Robustness of classifiers to uniform ℓ_p and Gaussian noise Supplementary material

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

In these appendices, we prove the theoretical results stated in the main article.

A Preliminary Results

In this section, we explicitly compute $\|\boldsymbol{r}_p^*(\boldsymbol{x})\|_p$ for a linear classifier as described in the main article.

Lemma 1. *For all $p \in [1, \infty]$, the ℓ_p-distance from any point $\boldsymbol{x}$ to the decision hyperplane $\mathcal{H}$ defined by $f(\boldsymbol{z}) = 0$ is:*

- *if $p = \infty$:*
$$\|\boldsymbol{r}_\infty^*(\boldsymbol{x})\|_\infty = \frac{|f(\boldsymbol{x})|}{\|\boldsymbol{w}\|_1};$$
- *if $p = 1$:*
$$\|\boldsymbol{r}_1^*(\boldsymbol{x})\|_1 = \frac{|f(\boldsymbol{x})|}{\|\boldsymbol{w}\|_\infty};$$
- *if $p \in (1, \infty)$:*
$$\|\boldsymbol{r}_p^*(\boldsymbol{x})\|_p = \frac{|f(\boldsymbol{x})|}{\|\boldsymbol{w}\|_{\frac{p}{p-1}}}.$$

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

$$\|\boldsymbol{r}_p^*(\boldsymbol{x})\|_p = \frac{|f(\boldsymbol{x})|}{\|\boldsymbol{w}\|_{\frac{p}{p-1}}}.$$

Proof. We distinguish between the three cases.

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

*Now at DeepMind.

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

- Suppose $p \in (1, \infty)$. The distance from $\boldsymbol{x}$ to $\mathcal{H}$ is equal to the minimum radius α of an ℓ_p ball $\mathcal{B}_p$ centered at $\boldsymbol{x}$ that intersects $\mathcal{H}$. This ball is described by the following equation (where $\boldsymbol{z}$ is the variable):
$$\sum_{i=1}^{d} (z_i - x_i)^p \leq \alpha^p.$$
For such a minimum radius, the plane described by $f(\boldsymbol{z}) = 0$ is tangent to $\mathcal{B}_p$ at some point $\boldsymbol{x} + \boldsymbol{n}$. Let us assume without loss of generality that every coordinate of $\boldsymbol{n}$ is non-negative. We also know that this hyperplane is described by the following equations (where $\boldsymbol{z}$ is the variable):
$$\begin{aligned} \nabla_{\boldsymbol{x}+\boldsymbol{n}} \left(\sum_{i=1}^{d} (z_i - x_i)^p - \alpha^p \right)^T (\boldsymbol{z} - (\boldsymbol{x} + \boldsymbol{n})) = 0 \quad &\Leftrightarrow \quad \sum_{i=1}^{d} n_i^{p-1} (z_i - x_i - n_i) = 0 \\ &\Leftrightarrow \quad \sum_{i=1}^{d} n_i^{p-1} (z_i - x_i) = \alpha^p, \end{aligned}$$
beacuse $\boldsymbol{n}$ belongs to the boundary of $\mathcal{B}_p$. The last equation thus describes the same hyperplane as $\boldsymbol{w}^T \boldsymbol{z} = -\boldsymbol{b}$. Therefore, there exists $\lambda \in \mathbb{R} \setminus \{0\}$ such that $\forall i, n_i^{p-1} = \lambda w_i$. Then, since $\boldsymbol{x} + \boldsymbol{n} \in \mathcal{B}_p$:
$$\sum_{i=1}^{d} n_i^{p-1} ((x_i + n_i) - x_i) = \lambda \sum_{i=1}^{d} w_i n_i = \alpha^p,$$
and, since $\boldsymbol{x} + \boldsymbol{n} \in \mathcal{H}$:
$$\sum_{i=1}^{d} w_i (x_i + n_i) + \boldsymbol{b} = f(x) + \boldsymbol{w}^T \boldsymbol{n} = 0,$$
we have $\lambda = -\frac{\alpha^p}{f(\boldsymbol{x})}$. Finally:
$$\begin{aligned} \alpha &= \left(\sum_{i=1}^{d} n_i^p \right)^{\frac{1}{p}} = \left(\sum_{i=1}^{d} (\lambda w_i)^{\frac{p}{p-1}} \right)^{\frac{1}{p}} = \left(\frac{\alpha^p}{|f(\boldsymbol{x})|} \right)^{\frac{1}{p-1}} \|\boldsymbol{w}\|_{\frac{p}{p-1}}^{\frac{1}{p-1}} \\ \alpha &= \frac{|f(\boldsymbol{x})|}{\|\boldsymbol{w}\|_{\frac{p}{p-1}}}. \end{aligned}$$

□

B Robustness of Linear Classifiers to ℓ_p Noise

B.1 Main Theorem

Theorem 1. *Let $p \in [1, \infty]$. Let $p' \in [1, \infty]$ be such that $\frac{1}{p} + \frac{1}{p'} = 1$. Then there exist universal constants $C, c, c' > 0$ such that, for all $\varepsilon < \frac{c^2}{c'}$:*
$$\zeta_1(\varepsilon) d^{1/p} \frac{\|\boldsymbol{w}\|_{p'}}{\|\boldsymbol{w}\|_2} \leq \frac{r_{p,\varepsilon}(\boldsymbol{x})}{\|\boldsymbol{r}_p^*(\boldsymbol{x})\|_p} \leq \zeta_2(\varepsilon) d^{1/p} \frac{\|\boldsymbol{w}\|_{p'}}{\|\boldsymbol{w}\|_2},$$
where $\zeta_1(\varepsilon) = C\sqrt{\varepsilon}$ and $\zeta_2(\varepsilon) = \frac{1}{\sqrt{c - \sqrt{c'\varepsilon}}}$.

Theorem 1 is proved by the following lemmas.

Lemma 2. *There exists a universal constant $C > 0$ such that*

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

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

Proof. Let us first express conveniently $\mathbb{P}_{\boldsymbol{v}\sim\mathcal{B}_p}\left\{g\left(\boldsymbol{x}\right) \neq g\left(\boldsymbol{x}+\alpha\boldsymbol{v}\right)\right\}$, where $\boldsymbol{v}\sim\mathcal{B}_p$ means that $\boldsymbol{v}$ is chosen uniformly at random in $\mathcal{B}_p$:

$$\begin{aligned}
\mathbb{P}_{\boldsymbol{v}\sim\mathcal{B}_p}\left\{g\left(\boldsymbol{x}\right) \neq g\left(\boldsymbol{x}+\alpha\boldsymbol{v}\right)\right\} &= \mathbb{P}_{\boldsymbol{v}\sim\mathcal{B}_p}\left(f\left(\boldsymbol{x}\right)f\left(\boldsymbol{x}+\alpha\boldsymbol{v}\right) \leq 0\right\} \\
&= \mathbb{P}_{\boldsymbol{v}\sim\mathcal{B}_p}\left\{\operatorname{sign}\left(f\left(\boldsymbol{x}\right)\right)\left(\boldsymbol{w}^T x + \boldsymbol{b}\right) \leq -\operatorname{sign}\left(f\left(\boldsymbol{x}\right)\right)\alpha\boldsymbol{w}^T\boldsymbol{v}\right\} \\
&= \mathbb{P}_{\boldsymbol{v}\sim\mathcal{B}_p}\left\{\|\boldsymbol{w}\|_{\frac{p}{p-1}}\|\boldsymbol{r}_p^*(\boldsymbol{x})\|_p \leq -\operatorname{sign}\left(f\left(x\right)\right)\alpha\boldsymbol{w}^T\boldsymbol{v}\right\} \qquad (1) \\
&= \mathbb{P}_{\boldsymbol{v}\sim\mathcal{B}_p}\left\{\|\boldsymbol{w}\|_{p'}\frac{\|\boldsymbol{r}_p^*(\boldsymbol{x})\|_p}{|\alpha|} \leq \boldsymbol{w}^T\boldsymbol{v}\right\} \qquad (2) \\
&= \frac{1}{2}\mathbb{P}_{\boldsymbol{v}\sim\mathcal{B}_p}\left\{\|\boldsymbol{w}\|_{p'}\frac{\|\boldsymbol{r}_p^*(\boldsymbol{x})\|_p}{|\alpha|} \leq \left|\boldsymbol{w}^T\boldsymbol{v}\right|\right\}, \qquad (3)
\end{aligned}$$

where Eq. (1) is given by Lemma 1, and Eq. (2) and (3) follow from $\boldsymbol{v}\sim\mathcal{B}_p \Rightarrow -\boldsymbol{v}\sim\mathcal{B}_p$.

Markov's inequality gives, from Eq. (3):

$$\mathbb{P}_{\boldsymbol{v}\sim\mathcal{B}_p}\left\{g\left(\boldsymbol{x}\right) \neq g\left(\boldsymbol{x}+\alpha\boldsymbol{v}\right)\right\} \leq \frac{1}{2}\frac{\mathbb{E}_{\boldsymbol{v}\sim\mathcal{B}_p}\left[\left(\sum_{i=1}^d w_i v_i\right)^2\right]}{\left(\|\boldsymbol{w}\|_{p'}\frac{\|\boldsymbol{r}_p^*(\boldsymbol{x})\|_p}{|\alpha|}\right)^2}.$$

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

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

Therefore:

$$\mathbb{P}_{\boldsymbol{v}\sim\mathcal{B}_p}\left\{g\left(\boldsymbol{x}\right) \neq g\left(\boldsymbol{x}+\alpha\boldsymbol{v}\right)\right\} \leq \frac{1}{2}\frac{\left(\frac{2C_0}{d^{\frac{1}{p}}}\|\boldsymbol{w}\|_2\right)^2}{\left(\|\boldsymbol{w}\|_{p'}\frac{\|\boldsymbol{r}_p^*(\boldsymbol{x})\|_p}{|\alpha|}\right)^2}.$$

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

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

□

Lemma 3. *There exist universal constants $c, c' > 0$ such that, for all $\varepsilon < \frac{c^2}{c'}$:*

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

where $\zeta_2(\varepsilon) = \frac{1}{\sqrt{c - \sqrt{c'\varepsilon}}}$.

Proof. We first transform the expression of $\mathbb{P}_{\boldsymbol{v}\sim\mathcal{B}_p}\left\{g\left(\boldsymbol{x}\right)\neq g\left(\boldsymbol{x}+\alpha\boldsymbol{v}\right)\right\}$:

$$\begin{aligned}\mathbb{P}_{\boldsymbol{v}\sim\mathcal{B}_p}\left\{g\left(\boldsymbol{x}\right)\neq g\left(\boldsymbol{x}+\alpha\boldsymbol{v}\right)\right\} &= \mathbb{P}_{\boldsymbol{v}\sim\mathcal{B}_p}\left\{\|\boldsymbol{w}\|_{\frac{p}{p-1}}\frac{\|\boldsymbol{r}_p^*(\boldsymbol{x})\|_p}{|\alpha|}\leq \boldsymbol{w}^T\boldsymbol{v}\right\}\\ &= \frac{1}{2}\mathbb{P}_{\boldsymbol{v}\sim\mathcal{B}_p}\left\{\|\boldsymbol{w}\|_{p'}\frac{\|\boldsymbol{r}_p^*(\boldsymbol{x})\|_p}{|\alpha|}\leq \left|\boldsymbol{w}^T\boldsymbol{v}\right|\right\}\\ &= \frac{1}{2}\mathbb{P}_{\boldsymbol{v}\sim\mathcal{B}_p}\left\{\frac{1}{\mathrm{Var}\left(\boldsymbol{w}^T\boldsymbol{v}\right)}\left(\|\boldsymbol{w}\|_{p'}\frac{\|\boldsymbol{r}_p^*(\boldsymbol{x})\|_p}{|\alpha|}\right)^2\leq \frac{\left(\boldsymbol{w}^T\boldsymbol{v}\right)^2}{\mathrm{Var}\left(\boldsymbol{w}^T\boldsymbol{v}\right)}\right\}.\end{aligned}$$

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

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

Note that $\mathbb{E}_{\boldsymbol{v}\sim\mathcal{B}_p}\left(\frac{\left(\boldsymbol{w}^T\boldsymbol{v}\right)^2}{\mathrm{Var}(\boldsymbol{w}^T\boldsymbol{v})}\right)=1$, because $\mathbb{E}_{\boldsymbol{v}\sim\mathcal{B}_p}\left(\boldsymbol{w}^T\boldsymbol{v}\right)=0$. So, by using Paley-Zygmund's inequality with $X=\frac{\left(\boldsymbol{w}^T\boldsymbol{v}\right)^2}{\mathrm{Var}(\boldsymbol{w}^T\boldsymbol{v})}$ and $t=\frac{1}{\mathrm{Var}(\boldsymbol{w}^T\boldsymbol{v})}\left(\|\boldsymbol{w}\|_{p'}\frac{\|\boldsymbol{r}_p^*(\boldsymbol{x})\|_p}{|\alpha|}\right)^2$, when $|\alpha|\geq\frac{\|\boldsymbol{w}\|_{p'}}{\sqrt{\mathrm{Var}(\boldsymbol{w}^T\boldsymbol{v})}}\|\boldsymbol{r}_p^*(\boldsymbol{x})\|_p$:

$$\mathbb{P}_{\boldsymbol{v}\sim\mathcal{B}_p}\left\{g\left(\boldsymbol{x}\right)\neq g\left(\boldsymbol{x}+\alpha\boldsymbol{v}\right)\right\} \quad\geq\quad \frac{\left(1-\frac{1}{\mathrm{Var}(\boldsymbol{w}^T\boldsymbol{v})}\left(\|\boldsymbol{w}\|_{p'}\frac{\|\boldsymbol{r}_p^*(\boldsymbol{x})\|_p}{|\alpha|}\right)^2\right)^2}{2\mathbb{E}_{\boldsymbol{v}\sim\mathcal{B}_p}\left[\frac{(\boldsymbol{w}^T\boldsymbol{v})^4}{\mathrm{Var}(\boldsymbol{w}^T\boldsymbol{v})^2}\right]}.$$

So, if $|\alpha|>\frac{1}{\sqrt{\mathrm{Var}(\boldsymbol{w}^T\boldsymbol{v})-\sqrt{2\varepsilon\mathbb{E}\left[(\boldsymbol{w}^T\boldsymbol{v})^4\right]}}}\|\boldsymbol{w}\|_{p'}\|\boldsymbol{r}_p^*(\boldsymbol{x})\|_p$, then $\mathbb{P}_{\boldsymbol{v}\sim\mathcal{B}_p}\left\{g\left(\boldsymbol{x}\right)\neq g\left(\boldsymbol{x}+\alpha\boldsymbol{v}\right)\right\}>\varepsilon$. According to [Barthe et al., 2005, Theorem 7], there is a universal constant $c_0>0$ such that:

- for $\mathrm{Var}\left(\boldsymbol{w}^T\boldsymbol{v}\right)$:

$$\mathrm{Var}\left(\boldsymbol{w}^T\boldsymbol{v}\right)\geq\left(\frac{c_0}{d^{\frac{1}{p}}}\|\boldsymbol{w}\|_2\right)^2;$$

- for $\mathbb{E}\left[\left(\boldsymbol{w}^T\boldsymbol{v}\right)^4\right]$:

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

So there are universal constants $c=c_0^2, c'=512C_0^4>0$ such that:

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

□

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*

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

where $\zeta_1(\varepsilon)=\frac{C'}{\sqrt{\log\frac{3}{\varepsilon}}}\left(1-\frac{1}{\min(p,2)}\right)$.

Proof. Let $p_2 = \min(p, 2)$. We have:

$$\begin{aligned}
\mathbb{P}_{\boldsymbol{v}\sim\mathcal{B}_p}\left\{g(\boldsymbol{x}) \neq g(\boldsymbol{x}+\alpha\boldsymbol{v})\right\} &= \mathbb{P}_{\boldsymbol{v}\sim\mathcal{B}_p}\left\{\|\boldsymbol{w}\|_{\frac{p}{p-1}} \frac{\|\boldsymbol{r}_p^*(\boldsymbol{x})\|_p}{|\alpha|} \leq \boldsymbol{w}^T\boldsymbol{v}\right\} \\
&= \mathbb{P}_{\boldsymbol{v}\sim\mathcal{B}_p}\left\{e^{\theta t} \leq \exp\left(\theta\boldsymbol{w}^T\boldsymbol{v}\right)\right\},
\end{aligned}$$

where $t = \|\boldsymbol{w}\|_{p'} \frac{\|\boldsymbol{r}_p^*(\boldsymbol{x})\|_p}{|\alpha|}$, for any $\theta > 0$. Markov's inequality gives:

$$\begin{aligned}
\mathbb{P}_{\boldsymbol{v}\sim\mathcal{B}_p}\left\{g(\boldsymbol{x}) \neq g(\boldsymbol{x}+\alpha\boldsymbol{v})\right\} &\leq \frac{1}{e^{\theta t}}\mathbb{E}_{\boldsymbol{v}\sim\mathcal{B}_p}\left[\exp\left(\theta\boldsymbol{w}^T\boldsymbol{v}\right)\right] = \frac{1}{e^{\theta t}}\sum_{k=0}^{\infty}\frac{1}{k!}\mathbb{E}_{\boldsymbol{v}\sim\mathcal{B}_p}\left[\left(\theta\boldsymbol{w}^T\boldsymbol{v}\right)^k\right]. \\
&\leq \frac{1}{e^{\theta t}}\sum_{k=0}^{\infty}\frac{1}{(2k)!}\mathbb{E}_{\boldsymbol{v}\sim\mathcal{B}_p}\left[\left(\theta\boldsymbol{w}^T\boldsymbol{v}\right)^{2k}\right],
\end{aligned}$$

since $\boldsymbol{w}^T\boldsymbol{v}$ is symmetric. In [Barthe et al., 2005, Theorem 7], it is proved that:

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

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

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

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

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

$$\mathbb{E}_{\boldsymbol{v}\sim\mathcal{B}_p}\left[\left|\sum_{i=1}^{d} w_i v_i\right|^k\right] \leq \left(\frac{C_0 k^{\frac{1}{p_2}}}{d^{\frac{1}{p}}}\|\boldsymbol{w}\|_2\right)^k.$$

Thus:

$$\mathbb{P}_{\boldsymbol{v}\sim\mathcal{B}_p}\left\{g(\boldsymbol{x}) \neq g(\boldsymbol{x}+\alpha\boldsymbol{v})\right\} \leq \frac{1}{e^{\theta t}}\sum_{k=0}^{\infty}\frac{1}{(2k)!}\left(\theta\frac{C_0(2k)^{\frac{1}{p_2}}}{d^{\frac{1}{p}}}\|\boldsymbol{w}\|_2\right)^{2k}.$$

We can bound the following power series using Stirling-like bounds [Robbins, 1955] in (4) and (5):

$$\begin{aligned}
\sum_{k=0}^{\infty} \frac{(2k)^{\frac{2k}{p_2}}}{(2k)!} x^k &\leq 1 + \frac{1}{\sqrt{2\pi}} \sum_{k=1}^{\infty} \frac{(2k)^{\frac{2k}{p_2}}}{(2k)^{2k+\frac{1}{2}}} \left(e^2 x\right)^k & (4) \\
&= 1 + \frac{1}{\sqrt{2\pi}} \sum_{k=1}^{\infty} (2k)^{-2\left(1-\frac{1}{p_2}\right)k-\frac{1}{2}} \left(e^2 x\right)^k \\
&\leq 1 + \frac{1}{\sqrt{2\pi}} \sum_{k=1}^{\infty} \left(\left\lfloor 2\left(1-\frac{1}{p_2}\right)k\right\rfloor\right)^{-\left\lfloor 2\left(1-\frac{1}{p_2}\right)k\right\rfloor-\frac{1}{2}} \left(e^2 x\right)^k \\
&\leq 1 + \frac{e}{\sqrt{2\pi}} \sum_{k=1}^{\infty} \frac{\exp\left(-\left\lfloor 2\left(1-\frac{1}{p_2}\right)k\right\rfloor\right)}{\left\lfloor 2\left(1-\frac{1}{p_2}\right)k\right\rfloor!} \left(e^2 x\right)^k & (5) \\
&\leq 1 + \frac{e^2}{\sqrt{2\pi}} \sum_{k=1}^{\infty} \frac{\exp\left(-2\left(1-\frac{1}{p_2}\right)k\right)}{k!} \left(e^2 x\right)^k .
\end{aligned}$$

Therefore:

$$\mathbb{P}_{\boldsymbol{v}\sim\mathcal{B}_p}\left\{g\left(\boldsymbol{x}\right) \neq g\left(\boldsymbol{x}+\alpha\boldsymbol{v}\right)\right\} \leq 3e^{-\theta t} \exp\left(\theta^2 \frac{p_2}{p_2-1} \frac{e^2 C_0^2}{d^{\frac{2}{p}}} \left\|\boldsymbol{w}\right\|_2^2\right).$$

By choosing $\theta = \frac{1}{2} t \left(\frac{p_2}{p_2-1} \frac{e^2 C_0 k^{\frac{1}{p_2}}}{d^{\frac{2}{p}}} \left\|\boldsymbol{w}\right\|_2\right)^{-1}$:

$$\begin{aligned}
\mathbb{P}_{\boldsymbol{v}\sim\mathcal{B}_p}\left\{g\left(\boldsymbol{x}\right) \neq g\left(\boldsymbol{x}+\alpha\boldsymbol{v}\right)\right\} &\leq 3\exp\left(-t^2\left(1-\frac{1}{p_2}\right)\frac{d^{\frac{2}{p}}}{2e^2 C_0^2 \left\|\boldsymbol{w}\right\|_2^2}\right) \\
&= 3\exp\left(-\left(\frac{\|\boldsymbol{r}_p^*(\boldsymbol{x})\|_p}{|\alpha|}\right)^2 \left(1-\frac{1}{p_2}\right)\frac{d^{\frac{2}{p}} \left\|\boldsymbol{w}\right\|_{p'}^2}{2e^2 C_0^2 \left\|\boldsymbol{w}\right\|_2^2}\right).
\end{aligned}$$

So, if $|\alpha| < \frac{C'}{\sqrt{\ln\frac{3}{\varepsilon}}} \left(1-\frac{1}{p_2}\right) d^{\frac{1}{p}} \frac{\|\boldsymbol{w}\|_{p'}}{\|\boldsymbol{w}\|_2} \|\boldsymbol{r}_p^*(\boldsymbol{x})\|_p$, then $\mathbb{P}_{\boldsymbol{v}\sim\mathcal{B}_p}\left\{g\left(\boldsymbol{x}\right) \neq g\left(\boldsymbol{x}+\alpha\boldsymbol{v}\right)\right\} < \varepsilon$, where $C = \frac{1}{2e^2C_0^2} > 0$ is a universal constant, and:

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

□

B.3 Typical Value of the Multiplicative Factor

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 $p = 1$,

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

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

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

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

Thus:

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

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

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

because $\frac{\|\boldsymbol{g}\|_2}{\sqrt{d}} \xrightarrow[\text{a.s.}]{} \frac{1}{\sqrt{2}}$.

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

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

Using the previous computations for $p = 2$, we find:

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

□

C Robustness of Linear Classifiers to Gaussian Noise

C.1 Main Theorem

Theorem 2. *For $\varepsilon < \frac{1}{3}$, $\zeta_1'(\varepsilon) = \sqrt{\frac{1}{2\ln\left(\frac{1}{\varepsilon}\right)}}$ and $\zeta_2'(\varepsilon) = \sqrt{\frac{1}{1-\sqrt{3\varepsilon}}}$:*

$$\zeta_1'(\varepsilon)\frac{\|\boldsymbol{w}\|_2}{\|\sqrt{\Sigma}\boldsymbol{w}\|_2} \le \frac{r_{\Sigma,\varepsilon}(\boldsymbol{x})}{\|\boldsymbol{r}_2^*(\boldsymbol{x})\|_2} \le \zeta_2'(\varepsilon)\frac{\|\boldsymbol{w}\|_2}{\|\sqrt{\Sigma}\boldsymbol{w}\|_2}.$$

Theorem 2 is proved by the following lemmas.

Lemma 5. *For $\zeta_1'(\varepsilon) = \sqrt{\frac{1}{2\ln\left(\frac{1}{\varepsilon}\right)}}$,*

$$\frac{r_{\Sigma,\varepsilon}(\boldsymbol{x})}{\|\boldsymbol{r}_2^*(\boldsymbol{x})\|_2} \ge \zeta_1'(\varepsilon)\frac{\|\boldsymbol{w}\|_2}{\|\sqrt{\Sigma}\boldsymbol{w}\|_2}.$$

Proof. As in the proof of Lemma 2:

$$\mathbb{P}_{\boldsymbol{v}\sim\mathcal{N}(\boldsymbol{0},\Sigma)}\left\{g\left(\boldsymbol{x}+\alpha\boldsymbol{v}\right) \neq g\left(\boldsymbol{x}\right)\right\} = \mathbb{P}_{\boldsymbol{v}\sim\mathcal{N}(\boldsymbol{0},\Sigma)}\left\{\|w\|_2\left\|\boldsymbol{r}_2^*(\boldsymbol{x})\right\|_2 \le |\alpha|\,\boldsymbol{w}^T\boldsymbol{v}\right\}.$$

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

$$\begin{aligned}\mathbb{P}_{\boldsymbol{v}\sim\mathcal{N}(\boldsymbol{0},\Sigma)}\left\{g\left(\boldsymbol{x}+\alpha\boldsymbol{v}\right) \neq g\left(\boldsymbol{x}\right)\right\} &= \mathbb{P}_{\boldsymbol{v}\sim\mathcal{N}(\boldsymbol{0},I_d)}\left\{\|\boldsymbol{w}\|_2\left\|\boldsymbol{r}_2^*(\boldsymbol{x})\right\|_2 \le |\alpha|\,\boldsymbol{w}^T\sqrt{\Sigma}\boldsymbol{v}\right\} \\ &= \mathbb{P}_{\boldsymbol{v}\sim\mathcal{N}(\boldsymbol{0},I_d)}\left\{\|\boldsymbol{w}\|_2\left\|\boldsymbol{r}_2^*(\boldsymbol{x})\right\|_2 \le \left(|\alpha|\sqrt{\Sigma}\boldsymbol{w}\right)^T\boldsymbol{v}\right\}.\end{aligned}$$

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

$$\mathbb{P}_{\boldsymbol{v}\sim\mathcal{N}(\boldsymbol{0},\Sigma)}\left\{g\left(\boldsymbol{x}+\alpha\boldsymbol{v}\right) \neq g\left(\boldsymbol{x}\right)\right\} \le \exp\left(-\frac{1}{2}\left(\frac{\|\boldsymbol{w}\|_2\left\|\boldsymbol{r}_2^*(\boldsymbol{x})\right\|_2}{\alpha\|\sqrt{\Sigma}\boldsymbol{w}\|_2}\right)^2\right).$$

So, if $|\alpha| < \sqrt{\frac{1}{2\ln\frac{1}{\varepsilon}}} \frac{\|\boldsymbol{w}\|_2}{\|\sqrt{\Sigma}\boldsymbol{w}\|_2} \|\boldsymbol{r}_2^*(\boldsymbol{x})\|_2$, then $\mathbb{P}_{\boldsymbol{v}\sim\mathcal{N}(\boldsymbol{0},\Sigma)}\left\{g\left(\boldsymbol{x}+\alpha\boldsymbol{v}\right) \neq g\left(\boldsymbol{x}\right)\right\} < \varepsilon$. Thus,

$$\frac{r_{\Sigma,\varepsilon}(\boldsymbol{x})}{\|\boldsymbol{r}_2^*(\boldsymbol{x})\|_2} \geq \zeta_1'(\varepsilon)\frac{\|\boldsymbol{w}\|_2}{\|\sqrt{\Sigma}\boldsymbol{w}\|_2}.$$

□

Lemma 6. *For $\varepsilon < \frac{1}{3}$ and $\zeta_2'(\varepsilon) = \sqrt{\frac{1}{1-\sqrt{3\varepsilon}}}$,*

$$\frac{r_{\Sigma,\varepsilon}(\boldsymbol{x})}{\|\boldsymbol{r}_2^*(\boldsymbol{x})\|_2} \leq \zeta_2'(\varepsilon)\frac{\|\boldsymbol{w}\|_2}{\|\sqrt{\Sigma}\boldsymbol{w}\|_2}.$$

Proof.

$$\begin{aligned}
\mathbb{P}_{\boldsymbol{v}\sim\mathcal{N}(\boldsymbol{0},\Sigma)}\left\{g\left(\boldsymbol{x}+\alpha\boldsymbol{v}\right) \neq g\left(\boldsymbol{x}\right)\right\} &= \mathbb{P}_{\boldsymbol{v}\sim\mathcal{N}(\boldsymbol{0},I_d)}\left\{\|\boldsymbol{w}\|_2\|\boldsymbol{r}_2^*(\boldsymbol{x})\|_2 \leq \left(|\alpha|\sqrt{\Sigma}\boldsymbol{w}\right)^T\boldsymbol{v}\right\} \\
&= \frac{1}{2}\mathbb{P}_{\boldsymbol{v}\sim\mathcal{N}(\boldsymbol{0},I_d)}\left\{\left(\|\boldsymbol{w}\|_2\|\boldsymbol{r}_2^*(\boldsymbol{x})\|_2\right)^2 \leq \left(\left(\alpha\sqrt{\Sigma}\boldsymbol{w}\right)^T\boldsymbol{v}\right)^2\right\} \\
&= \frac{1}{2}\mathbb{P}_{\boldsymbol{v}\sim\mathcal{N}(\boldsymbol{0},I_d)}\left\{\left(\frac{1}{\alpha}\frac{\|\boldsymbol{w}\|_2}{\|\sqrt{\Sigma}\boldsymbol{w}\|_2}\|\boldsymbol{r}_2^*(\boldsymbol{x})\|_2\right)^2 \leq \frac{\left(\left(\sqrt{\Sigma}\boldsymbol{w}\right)^T\boldsymbol{v}\right)^2}{\|\sqrt{\Sigma}\boldsymbol{w}\|_2^2}\right\}.
\end{aligned}$$

Note that $\mathbb{E}_{\boldsymbol{v}\sim\mathcal{N}(\boldsymbol{0},I_d)}\left(\frac{\left(\left(\sqrt{\Sigma}\boldsymbol{w}\right)^T\boldsymbol{v}\right)^2}{\|\sqrt{\Sigma}\boldsymbol{w}\|_2^2}\right) = \frac{\mathrm{Var}_{\boldsymbol{v}\sim\mathcal{N}(\boldsymbol{0},I_d)}\left(\left(\sqrt{\Sigma}\boldsymbol{w}\right)^T\boldsymbol{v}\right)}{\|\sqrt{\Sigma}\boldsymbol{w}\|_2^2} = 1$. So, by using Paley-Zygmund's inequality, when $|\alpha| \geq \frac{\|\boldsymbol{w}\|_2}{\|\sqrt{\Sigma}\boldsymbol{w}\|_2}\|\boldsymbol{r}_2^*(\boldsymbol{x})\|_2$:

$$\begin{aligned}
\mathbb{P}_{\boldsymbol{v}\sim\mathcal{N}(\boldsymbol{0},\Sigma)}\left(g\left(\boldsymbol{x}+\alpha\boldsymbol{v}\right) \neq g\left(\boldsymbol{x}\right)\right) &\geq \frac{\left(1-\left(\frac{1}{\alpha}\frac{\|\boldsymbol{w}\|_2}{\|\sqrt{\Sigma}\boldsymbol{w}\|_2}\|\boldsymbol{r}_2^*(\boldsymbol{x})\|_2\right)^2\right)^2}{\frac{2\|\sqrt{\Sigma}\boldsymbol{w}\|_4^4+\|\sqrt{\Sigma}\boldsymbol{w}\|_2^4}{\|\sqrt{\Sigma}\boldsymbol{w}\|_2^4}} = \frac{\left(1-\left(\frac{1}{\alpha}\frac{\|\boldsymbol{w}\|_2}{\|\sqrt{\Sigma}\boldsymbol{w}\|_2}\|\boldsymbol{r}_2^*(\boldsymbol{x})\|_2\right)^2\right)^2}{2\left(\frac{\|\sqrt{\Sigma}\boldsymbol{w}\|_4}{\|\sqrt{\Sigma}\boldsymbol{w}\|_2}\right)^4+1} \\
&\geq \frac{\left(1-\left(\frac{1}{\alpha}\frac{\|\boldsymbol{w}\|_2}{\|\sqrt{\Sigma}\boldsymbol{w}\|_2}\|\boldsymbol{r}_2^*(\boldsymbol{x})\|_2\right)^2\right)^2}{3}.
\end{aligned}$$

So, if $|\alpha| > \frac{1}{\sqrt{1-\sqrt{3\varepsilon}}}\frac{\|\boldsymbol{w}\|_2}{\|\sqrt{\Sigma}\boldsymbol{w}\|_2}\|\boldsymbol{r}_2^*(\boldsymbol{x})\|_2$, then $\mathbb{P}_{\boldsymbol{v}\sim\mathcal{N}(\boldsymbol{0},\Sigma)}\left\{g\left(\boldsymbol{x}+\alpha\boldsymbol{v}\right) \neq g\left(\boldsymbol{x}\right)\right\} > \varepsilon$. Therefore,

$$\frac{r_{\Sigma,\varepsilon}(\boldsymbol{x})}{\|\boldsymbol{r}_2^*(\boldsymbol{x})\|_2} \leq \zeta_2'(\varepsilon)\frac{\|\boldsymbol{w}\|_2}{\|\sqrt{\Sigma}\boldsymbol{w}\|_2}.$$

□

C.2 Typical Value of the Multiplicative Factor

Proposition 2. *Let Σ be a $d\times d$ positive semidefinite matrix with $\mathrm{Tr}\left(\Sigma\right) = 1$. If $\boldsymbol{w}$ is a random direction uniformly distributed over the unit ℓ_2-sphere, then, for $t \leq \frac{\sqrt{\pi}}{8}d$:*

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

where $t' = \frac{5}{2}t$.

Proof. Suppose that $\boldsymbol{w}$ is a random direction uniformly distributed over the unit ℓ_2 sphere.

Then $\boldsymbol{w}$ can be written as $\frac{\boldsymbol{g}}{\|\boldsymbol{g}\|_2}$, where $\boldsymbol{g} = (g_1, \ldots, g_d)$ 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

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

where $\sqrt{\Sigma} = \mathrm{Diag}\left((\lambda_i)\right)$ in the previously mentioned orthogonal basis.

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

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

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

$$\mathbb{P}\left\{\left|\sum_{i=1}^d \lambda_i^2 \left(g_i^2 - \mathbb{E}\left(g_i^2\right)\right)\right| \geq 2t\sqrt{\mathrm{Var}\left(g_i^2 - \mathbb{E}\left(g_i^2\right)\right)\sum_{i=1}^d \lambda_i^4}\right\} \leq 2e^{-t^2},$$

for $t \leq \beta\sqrt{\mathrm{Tr}\left(\Sigma^2\right)}$ where $\beta = \frac{\sqrt{\pi}}{8}$ is a constant[1], i.e., for $t \leq \frac{\beta}{2}$:

$$\mathbb{P}\left\{\left|\|\sqrt{\Sigma}\boldsymbol{g}\|_2 - \frac{1}{2}\right| \geq t\right\} \leq 2\exp\left(-\frac{t^2}{2\,\mathrm{Tr}\left(\Sigma^2\right)}\right).$$

$2\|\boldsymbol{g}\|_2^2$ has a chi-squared distribution, so using a simple concentration inequality for the chi-squared distribution[2]:

$$\mathbb{P}\left\{\left|\frac{1}{d}\|\boldsymbol{g}\|_2^2 - \frac{1}{2}\right| \geq t\right\} \leq 2\exp\left(-\frac{dt^2}{2}\right).$$

Overall, for $t \leq \beta d$ and $t' = \frac{5}{2}t$:

$$\begin{aligned}
\mathbb{P}\left\{\left|\left(\frac{\|\boldsymbol{g}\|_2}{\|\sqrt{\Sigma}\boldsymbol{g}\|_2}\right)^2 - d\right| \geq t'\right\} &= \mathbb{P}\left\{\left|\frac{\|\boldsymbol{g}\|_2^2 - d\|\sqrt{\Sigma}\boldsymbol{g}\|_2^2}{\|\sqrt{\Sigma}\boldsymbol{g}\|_2}\right| \geq t'\right\} \\
&= \mathbb{P}\left\{\left|\frac{\left(\|\boldsymbol{g}\|_2^2 - \frac{d}{2}\right) - d\left(\|\sqrt{\Sigma}\boldsymbol{g}\|_2^2 - \frac{1}{2}\right)}{\|\sqrt{\Sigma}\boldsymbol{g}\|_2^2}\right| \geq t'\right\} \\
&\leq \mathbb{P}\left\{\frac{\left|\frac{1}{d}\|\boldsymbol{g}\|_2^2 - \frac{1}{2}\right| + \left|\|\sqrt{\Sigma}\boldsymbol{g}\|_2^2 - \frac{1}{2}\right|}{\|\sqrt{\Sigma}\boldsymbol{g}\|_2^2} \geq \frac{t'}{d}\right\} \\
&\leq \mathbb{P}\left\{\left|\frac{1}{d}\|\boldsymbol{g}\|_2^2 - \frac{1}{2}\right| \geq \frac{t}{2d}\right\} + \mathbb{P}\left\{\left|\|\sqrt{\Sigma}\boldsymbol{g}\|_2^2 - \frac{1}{2}\right| \geq \frac{t}{2d}\right\} \\
&\quad + \mathbb{P}\left\{\left|\|\sqrt{\Sigma}\boldsymbol{g}\|_2^2 - \frac{1}{2}\right| \geq \frac{1}{10}\right\},
\end{aligned}$$

so, using the previous inequalities:

$$\mathbb{P}\left\{\left|\left(\frac{\|\boldsymbol{g}\|_2}{\|\sqrt{\Sigma}\boldsymbol{g}\|_2}\right)^2 - d\right| \geq t'\right\} \leq 2\exp\left(-\frac{t^2}{8d}\right) + 2\exp\left(-\frac{t^2}{8d^2\,\mathrm{Tr}\left(\Sigma^2\right)}\right) + 2\exp\left(-\frac{1}{200\,\mathrm{Tr}\left(\Sigma^2\right)}\right).$$

□

[1] Because $\frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{\pi}} = \mathbb{E}\left(|g_i|^k\right) \leq \frac{1}{2}\mathbb{E}\left(g_i^2\right)\left(\frac{4}{\sqrt{\pi}}\right)^{k-2} k!$ for all $k > 1$.

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

D Robustness of LAF Classifiers to ℓ_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 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:

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

and

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

provided

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

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

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

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

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

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

□

Theorem 4. *Let Σ be a $d \times d$ positive semidefinite matrix with $\operatorname{Tr}(\Sigma) = 1$. 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

$$\eta \geq (1+\gamma)\left(1 + 8\operatorname{Tr}\left(\Sigma^2\right)\ln\frac{4}{\varepsilon}\right)\zeta_2'\left(\frac{3\varepsilon}{2}\right)\frac{\|\nabla f(\boldsymbol{x}^*)\|_2}{\|\sqrt{\Sigma}\nabla f(\boldsymbol{x}^*)\|_2}\|\boldsymbol{r}_2^*(\boldsymbol{x})\|_2 = \eta_{\text{lim}}.$$

Proof. This proof can be directly adapted from the proof of Theorem 3. The difference in the Gaussian case is that $\boldsymbol{v}$ 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 3 to the Gaussian case.

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

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

for $t = \psi(\varepsilon) = 8\operatorname{Tr}\left(\Sigma^2\right)\ln\frac{4}{\varepsilon}$.

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 $\eta' \leq \eta_{\lim}$, $z \in \mathcal{H}_\gamma^+(\boldsymbol{x}, \boldsymbol{x}^*) \cap \mathcal{B}_p(\boldsymbol{x}, \eta') \Rightarrow f(\boldsymbol{z})\, f(\boldsymbol{x}) < 0$, we have $r_{\Sigma,\varepsilon}(\boldsymbol{x}) \leq r_{\Sigma,\frac{3\varepsilon}{2}}(f_+, \boldsymbol{x})$; indeed, if $\boldsymbol{x} + \alpha\boldsymbol{v}$ with $\alpha \leq \frac{\eta_{\lim}}{1+\psi(\varepsilon)}$ is misclassified by f_+, then it is misclassified by f if $\|\alpha\boldsymbol{v}\|_2 \leq \eta_{\lim}$. Therefore, by applying Lemma 6 to f_+:

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

□

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 $\boldsymbol{w}_k, \boldsymbol{b}_k$ such that $f_k(\boldsymbol{x}) = \boldsymbol{w}_k^T\boldsymbol{x} + \boldsymbol{b}_k$. 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 1. Let $k = g(\boldsymbol{x})$ (the label attributed to $\boldsymbol{x}$ by f), j be a class such that $\boldsymbol{x} + \boldsymbol{r}_p^*(\boldsymbol{x})$ lies on the decision boundary between classes k and j (i.e., the class of the adversarial pertubation of $\boldsymbol{x}$) and $j' = \operatorname{argmin}_l \frac{\|\boldsymbol{w}_k - \boldsymbol{w}_l\|_{p'}}{\|\boldsymbol{w}_k - \boldsymbol{w}_l\|_2}$. Then, for all $\varepsilon < \varepsilon_0$:*

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

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:

$$\boldsymbol{r}_p^*(\boldsymbol{x}, l) = \operatorname*{argmin}_{\boldsymbol{r}} \|\boldsymbol{r}\|_p \text{ s.t. } f_k(\boldsymbol{x}+\boldsymbol{r}) < f_j(\boldsymbol{x}+\boldsymbol{r}).$$

It is then possible to express conveniently $\mathbb{P}_{\boldsymbol{v}\sim\mathcal{B}_p}\{g(\boldsymbol{x}) \neq g(\boldsymbol{x}+\alpha\boldsymbol{v})\}$:

$$\begin{aligned}\mathbb{P}_{\boldsymbol{v}\sim\mathcal{B}_p}\{g(\boldsymbol{x}) \neq g(\boldsymbol{x}+\alpha\boldsymbol{v})\} &= \mathbb{P}_{\boldsymbol{v}\sim\mathcal{B}_p}\{\exists l \neq k, f_k(\boldsymbol{x}) < f_l(\boldsymbol{x}+\alpha\boldsymbol{v})\}\\ &= \mathbb{P}_{\boldsymbol{v}\sim\mathcal{B}_p}\left\{\exists l \neq k, (\boldsymbol{w}_l - \boldsymbol{w}_k)^T\boldsymbol{v} \geq \frac{f_k(\boldsymbol{x}) - f_l(\boldsymbol{x})}{|\alpha|}\right\}\\ &= \mathbb{P}_{\boldsymbol{v}\sim\mathcal{B}_p}\left\{\exists l \neq k, \frac{(\boldsymbol{w}_l - \boldsymbol{w}_k)^T}{\|\boldsymbol{w}_l - \boldsymbol{w}_k\|_{p'}}\boldsymbol{v} \geq \frac{\boldsymbol{r}_p^*(\boldsymbol{x}, l)}{|\alpha|}\right\}.\end{aligned}$$

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

$$\begin{aligned}\mathbb{P}_{\boldsymbol{v}\sim\mathcal{B}_p}\{g(\boldsymbol{x}) \neq g(\boldsymbol{x}+\alpha\boldsymbol{v})\} &\geq \mathbb{P}_{\boldsymbol{v}\sim\mathcal{B}_p}\left\{\frac{(\boldsymbol{w}_j - \boldsymbol{w}_k)^T}{\|\boldsymbol{w}_j - \boldsymbol{w}_k\|_{p'}}\boldsymbol{v} \geq \frac{\boldsymbol{r}_p^*(\boldsymbol{x}, j)}{|\alpha|}\right\}\\ &= \mathbb{P}_{\boldsymbol{v}\sim\mathcal{B}_p}\left\{\frac{(\boldsymbol{w}_j - \boldsymbol{w}_k)^T}{\|\boldsymbol{w}_j - \boldsymbol{w}_k\|_{p'}}\boldsymbol{v} \geq \frac{\boldsymbol{r}_p^*(\boldsymbol{x})}{|\alpha|}\right\},\end{aligned}$$

by definition of j. Then using the same reasoning as in Lemma 3 leads to

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

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

$$\begin{aligned}\mathbb{P}_{\boldsymbol{v}\sim\mathcal{B}_p}\{g(\boldsymbol{x}) \neq g(\boldsymbol{x}+\alpha\boldsymbol{v})\} &\leq \sum_{l\neq k}\mathbb{P}_{\boldsymbol{v}\sim\mathcal{B}_p}\left\{\frac{(\boldsymbol{w}_l - \boldsymbol{w}_k)^T}{\|\boldsymbol{w}_l - \boldsymbol{w}_k\|_{p'}}\boldsymbol{v} \geq \frac{\boldsymbol{r}_p^*(\boldsymbol{x}, l)}{|\alpha|}\right\}\\ &\leq \sum_{l\neq k}\mathbb{P}_{\boldsymbol{v}\sim\mathcal{B}_p}\left\{\frac{(\boldsymbol{w}_l - \boldsymbol{w}_k)^T}{\|\boldsymbol{w}_l - \boldsymbol{w}_k\|_{p'}}\boldsymbol{v} \geq \frac{\boldsymbol{r}_p^*(\boldsymbol{x})}{|\alpha|}\right\},\end{aligned}$$

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

$$\mathbb{P}_{\boldsymbol{v} \sim \mathcal{B}_p}\left\{g(\boldsymbol{x}) \neq g(\boldsymbol{x} + \alpha \boldsymbol{v})\right\} \leq \sum_{l \neq k} \frac{\varepsilon}{L-1} = \varepsilon.$$

Therefore:

$$\frac{r_{p,\varepsilon}(\boldsymbol{x})}{\|\boldsymbol{r}_p^*(\boldsymbol{x})\|_p} \geq \zeta_1\left(\frac{\varepsilon}{L-1}\right) d^{1/p} \frac{\|\boldsymbol{w}_k - \boldsymbol{w}_{j'}\|_{p'}}{\|\boldsymbol{w}_k - \boldsymbol{w}_{j'}\|_2}.$$

□

The proof of this theorem uses the union bound to obtain the lower bound, explaining that $\zeta_1(\varepsilon)$ in the binary case becomes $\zeta_1(\frac{\varepsilon}{L-1})$ 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 $\zeta_1(\frac{\varepsilon}{L-1})$ gives a proper lower bound on $\frac{r_{p,\varepsilon}(\boldsymbol{x})}{\|\boldsymbol{r}_p^*(\boldsymbol{x})\|_p}$.

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 transormation of the inequalities (replacing $\zeta_1(\varepsilon)$ by $\zeta_1(\frac{\varepsilon}{L-1})$ and using similar definitions of j and j').

References

[Barthe et al., 2005] Barthe, F., Guédon, O., Mendelson, S., and Naor, A. (2005). A probabilistic approach to the geometry of the ℓ_p^n-ball. *The Annals of Probability*, 33(2):480–513.

[Bernstein, 1927] Bernstein, S. (1927). *Theory of Probability.*

[Galambos, 1987] Galambos, J. (1987). *The Asymptotic Theory of Extreme Order Statistics.* R. E. Krieger, second edition.

[Robbins, 1955] Robbins, H. (1955). A remark on stirling's formula. *The American Mathematical Monthly*, 62(1):26–29.