < \epsilon \). Thus,

\[
\frac{r_{\Sigma,\epsilon}(x)}{\| r_2^*(x) \|_2} \geq \zeta_\epsilon(\epsilon) \frac{\| w \|_2}{\| \sqrt{\Sigma} w \|_2}.
\]

Lemma 6. For \( \epsilon < \frac{1}{3} \) and \( \zeta_\epsilon(\epsilon) = \sqrt{\frac{1}{1 - \sqrt{\epsilon}}} \),

\[
\frac{r_{\Sigma,\epsilon}(x)}{\| r_2^*(x) \|_2} \leq \zeta_\epsilon(\epsilon) \frac{\| w \|_2}{\| \sqrt{\Sigma} w \|_2}.
\]

Proof.

\[
P_{v \sim \mathcal{N}(0, \Sigma)} \{ g(x + \alpha v) \neq g(x) \} = \mathbb{P}_{v \sim \mathcal{N}(0, I_d)} \left\{ \| w \|_2 \| r_2^*(x) \|_2 \leq \left( |\alpha| \sqrt{\Sigma} w \right)^T v \right\}
\]

\[
= \frac{1}{2} \mathbb{P}_{v \sim \mathcal{N}(0, I_d)} \left\{ \left( \frac{1}{\alpha \| \sqrt{\Sigma} w \|_2} \right) \| \sqrt{\Sigma} w \|_2 \| r_2^*(x) \|_2 \right\} \leq \frac{\left( \left( \frac{\| \sqrt{\Sigma} w \|_2}{\| \sqrt{\Sigma} w \|_2} \right)^2 \right)}{\frac{1}{\alpha \| \sqrt{\Sigma} w \|_2}} \frac{\| w \|_2}{\| \sqrt{\Sigma} w \|_2}.
\]

Note that \( \mathbb{E}_{v \sim \mathcal{N}(0, I_d)} \left( \left. \left( \frac{\| \sqrt{\Sigma} w \|_2}{\| \sqrt{\Sigma} w \|_2} \right)^2 \right| \| \sqrt{\Sigma} w \|_2 \right) = 1 \). So, by using Paley-Zygmund’s inequality, when \( |\alpha| \geq \frac{\| w \|_2}{\| \sqrt{\Sigma} w \|_2} \| r_2^*(x) \|_2 \):

\[
P_{v \sim \mathcal{N}(0, \Sigma)} \{ g(x + \alpha v) \neq g(x) \} \geq \frac{1 - \left( \frac{\| w \|_2}{\| \sqrt{\Sigma} w \|_2} \| r_2^*(x) \|_2 \right)^2}{2 \| \sqrt{\Sigma} w \|_2^4 + \| \sqrt{\Sigma} w \|_2^4} = \frac{1 - \left( \frac{\| w \|_2}{\| \sqrt{\Sigma} w \|_2} \| r_2^*(x) \|_2 \right)^2}{\frac{1}{3}},
\]

So, if \( |\alpha| > \frac{\| w \|_2}{\| \sqrt{\Sigma} w \|_2} \| r_2^*(x) \|_2 \), then \( P_{v \sim \mathcal{N}(0, \Sigma)} \{ g(x + \alpha v) \neq g(x) \} > \epsilon \). Therefore,

\[
\frac{r_{\Sigma,\epsilon}(x)}{\| r_2^*(x) \|_2} \leq \zeta_\epsilon(\epsilon) \frac{\| w \|_2}{\| \sqrt{\Sigma} w \|_2}.
\]

\[\square\]
C.2 Typical Value of the Multiplicative Factor

**Proposition 2.** Let \( \Sigma \) be a \( d \times d \) positive semidefinite matrix with \( \text{Tr}(\Sigma) = 1 \). If \( w \) is a random direction uniformly distributed over the unit \( \ell_2 \)-sphere, then, for \( t \leq \frac{\sqrt{2}\pi}{8} d \):

\[
\mathbb{P} \left\{ \left( \frac{||w||_2}{\sqrt{\Sigma w}} \right)^2 - d \geq t' \right\} \leq 2 \exp \left( -\frac{t^2}{8d} \right) + 2 \exp \left( -\frac{1}{200 \text{Tr}(\Sigma^2)} \right),
\]

where \( t' = \frac{t}{2} \).

**Proof.** Suppose that \( w \) is a random direction uniformly distributed over the unit \( \ell_2 \) sphere.

Then \( w \) can be written as \( \frac{g}{||g||_2} \), where \( g = (g_1, \ldots, g_d) \) are i.i.d. with normal distribution (\( \mu = 0, \sigma^2 = \frac{1}{2} \)). By using this representation in the orthogonal basis in which \( \sqrt{\Sigma} \) is diagonal, we get

\[
\frac{||g||_2}{||\sqrt{\Sigma}g||_2} = \sqrt{\frac{\sum_{i=1}^{d} g_i^2}{\sum_{i=1}^{d} (\lambda_i g_i)^2}},
\]

where \( \sqrt{\Sigma} = \text{Diag} ((\lambda_i)) \) in the previously mentioned orthogonal basis.

Let us focus on the concentration of \( \sum_{i=1}^{d} (\lambda_i g_i)^2 \). We have:

\[
\sum_{i=1}^{d} (\lambda_i g_i)^2 - \frac{1}{2} = \sum_{i=1}^{d} (\lambda_i g_i)^2 - \frac{1}{2} \sum_{i=1}^{d} \lambda_i^2 = \sum_{i=1}^{d} \lambda_i^2 \left( g_i^2 - \mathbb{E} (g_i^2) \right).
\]

One of Bernstein-type inequalities [Bernstein, 1927] can be applied:

\[
\mathbb{P} \left\{ \left| \sum_{i=1}^{d} \lambda_i^2 (g_i^2 - \mathbb{E} (g_i^2)) \right| \geq 2t \sqrt{\text{Var} \left( g_i^2 - \mathbb{E} (g_i^2) \right) \sum_{i=1}^{d} \lambda_i^2} \right\} \leq 2 \exp \left( -\frac{t^2}{2 \text{Tr}(\Sigma^2)} \right),
\]

for \( t \leq \beta \sqrt{\text{Tr}(\Sigma^2)} \) where \( \beta = \frac{\sqrt{2}}{8} \) is a constant\(^8\) i.e., for \( t \leq \frac{\beta}{2} \):

\[
\mathbb{P} \left\{ \left| \frac{1}{d} \frac{||g||_2^2 - 1}{2} \right| \right\} \leq 2 \exp \left( -\frac{dt^2}{2} \right).
\]

Using the fact that \( 2 ||g||_2^2 \) has a chi-squared distribution, so using a simple concentration inequality for the chi-squared distribution\(^9\):

\[
\mathbb{P} \left\{ \left| \frac{1}{d} \frac{||g||_2^2 - 1}{2} \right| \right\} \leq 2 \exp \left( -\frac{dt^2}{2} \right).
\]

Overall, for \( t \leq \beta d \) and \( t' = \frac{t}{2} \):

\[
\mathbb{P} \left\{ \left( \frac{||g||_2^2}{\sqrt{\Sigma g}} \right)^2 - d \geq t' \right\} = \mathbb{P} \left\{ \left( \frac{||g||_2^2 - d \frac{||\sqrt{\Sigma} g||_2^2}{||\sqrt{\Sigma} g||_2^2}}{||\sqrt{\Sigma} g||_2^2} \right)^2 \geq t' \right\} = \mathbb{P} \left\{ \left( \frac{||g||_2^2 - d \frac{||\sqrt{\Sigma} g||_2^2}{||\sqrt{\Sigma} g||_2^2}}{||\sqrt{\Sigma} g||_2^2} \right)^2 \geq t' \right\} \leq \mathbb{P} \left\{ \left( \frac{1}{d} \frac{||g||_2^2 - 1}{2} \geq \frac{t}{2d} \text{ or } \frac{1}{d} \frac{||\sqrt{\Sigma} g||_2^2 - 1}{2} \geq \frac{t}{2d} \right) \right\}.
\]

---

\(^8\)Because \( \frac{1}{d} \frac{1}{2} \geq \mathbb{E} \left( \frac{1}{d} \right) \leq \frac{1}{d} \mathbb{E} (g_i^2) \left( \frac{1}{2d} \right)^{k-2} k! \) for all \( k > 1 \).

\(^9\)Using the fact that \( 2 ||g||_2^2 \) is a sum of independent sub-exponential random variables (see [https://www.stat.berkeley.edu/~mjwain/stat210b/Chap2_TailBounds_Jan22_2015.pdf](https://www.stat.berkeley.edu/~mjwain/stat210b/Chap2_TailBounds_Jan22_2015.pdf) Example 2.5, for instance).
so, using the previous inequalities:

\[
P\left\{ \left( \frac{\|g\|_2}{\sqrt{\Sigma g}} \right)^2 - d \geq t' \right\} \leq 2 \exp \left( - \frac{t'^2}{8d} \right) + 2 \exp \left( - \frac{t^2}{8d^2 \text{Tr}(\Sigma^2)} \right) + 2 \exp \left( - \frac{1}{200 \text{Tr}(\Sigma^2)} \right).
\]

\[\square\]

## D Robustness of LAF Classifiers to \(\ell_p\) and Gaussian Noise

**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 \(\text{[7]}\). Then, for all \(\varepsilon <\varepsilon_0\), the following holds.

Assume \(f\) is a classifier that is \((\gamma, \eta)\)-LAF at point \(\mathbf{x}\) and \(\mathbf{x}^*\) be such that \(r_{p,\varepsilon}(\mathbf{x}) = \mathbf{x}^* - \mathbf{x}\). Then:

\[
(1 - \gamma) \zeta_1(\varepsilon) d^{1/p} \frac{\|\nabla f(\mathbf{x}^*)\|_{p'}}{\|\nabla f(\mathbf{x}^*)\|_2} \leq \frac{r_{p,\varepsilon}(\mathbf{x})}{\|r_{p,\varepsilon}(\mathbf{x})\|_p}
\]

and

\[
\frac{r_{p,\varepsilon}(\mathbf{x})}{\|r_{p,\varepsilon}(\mathbf{x})\|_p} \leq (1 + \gamma) \zeta_2(\varepsilon) d^{1/p} \frac{\|\nabla f(\mathbf{x}^*)\|_{p'}}{\|\nabla f(\mathbf{x}^*)\|_2}
\]

provided

\[
\eta \geq (1 + \gamma) \zeta_2(\varepsilon) d^{1/p} \frac{\|\nabla f(\mathbf{x}^*)\|_{p'}}{\|\nabla f(\mathbf{x}^*)\|_2} \|r_{p,\varepsilon}(\mathbf{x})\|_p = \eta_{\text{lim}}.
\]

**Proof.** Let \(f_- \) and \(f_+\) be functions such that the separating hyperplanes of, respectively, \(\mathcal{H}^-_\gamma(\mathbf{x}, \mathbf{x}^*)\) and \(\mathcal{H}^+_\gamma(\mathbf{x}, \mathbf{x}^*)\) are described by equations, respectively, \(f_-(\mathbf{z}) = 0\) and \(f_+(\mathbf{z}) = 0\). By definition, we know that \(\|r_{p,\varepsilon}(f_-, \mathbf{x})\|_p = (1 - \gamma) \|r_{p,\varepsilon}(\mathbf{x})\|_p\) and \(\|r_{p,\varepsilon}(f_+, \mathbf{x})\|_p = (1 + \gamma) \|r_{p,\varepsilon}(\mathbf{x})\|_p\).

From the definition of LAF classifiers, since for all \(\eta' \leq \frac{1}{1 + \gamma} \eta_{\text{lim}}\), \(z \in \mathcal{H}^-_\gamma(\mathbf{x}, \mathbf{x}^*) \cap \mathcal{B}_p(\mathbf{x}, \eta') \Rightarrow f(\mathbf{z}) f(\mathbf{x}) > 0\), we have \(r_{p,\varepsilon}(f_-, \mathbf{x}) \leq r_{p,\varepsilon}(\mathbf{x})\); indeed, if \(\mathbf{x} + \alpha \mathbf{v}\) with \(\alpha \leq \frac{1}{1 + \gamma} \eta_{\text{lim}}\) is not misclassified by \(f_-\), then it is not misclassified by \(f\). Therefore, by applying Lemma \(\text{[2]}\) to \(f_-\), we get:

\[
(1 - \gamma) \zeta_1(\varepsilon) d^{1/p} \frac{\|\nabla f(\mathbf{x}^*)\|_{p'}}{\|\nabla f(\mathbf{x}^*)\|_2} \leq \frac{r_{p,\varepsilon}(\mathbf{x})}{\|r_{p,\varepsilon}(\mathbf{x})\|_p}
\]

Since as long as \(\eta' \leq \eta_{\text{lim}}\), \(z \in \mathcal{H}^+_\gamma(\mathbf{x}, \mathbf{x}^*) \cap \mathcal{B}_p(\mathbf{x}, \eta') \Rightarrow f(\mathbf{z}) f(\mathbf{x}) < 0\), we can apply a symmetric reasoning for \(f_+\), and get:

\[
\frac{r_{p,\varepsilon}(\mathbf{x})}{\|r_{p,\varepsilon}(\mathbf{x})\|_p} \leq (1 + \gamma) \zeta_2(\varepsilon) d^{1/p} \frac{\|\nabla f(\mathbf{x}^*)\|_{p'}}{\|\nabla f(\mathbf{x}^*)\|_2}
\]

\[\square\]

**Theorem 4.** Let \(\Sigma\) be a \(d \times d\) positive semidefinite matrix with \(\text{Tr}(\Sigma) = 1\). Let \(\varepsilon_0, \zeta_1'(\varepsilon), \zeta_2'(\varepsilon)\) as in Theorem \(\text{[3]}\). Then, for all \(\varepsilon < \frac{1}{2} \varepsilon_0\), the following holds.

Assume \(f\) is a classifier that is \((\gamma, \eta)\)-LAF at point \(\mathbf{x}\) and \(\mathbf{x}^*\) be such that \(r_{2,\varepsilon}(\mathbf{x}) = \mathbf{x}^* - \mathbf{x}\). Then:

\[
(1 - \gamma) \zeta_1'(\varepsilon) \frac{\|\nabla f(\mathbf{x}^*)\|_2}{\|\sqrt{\Sigma} \nabla f(\mathbf{x}^*)\|_2} \leq \frac{r_{2,\varepsilon}(\mathbf{x})}{\|r_{2,\varepsilon}(\mathbf{x})\|_2}
\]

and

\[
\frac{r_{2,\varepsilon}(\mathbf{x})}{\|r_{2,\varepsilon}(\mathbf{x})\|_2} \leq (1 + \gamma) \zeta_2'(\varepsilon) \frac{3\varepsilon}{2} \frac{\|\nabla f(\mathbf{x}^*)\|_2}{\|\sqrt{\Sigma} \nabla f(\mathbf{x}^*)\|_2},
\]

provided

\[
\eta \geq (1 + \gamma) \left( 1 + 8 \text{Tr}(\Sigma^2) \ln \frac{4}{\varepsilon} \right) \zeta_2'(\varepsilon) \frac{3\varepsilon}{2} \frac{\|\nabla f(\mathbf{x}^*)\|_2}{\|\sqrt{\Sigma} \nabla f(\mathbf{x}^*)\|_2} \|r_{2,\varepsilon}(\mathbf{x})\|_2 = \eta_{\text{lim}}.
\]
Proof. This proof can be directly adapted from the proof of Theorem 2. The difference in the Gaussian case is that \( v \) is no longer sampled from the unit ball, and its norm is not limited anymore. However, its norm can be bounded with high probability, and this enables to adapt the bounds of Theorem 2 to the Gaussian case.

Indeed, using a Bernstein inequality as in the proof of Proposition 2, we have:

\[
\mathbb{P} \left\{ \left\| \sqrt{\Sigma} v \right\|_2 - 1 \geq t \right\} \leq 2 \exp \left( - \frac{t^2}{8 \text{Tr} (\Sigma^2)} \right) \leq \frac{\varepsilon}{2},
\]

for \( t = \psi(\varepsilon) = 8 \text{Tr} (\Sigma^2) \ln \frac{1}{\delta} \).

Let us focus on the upper bound for this proof; the lower bound follows by a similar reasoning. From the definition of LAF classifiers, since for all for all \( \eta' \leq \eta_{\text{lim}} \), \( v \in \mathcal{H}^+(x, x') \cap B_{\eta'}(x, \eta') \Rightarrow f(x) f(x') < 0 \), we have \( r_{\Sigma, \varepsilon}(x) \leq r_{\Sigma, \frac{1}{\eta_{\text{lim}}}}(f_{\varepsilon}, x) \); indeed, if \( x + \alpha v \) with \( \alpha \leq \frac{1}{\eta_{\text{lim}}} \) is misclassified by \( f_{\varepsilon} \), then it is misclassified by \( f \) if \( \| \alpha v \|_2 \leq \eta_{\text{lim}} \). Therefore, by applying Lemma 2 to \( f_{\varepsilon} \):

\[
\frac{r_{\Sigma, \varepsilon}(x)}{\| r^*_\varepsilon(x) \|_2} \leq (1 + \gamma) \zeta_2 \left( \frac{\varepsilon}{2} \right) \frac{\| \nabla f(x^*) \|_2}{\| \sqrt{\Sigma} \nabla f(x^*) \|_2}.
\]

\( \square \)

E Generalization to Multi-class Classifiers

We present in this section a generalization of Theorem 1 to multi-class linear classifiers, and discuss about the generalization of the other results to the multi-class case.

A classifier \( f \) is said to be linear if for all \( k \in [1, L] \), there are vector \( w_k, b_k \) such that \( f_k(x) = w_k^T x + b_k \). In this setting, Theorem 1 can be generalized by replacing \( \zeta_1(\varepsilon) \) by \( \zeta_1 \left( \frac{\varepsilon}{L - 1} \right) \) in the lower bound.

**Theorem 5.** 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 the constants as defined in Theorem 2.

Let \( k = g(x) \) (the label attributed to \( x \) by \( f \) ), \( j \) be a class such that \( x + r^*_\varepsilon(x) \) lies on the decision boundary between classes \( k \) and \( j \) (i.e., the class of the adversarial perturbation of \( x \)) and \( j' = \text{argmin}_i \frac{\| w_k - w_l \|}{\| w_k - w_{l'} \|} \). Then, for all \( \varepsilon < \varepsilon_0 \):

\[
\zeta_1 \left( \frac{\varepsilon}{L - 1} \right) d_1 \frac{\| w_k - w_{j'} \|_{p'}}{\| w_k - w_j \|_{p'}} \leq \frac{\| w_k - w_{j'} \|_{p'}}{\| w_k - w_j \|_{2}} \leq \zeta_2(\varepsilon) d_1 \frac{\| w_k - w_{j'} \|_{p'}}{\| w_k - w_j \|_{2}}.
\]

**Proof.** We first define for the sake of the demonstration for any class \( l \) the adversarial perturbation in the binary case where only classes \( k \) and \( l \) are considered:

\[
r^*_\varepsilon(x, l) = \text{argmin}_r \| r \|_p \text{ s.t. } f_k(x + r) < f_j(x + r).
\]

It is then possible to express conveniently \( \mathbb{P}_{v \sim B_p} \{ g(x) \neq g(x + \alpha v) \} \):

\[
\mathbb{P}_{v \sim B_p} \{ g(x) \neq g(x + \alpha v) \} = \mathbb{P}_{v \sim B_p} \{ \exists l \neq k, f_k(x) < f_l(x + \alpha v) \}
\]

\[
= \mathbb{P}_{v \sim B_p} \left\{ \exists l \neq k, (w_l - w_k)^T v \geq \frac{f_k(x) - f_l(x)}{\| \alpha \|} \right\}
\]

\[
= \mathbb{P}_{v \sim B_p} \left\{ \exists l \neq k, \left\| \frac{w_l - w_k}{\| w_k - w_{l'} \|} \right\|_{p'} \geq \frac{r^*_\varepsilon(x, l)}{\| \alpha \|} \right\}.
\]

Let us first prove the inequality on the upper bound, as in Lemma 3:

\[
\mathbb{P}_{v \sim B_p} \{ g(x) \neq g(x + \alpha v) \} \geq \mathbb{P}_{v \sim B_p} \left\{ \left\| \frac{w_j - w_k}{\| w_k - w_{l'} \|} \right\|_{p'} \geq \frac{r^*_\varepsilon(x, j)}{\| \alpha \|} \right\}
\]

\[
= \mathbb{P}_{v \sim B_p} \left\{ \left\| \frac{w_j - w_k}{\| w_k - w_{l'} \|} \right\|_{p'} \geq \frac{r^*_\varepsilon(x, j)}{\| \alpha \|} \right\}.
\]

\[
\mathbb{P}_{v \sim B_p} \{ g(x) \neq g(x + \alpha v) \} \leq \mathbb{P}_{v \sim B_p} \left\{ \left\| \frac{w_j - w_k}{\| w_k - w_{l'} \|} \right\|_{p'} \geq \frac{r^*_\varepsilon(x, j)}{\| \alpha \|} \right\}. 
\]
by definition of $j$. Then using the same reasoning as in Lemma 3 leads to

$$\frac{r_{p,\varepsilon}(x)}{\|r_p^*(x)\|_p} \leq \zeta_2(\varepsilon) d^{1/p} \frac{\|w_k - w_j\|_p'}{\|w_k - w_j\|_2}.$$

Let us then prove the inequality on the lower bounds, as in Lemma 2. We use the union bound to derive the inequality:

$$\mathbb{P}_{v \sim B_p \{ \alpha \neq g(x) \}} \leq \sum_{l \neq k} \mathbb{P}_{v \sim B_p \{ \alpha \neq g(x) \}} \left\{ \frac{(w_l - w_k)^T v}{\|w_l - w_k\|_p'} \geq \frac{r^*_p(x, l)}{|\alpha|} \right\},$$

because $r^*_p(x) \geq r^*_p(x, l)$ for all $l$. Moreover, for $|\alpha| < \zeta_1 \left( \frac{\varepsilon}{L - 1} \right) d^{1/2} \frac{\|w_k - w_j\|_p'}{\|w_k - w_j\|_2} \|r^*_p(x)\|_p$, by following the reasoning of Lemma 2 for each $l \neq k$:

$$\mathbb{P}_{v \sim B_p \{ \alpha \neq g(x) \}} \leq \sum_{l \neq k} \frac{\varepsilon}{L - 1} = \varepsilon.$$

Therefore:

$$\frac{r_{p,\varepsilon}(x)}{\|r_p^*(x)\|_p} \geq \zeta_1 \left( \frac{\varepsilon}{L - 1} \right) d^{1/p} \frac{\|w_k - w_j\|_p'}{\|w_k - w_j\|_2}.$$

The proof of this theorem uses the union bound to obtain the lower bound, explaining that $\zeta_1(\varepsilon)$ in the binary case becomes $\zeta_1 \left( \frac{\varepsilon}{L - 1} \right)$ in the multi-class setting. However, this inequality represents a worst case in the majoration used in the proof, and we observed in our experiments that using the coefficient $\zeta_1(\varepsilon)$ instead of $\zeta_1 \left( \frac{\varepsilon}{L - 1} \right)$ gives a proper lower bound on $\frac{r_{p,\varepsilon}(x)}{\|r_p^*(x)\|_p}$.

Notice that it is possible to generalize other results that we proved in the binary case (Lemma 4, Theorems 2 and 3) to the multi-class problem with a similar transformation of the inequalities (replacing $\zeta_1(\varepsilon)$ by $\zeta_1 \left( \frac{\varepsilon}{L - 1} \right)$ and using similar definitions of $j$ and $j'$).