A THEORETICAL STATEMENTS FOR THE LINEAR MODEL

Before we present the proof of the theorem, we introduce two lemmas are of separate interest that are used throughout the proof of Theorem 1. Recall that the definition of the (standard normalized) maximum-ℓ_2-margin solution (max-margin solution in short) of a dataset $D = \{(x_i, y_i)\}_{i=1}^n$ corresponds to

$$\widehat{\theta} := \arg\max_{\|\theta\|_2 \leq 1} \min_{i \in [n]} y_i \theta^\top x_i, \tag{10}$$

by simply setting $\epsilon_{\text{tr}} = 0$ in Equation 4. The ℓ_2-margin of $\widehat{\theta}$ then reads $\min_{i\in[n]} y_i \widehat{\theta}^\top x_i$. Furthermore for a dataset $D = \{(x_i, y_i)\}_{i=1}^n$ we refer to the induced dataset $\widetilde{D}$ as the dataset with covariate vectors stripped of the first element, i.e.

$$\widetilde{D} = \{(\tilde{x}_i, y_i)\}_{i=1}^n := \{((x_i)_{[2:d]}, y_i)\}_{i=1}^n, \tag{11}$$

where $(x_i)_{[2:d]}$ refers to the last $d-1$ elements of the vector x_i. Furthermore, remember that for any vector z, $z_{[j]}$ refers to the j-th element of z and e_j denotes the j-th canonical basis vector. Further, recall the distribution $\mathbb{P}_r$ as defined in Section 3.1: the label $y \in \{+1, -1\}$ is drawn with equal probability and the covariate vector is sampled as $x = [y\frac{r}{2}, \tilde{x}]$ where $\tilde{x} \in \mathbb{R}^{d-1}$ is a random vector drawn from a standard normal distribution, i.e. $\tilde{x} \sim \mathcal{N}(0, \sigma^2 I_{d-1})$. We generally allow r, used to sample the training data, to differ from r_{test}, which is used during test time.

The following lemma derives a closed-form expression for the normalized max-margin solution for any dataset with fixed separation r in the signal component, and that is linearly separable in the last $d-1$ coordinates with margin $\tilde{\gamma}$.

Lemma A.1. *Let $D = \{(x_i, y_i)\}_{i=1}^n$ be a dataset that consists of points $(x, y) \in \mathbb{R}^d \times \{\pm 1\}$ and $x_{[1]} = y\frac{r}{2}$, i.e. the covariates x_i are deterministic in their first coordinate given y_i with separation distance r. Furthermore, let the induced dataset $\widetilde{D}$ also be linearly separable by the normalized max-ℓ_2-margin solution $\tilde{\theta}$ with an ℓ_2-margin $\tilde{\gamma}$. Then, the normalized max-margin solution of the original dataset D is given by*

$$\widehat{\theta} = \frac{1}{\sqrt{r^2 + 4\tilde{\gamma}^2}} \left[r, 2\tilde{\gamma}\tilde{\theta}\right]. \tag{12}$$

Further, the standard accuracy of $\widehat{\theta}$ for data drawn from $\mathbb{P}_{r_{test}}$ reads

$$\mathbb{P}_{r_{test}}(Y\widehat{\theta}^\top X > 0) = \Phi\left(\frac{r\, r_{test}}{4\sigma\, \tilde{\gamma}}\right). \tag{13}$$

The proof can be found in Section A.3. The next lemma provides high probability upper and lower bounds for the margin $\tilde{\gamma}$ of $\widetilde{D}$ when $\tilde{x}_i$ are drawn from the normal distribution.

Lemma A.2. *Let $\widetilde{D} = \{(\tilde{x}_i, y_i)\}_{i=1}^n$ be a random dataset where $y_i \in \{\pm 1\}$ are equally distributed and $\tilde{x}_i \sim \mathcal{N}(0, \sigma I_{d-1})$ for all i, and $\tilde{\gamma}$ is the maximum ℓ_2 margin that can be written as*

$$\tilde{\gamma} = \max_{\|\tilde{\theta}\|_2 \leq 1} \min_{i \in [n]} y_i \tilde{\theta}^\top \tilde{x}_i.$$

Then, for any $t \geq 0$, with probability greater than $1 - 2e^{-\frac{t^2}{2}}$, we have $\tilde{\gamma}_{\min}(t) \leq \tilde{\gamma} \leq \tilde{\gamma}_{\max}(t)$ where

$$\tilde{\gamma}_{\max}(t) = \sigma\left(\sqrt{\frac{d-1}{n}} + 1 + \frac{t}{\sqrt{n}}\right), \quad \tilde{\gamma}_{\min}(t) = \sigma\left(\sqrt{\frac{d-1}{n}} - 1 - \frac{t}{\sqrt{n}}\right).$$

A.1 PROOF OF THEOREM 3.1

Given a dataset $D = \{(x_i, y_i)\}$ drawn from $\mathbb{P}_r$, it is easy to see that the (normalized) ϵ_{tr}-robust max-margin solution 4 of D with respect to signal-attacking perturbations $T(\epsilon_{\text{tr}}; x_i)$ as defined in Equation 3, can be written as

$$\begin{aligned}
\widehat{\theta}^{\epsilon_{\text{tr}}} &= \arg\max_{\|\theta\|_2 \leq 1} \min_{i \in [n], x_i' \in T(x_i; \epsilon_{\text{tr}})} y_i \theta^\top x_i' \\
&= \arg\max_{\|\theta\|_2 \leq 1} \min_{i \in [n], |\beta| \leq \epsilon_{\text{tr}}} y_i \theta^\top (x_i + \beta e_1) \\
&= \arg\max_{\|\theta\|_2 \leq 1} \min_{i \in [n]} y_i \theta^\top (x_i - y_i \epsilon_{\text{tr}} \operatorname{sign}(\theta_{[1]}) e_1).
\end{aligned}$$

Note that by definition, it is equivalent to the (standard normalized) max-margin solution $\widehat{\theta}$ of the shifted dataset $D_{\epsilon_{\text{tr}}} = \{(x_i - y_i \epsilon_{\text{tr}} \operatorname{sign}(\theta_{[1]}) e_1, y_i)\}_{i=1}^n$. Since $D_{\epsilon_{\text{tr}}}$ satisfies the assumptions of Lemma A.1, it then follows directly that the normalized ϵ_{tr}-robust max-margin solution reads

$$\widehat{\theta}^{\epsilon_{\text{tr}}} = \frac{1}{\sqrt{(r - 2\epsilon_{\text{tr}})^2 + 4\tilde{\gamma}^2}} \left[r - 2\epsilon_{\text{tr}}, 2\tilde{\gamma}\tilde{\theta} \right], \tag{14}$$

by replacing r by $r - 2\epsilon_{\text{tr}}$ in Equation 12. Similar to above, $\tilde{\theta} \in R^{d-1}$ is the (standard normalized) max-margin solution of $\{(\tilde{x}_i, y_i)\}_{i=1}^n$ and $\tilde{\gamma}$ the corresponding margin.

Proof of 1. We can now compute the ϵ_{te}-robust accuracy of the ϵ_{tr}-robust max-margin estimator $\widehat{\theta}^{\epsilon_{\text{tr}}}$ for a given dataset D as a function of $\tilde{\gamma}$. Note that in the expression of $\widehat{\theta}^{\epsilon_{\text{tr}}}$, all values are fixed for a fixed dataset, while $0 \leq \epsilon_{\text{tr}} \leq r - 2\tilde{\gamma}_{\max}$ can be chosen. First note that for a test distribution $\mathbb{P}_r$, the ϵ_{te}-robust accuracy, defined as one minus the robust error (Equation 1), for a classifier associated with a vector θ, can be written as

$$\begin{aligned} \operatorname{Acc}(\theta; \epsilon_{\text{te}}) &= \mathbb{E}_{X,Y \sim \mathbb{P}_r} \left[\mathbb{I}\{ \min_{x' \in T(X; \epsilon_{\text{te}})} Y \theta^\top x' > 0 \} \right] \\ &= \mathbb{E}_{X,Y \sim \mathbb{P}_r} \left[\mathbb{I}\{ Y\theta^\top X - \epsilon_{\text{te}} \theta_{[1]} > 0 \} \right] = \mathbb{E}_{X,Y \sim \mathbb{P}_r} \left[\mathbb{I}\{ Y\theta^\top (X - Y\epsilon_{\text{te}} \operatorname{sign}(\theta_{[1]}) e_1) > 0 \} \right] \end{aligned} \tag{15}$$

Now, recall that by Equation 14 and the assumption in the theorem, we have $r - 2\epsilon_{\text{tr}} > 0$, so that $\operatorname{sign}(\widehat{\theta}^{\epsilon_{\text{tr}}}) = 1$. Further, using the definition of the $T(\epsilon_{\text{tr}}; x)$ in Equation 3 and by definition of the distribution $\mathbb{P}_r$, we have $X_{[1]} = Y\frac{r}{2}$. Plugging into Equation 15 then yields

$$\begin{aligned} \operatorname{Acc}(\widehat{\theta}^{\epsilon_{\text{tr}}}; \epsilon_{\text{te}}) &= \mathbb{E}_{X,Y \sim \mathbb{P}_r} \left[\mathbb{I}\{ Y \widehat{\theta}^{\epsilon_{\text{tr}}\top} (X - Y\epsilon_{\text{te}} e_1) > 0 \} \right] \\ &= \mathbb{E}_{X,Y \sim \mathbb{P}_r} \left[\mathbb{I}\{ Y \widehat{\theta}^{\epsilon_{\text{tr}}\top} (X_{-1} + Y \left(\frac{r}{2} - \epsilon_{\text{te}} \right) e_1) > 0 \} \right] \\ &= \mathbb{P}_{r - 2\epsilon_{\text{te}}} (Y \widehat{\theta}^{\epsilon_{\text{tr}}\top} X > 0) \end{aligned}$$

where X_{-1} is a shorthand for the random vector $X_{-1} = (0; X_{[2]}, \ldots, X_{[d]})$. The assumptions in Lemma A.1 ($D_{\epsilon_{\text{tr}}}$ is linearly separable) are satisfied whenever the $n < d - 1$ samples are distinct, i.e. with probability one. Hence applying Lemma A.1 with $r_{\text{test}} = r - 2\epsilon_{\text{te}}$ and $r = r - 2\epsilon_{\text{tr}}$ yields

$$\operatorname{Acc}(\widehat{\theta}^{\epsilon_{\text{tr}}}; \epsilon_{\text{te}}) = \Phi \left(\frac{r(r - 2\epsilon_{\text{te}})}{4\sigma\tilde{\gamma}} - \epsilon_{\text{tr}} \frac{r - 2\epsilon_{\text{te}}}{2\sigma\tilde{\gamma}} \right). \tag{16}$$

Theorem statement a) then follows by noting that Φ is a monotically decreasing function in ϵ_{tr}. The expression for the robust error then follows by noting that $1 - \Phi(-z) = \Phi(z)$ for any $z \in \mathbb{R}$ and defining

$$\tilde{\varphi} = \frac{\sigma\tilde{\gamma}}{r/2 - \epsilon_{\text{te}}}. \tag{17}$$

Proof of 2. First define $\varphi_{\min}, \varphi_{\max}$ using $\tilde{\gamma}_{\min}, \tilde{\gamma}_{\max}$ as in Equation 17. Then we have by Equation 16

$$\begin{aligned} \operatorname{Err}(\widehat{\theta}^{\epsilon_{\text{tr}}}; \epsilon_{\text{te}}) - \operatorname{Err}(\widehat{\theta}^{0}; \epsilon_{\text{te}}) &= \operatorname{Acc}(\widehat{\theta}^{0}; \epsilon_{\text{te}}) - \operatorname{Acc}(\widehat{\theta}^{\epsilon_{\text{tr}}}; \epsilon_{\text{te}}) \\ &= \Phi \left(\frac{r/2}{\tilde{\varphi}} \right) - \Phi \left(\frac{r/2 - \epsilon_{\text{tr}}}{\tilde{\varphi}} \right) \\ &= \int_{r/2 - \epsilon_{\text{tr}}}^{r/2} \frac{1}{\sqrt{2\pi}\tilde{\varphi}} \mathbb{E}^{-\frac{x^2}{\tilde{\varphi}^2}} dx \end{aligned}$$

By plugging in $t = \sqrt{\frac{2\log 2/\delta}{n}}$ in Lemma A.2, we obtain that with probability at least $1 - \delta$ we have

$$\tilde{\gamma}_{\min} := \sigma \left[\sqrt{\frac{d-1}{n}} - \left(1 + \sqrt{\frac{2\log(2/\delta)}{n}} \right) \right] \leq \tilde{\gamma} \leq \sigma \left[\sqrt{\frac{d-1}{n}} + \left(1 + \sqrt{\frac{2\log(2/\delta)}{n}} \right) \right] =: \tilde{\gamma}_{\max}$$

and equivalently $\varphi_{\min} \leq \tilde{\varphi} \leq \varphi_{\max}$.

Now note the general fact that for all $\tilde{\varphi} \leq \sqrt{2}x$ the density function $f(\tilde{\varphi}; x) = \frac{1}{\sqrt{2\pi}\tilde{\varphi}}\mathbb{E}^{-\frac{x^2}{\tilde{\varphi}^2}}$ is monotonically increasing in $\tilde{\varphi}$.

By assumption of the theorem, $\tilde{\varphi} \leq \sqrt{2}(r/2 - \epsilon_{\text{tr}})(r/2 - \epsilon_{\text{te}})$ so that $f(\tilde{\varphi}; x) \geq f(\varphi_{\text{min}}; x)$ for all $x \in [r/2 - \epsilon_{\text{tr}}, r/2]$ and therefore

$$\int_{r/2-\epsilon_{\text{tr}}}^{r/2} \frac{1}{\sqrt{2\pi}\tilde{\varphi}}\mathbb{E}^{-\frac{x^2}{\tilde{\varphi}^2}}dx \geq \int_{r/2-\epsilon_{\text{tr}}}^{r/2} \frac{1}{\sqrt{2\pi}\varphi_{\text{min}}}\mathbb{E}^{-\frac{x^2}{\tilde{\varphi}^2}}dx = \Phi\left(\frac{r/2}{\varphi_{\text{min}}}\right) - \Phi\left(\frac{r/2 - \epsilon_{\text{tr}}}{\varphi_{\text{min}}}\right).$$

and the statement is proved.

A.2 Proof of Corollary 3.2

We now show that Theorem 3.1 also holds for ℓ_1-ball perturbations with at most radius ϵ. Following similar steps as in Equation 14, the ϵ_{tr}-robust max-margin solution for ℓ_1-perturbations can be written as

$$\widehat{\theta}^{\epsilon_{\text{tr}}} := \arg\max_{\|\theta\|_2 \leq 1} \min_{i \in [n]} y_i \theta^\top (x_i - y_i \epsilon_{\text{tr}} \operatorname{sign}(\theta_{[j^\star(\theta)]}) e_{j^\star(\theta)}) \tag{18}$$

where $j^\star(\theta) := \arg\max_j |\theta_j|$ is the index of the maximum absolute value of θ. We now prove by contradiction that the robust max-margin solution for this perturbation set 9 is equivalent to the solution 14 for the perturbation set 3. We start by assuming that $\widehat{\theta}^{\epsilon_{\text{tr}}}$ does not solve Equation 14, which is equivalent to assuming $1 \notin j^\star(\widehat{\theta}^{\epsilon_{\text{tr}}})$ by definition. We now show how this assumption leads to a contradiction.

Define the shorthand $j^\star := j^\star(\widehat{\theta}^{\epsilon_{\text{tr}}}) - 1$. Since $\widehat{\theta}^{\epsilon_{\text{tr}}}$ is the solution of 18, by definition, we have that $\widehat{\theta}^{\epsilon_{\text{tr}}}$ is also the max-margin solution of the shifted dataset $D_{\epsilon_{\text{tr}}} := (x_i - y_i \epsilon_{\text{tr}} \operatorname{sign}(\theta_{[j^\star+1]}) e_{j^\star+1}, y_i)$. Further, note that by the assumption that $1 \notin j^\star(\widehat{\theta}^{\epsilon_{\text{tr}}})$, this dataset $D_{\epsilon_{\text{tr}}}$ consists of input vectors $x_i = (y_i \frac{r}{2}, \tilde{x}_i - y_i \epsilon_{\text{tr}} \operatorname{sign}(\theta_{[j^\star+1]}) e_{j^\star+1})$. Hence via Lemma A.1, $\widehat{\theta}^{\epsilon_{\text{tr}}}$ can be written as

$$\widehat{\theta}^{\epsilon_{\text{tr}}} = \frac{1}{\sqrt{r^2 - 4(\tilde{\gamma}^{\epsilon_{\text{tr}}})^2}} [r, 2\tilde{\gamma}^{\epsilon_{\text{tr}}} \tilde{\theta}^{\epsilon_{\text{tr}}}], \tag{19}$$

where $\tilde{\theta}^{\epsilon_{\text{tr}}}$ is the normalized max-margin solution of $\widetilde{D} := (\tilde{x}_i - y_i \epsilon_{\text{tr}} \operatorname{sign}(\tilde{\theta}_{[j^\star]}) e_{j^\star}, y_i)$.

We now characterize $\tilde{\theta}^{\epsilon_{\text{tr}}}$. Note that by assumption, $j^\star = j^\star(\tilde{\theta}^{\epsilon_{\text{tr}}}) = \arg\max_j |\tilde{\theta}^{\epsilon_{\text{tr}}}_{[j]}|$. Hence, the normalized max-margin solution $\tilde{\theta}^{\epsilon_{\text{tr}}}$ is the solution of

$$\tilde{\theta}^{\epsilon_{\text{tr}}} := \arg\max_{\|\tilde{\theta}\|_2 \leq 1} \min_{i \in [n]} y_i \tilde{\theta}^\top \tilde{x}_i - \epsilon_{\text{tr}} |\tilde{\theta}_{[j^\star]}| \tag{20}$$

Observe that the minimum margin of this estimator $\tilde{\gamma}^{\epsilon_{\text{tr}}} = \min_{i \in [n]} y_i (\tilde{\theta}^{\epsilon_{\text{tr}}})^\top \tilde{x}_i - \epsilon_{\text{tr}} |\tilde{\theta}^{\epsilon_{\text{tr}}}_{[j^\star]}|$ decreases with ϵ_{tr} as the problem becomes harder $\tilde{\gamma}^{\epsilon_{\text{tr}}} \leq \tilde{\gamma}$, where the latter is equivalent to the margin of $\tilde{\theta}^{\epsilon_{\text{tr}}}$ for $\epsilon_{\text{tr}} = 0$. Since $r > 2\tilde{\gamma}_{\max}$ by assumption in the Theorem, by Lemma A.2 with probability at least $1 - 2\mathbb{E}^{-\frac{\alpha^2(d-1)}{n}}$, we then have that $r > 2\tilde{\gamma} \geq 2\tilde{\gamma}^{\epsilon_{\text{tr}}}$. Given the closed form of $\widehat{\theta}^{\epsilon_{\text{tr}}}$ in Equation 19, it directly follows that $\widehat{\theta}^{\epsilon_{\text{tr}}}_{[1]} = r > 2\tilde{\gamma}^{\epsilon_{\text{tr}}} \|\tilde{\theta}^{\epsilon_{\text{tr}}}\|_2 = \|\widehat{\theta}^{\epsilon_{\text{tr}}}_{[2:d]}\|_2$ and hence $1 \in j^\star(\widehat{\theta}^{\epsilon_{\text{tr}}})$. This contradicts the original assumption $1 \notin j^\star(\widehat{\theta}^{\epsilon_{\text{tr}}})$ and hence we established that $\widehat{\theta}^{\epsilon_{\text{tr}}}$ for the ℓ_1-perturbation set 9 has the same closed form 14 as for the perturbation set 3.

The final statement is proved by using the analogous steps as in the proof of 1. and 2. to obtain the closed form of the robust accuracy of $\widehat{\theta}^{\epsilon_{\text{tr}}}$.

A.3 Proof of Lemma A.1

We start by proving that $\widehat{\theta}$ is of the form

$$\widehat{\theta} = \left[a_1, a_2 \tilde{\theta}\right], \tag{21}$$

for $a_1, a_2 > 0$. Denote by $\mathcal{H}(\theta)$ the plane through the origin with normal θ. We define $d\left((x, y), \mathcal{H}(\theta)\right)$ as the signed euclidean distance from the point $(x, y) \in D \sim \mathbb{P}_r$ to the plane $\mathcal{H}(\theta)$. The signed

euclidean distance is the defined as the euclidean distance from x to the plane if the point (x, y) is correctly predicted by θ, and the negative euclidean distance from x to the plane otherwise. We rewrite the definition of the max l_2-margin classifier. It is the classifier induced by the normalized vector $\widehat{\theta}$, such that

$$\max_{\theta \in \mathbb{R}^d} \min_{(x,y)\in D} d\left((x,y), \mathcal{H}(\theta)\right) = \min_{(x,y)\in D} d\left((x,y), \mathcal{H}(\widehat{\theta})\right).$$

We use that D is deterministic in its first coordinate and get

$$\begin{aligned}\max_{\theta} \min_{(x,y)\in D} d\left((x,y), \mathcal{H}(\theta)\right) &= \max_{\theta} \min_{(x,y)\in D} y(\theta_{[1]} x_{[1]} + \tilde{\theta}^\top \tilde{x}) \\ &= \max_{\theta} \theta_1 \frac{r}{2} + \min_{(x,y)\in D} y\tilde{\theta}^\top \tilde{x}.\end{aligned}$$

Because $r > 0$, the maximum over all θ has $\widehat{\theta}_{[1]} \geq 0$. Take any $a > 0$ such that $\|\tilde{\theta}\|_2 = a$. By definition the max l_2-margin classifier, $\tilde{\theta}$, maximizes $\min_{(x,y)\in D} d\left((x,y), \mathcal{H}(\theta)\right)$. Therefore, $\widehat{\theta}$ is of the form of Equation 21.

Note that all classifiers induced by vectors of the form of Equation 21 classify D correctly. Next, we aim to find expressions for a_1 and a_2 such that Equation 21 is the normalized max l_2-margin classifier. The distance from any $x \in D$ to $\mathcal{H}(\widehat{\theta})$ is

$$d\left(x, \mathcal{H}(\widehat{\theta})\right) = \left| a_1 x_{[1]} + a_2 \tilde{\theta}^\top \tilde{x} \right|.$$

Using that $x_{[1]} = y\frac{r}{2}$ and that the second term equals $a_2 d\left(x, \mathcal{H}(\tilde{\theta})\right)$, we get

$$d\left(x, \mathcal{H}(\widehat{\theta})\right) = \left| a_1 \frac{r}{2} + a_2 d\left(x, \mathcal{H}(\tilde{\theta})\right) \right| = a_1 \frac{r}{2} + \sqrt{1 - a_1^2}\, d\left(x, \mathcal{H}(\tilde{\theta})\right). \tag{22}$$

Let $(\tilde{x}, y) \in \widetilde{D}$ be the point closest in Euclidean distance to $\tilde{\theta}$. This point is also the closest point in Euclidean distance to $\mathcal{H}(\widehat{\theta})$, because by Equation 22 $d\left(x, \mathcal{H}(\widehat{\theta})\right)$ is strictly decreasing for decreasing $d\left(x, \mathcal{H}(\tilde{\theta})\right)$. We maximize the minimum margin $d\left(x, \mathcal{H}(\widehat{\theta})\right)$ with respect to a_1. Define the vectors $a = [a_1, a_2]$ and $v = \left[\frac{r}{2}, d\left(x, \mathcal{H}(\tilde{\theta})\right)\right]$. We find using the dual norm that

$$a = \frac{v}{\|v\|_2}.$$

Plugging the expression of a into Equation 21 yields that $\widehat{\theta}$ is given by

$$\widehat{\theta} = \frac{1}{\sqrt{r^2 + 4\tilde{\gamma}^2}} \left[r, 2\tilde{\gamma}\tilde{\theta}\right].$$

For the second part of the lemma we first decompose

$$\mathbb{P}_{r_{\text{test}}}(Y\widehat{\theta}^\top X > 0) = \frac{1}{2}\mathbb{P}_{r_{\text{test}}}\left[\widehat{\theta}^\top X > 0 \mid Y = 1\right] + \frac{1}{2}\mathbb{P}_{r_{\text{test}}}\left[\widehat{\theta}^\top X < 0 \mid Y = -1\right]$$

We can further write

$$\begin{aligned}\mathbb{P}_{r_{\text{test}}}\left[\widehat{\theta}^\top X > 0 \mid Y = 1\right] &= \mathbb{P}_{r_{\text{test}}}\left[\sum_{i=2}^{d} \widehat{\theta}_{[i]} X_{[i]} > -\widehat{\theta}_{[1]}\, X_{[1]} \mid Y = 1\right] \\ &= \mathbb{P}_{r_{\text{test}}}\left[2\tilde{\gamma} \sum_{i=1}^{d-1} \tilde{\theta}_{[i]} X_{[i]} > -r\, \frac{r_{\text{test}}}{2} \mid Y = 1\right] \\ &= 1 - \Phi\left(-\frac{r\, r_{\text{test}}}{4\sigma\tilde{\gamma}}\right) = \Phi\left(\frac{r\, r_{\text{test}}}{4\sigma\tilde{\gamma}}\right)\end{aligned} \tag{23}$$

where Φ is the cumulative distribution function. The second equality follows by multiplying by the normalization constant on both sides and the third equality is due to the fact that $\sum_{i=1}^{d-1} \tilde{\theta}_{[i]} X_{[i]}$ is

a zero-mean Gaussian with variance $\sigma^2\|\tilde{\theta}\|_2^2 = \sigma^2$ since $\tilde{\theta}$ is normalized. Correspondingly we can write

$$\mathbb{P}_{r_{\text{test}}}\left[\widehat{\theta}^\top X < 0 \mid Y = -1\right] = \mathbb{P}_{r_{\text{test}}}\left[2\tilde{\gamma}\sum_{i=1}^{d-1}\tilde{\theta}_{[i]}X_{[i]} < -r\left(-\frac{r_{\text{test}}}{2}\right) \mid Y = -1\right] = \Phi\left(\frac{r\, r_{\text{test}}}{4\sigma\tilde{\gamma}}\right) \tag{24}$$

so that we can combine 23 and 23 and 24 to obtain $\mathbb{P}_{r_{\text{test}}}(Y\widehat{\theta}^\top X > 0) = \Phi\left(\frac{r\, r_{\text{test}}}{4\sigma\tilde{\gamma}}\right)$. This concludes the proof of the lemma.

A.4 Proof of Lemma A.2

The proof plan is as follows. We start from the definition of the max ℓ_2-margin of a dataset. Then, we rewrite the max ℓ_2-margin as an expression that includes a random matrix with independent standard normal entries. This allows us to prove the upper and lower bounds for the max-ℓ_2-margin in Sections A.4.1 and A.4.2 respectively, using non-asymptotic estimates on the singular values of Gaussian random matrices.

Given the dataset $\widetilde{D} = \{(\tilde{x}_i, y_i)\}_{i=1}^n$, we define the random matrix

$$X = \begin{pmatrix} \tilde{x}_1^\top \\ \tilde{x}_2^\top \\ \dots \\ \tilde{x}_n^\top \end{pmatrix}. \tag{25}$$

where $\tilde{x}_i \sim \mathcal{N}(0, \sigma I_{d-1})$. Let $\mathcal{V}$ be the class of all perfect predictors of $\widetilde{D}$. For a matrix A and vector b we also denote by $|Ab|$ the vector whose entries correspond to the absolute values of the entries of Ab. Then, by definition

$$\tilde{\gamma} = \max_{v\in\mathcal{V}, \|v\|_2=1} \min_{j\in[n]} |Xv|_{[j]} = \max_{v\in\mathcal{V}, \|v\|_2=1} \min_{j\in[n]} \sigma|Qv|_{[j]}, \tag{26}$$

where $Q = \frac{1}{\sigma}X$ is the scaled data matrix.

In the sequel we will use the operator norm of a matrix $A \in \mathbb{R}^{n\times d-1}$.

$$\|A\|_2 = \sup_{v\in\mathbb{R}^{d-1} | \|v\|_2=1} \|Av\|_2$$

and denote the maximum singular value of a matrix A as $s_{\max}(A)$ and the minimum singular value as $s_{\min}(A)$.

A.4.1 Upper bound

Given the maximality of the operator norm and since the minimum entry of the vector $|Qv|$ must be smaller than $\frac{\|Q\|_2}{\sqrt{n}}$, we can upper bound $\tilde{\gamma}$ by

$$\tilde{\gamma} \le \sigma\frac{1}{\sqrt{n}}\|Q\|_2.$$

Taking the expectation on both sides with respect to the draw of $\widetilde{D}$ and noting $\|Q\|_2 \le s_{\max}(Q)$, it follows from Corollary 5.35 of Vershynin (2010) that for all $t \ge 0$:

$$\mathbb{P}\left[\sqrt{d-1} + \sqrt{n} + t \ge s_{\max}(Q)\right] \ge 1 - 2e^{-\frac{t^2}{2}}.$$

Therefore, with a probability greater than $1 - 2e^{-\frac{t^2}{2}}$,

$$\tilde{\gamma} \le \sigma\left(1 + \frac{t + \sqrt{d-1}}{\sqrt{n}}\right).$$

A.4.2 LOWER BOUND

By the definition in Equation 26, if we find a vector $v \in \mathcal{V}$ with $\|v\|_2 = 1$ such that for an $a > 0$, it holds that $\min_{j \in n} \sigma |Xv|_{[j]} > a$, then $\tilde{\gamma} > a$.

Recall the definition of the max-ℓ_2-margin as in Equation 25. As $n < d-1$, the random matrix Q is a wide matrix, i.e. there are more columns than rows and therefore the minimal singular value is 0. Furthermore, Q has rank n almost surely and hence for all $c > 0$, there exists a $v \in \mathbb{R}^{d-1}$ such that

$$\sigma Q v = 1_{\{n\}} c > 0, \tag{27}$$

where $1_{\{n\}}$ denotes the all ones vector of dimension n. The smallest non-zero singular value of Q, $s_{\text{min, nonzero}}(Q)$, equals the smallest non-zero singular value of its transpose $Q^\top$. Therefore, there also exists a $v \in \mathcal{V}$ with $\|v\|_2 = 1$ such that

$$\tilde{\gamma} \geq \min_{j \in [n]} \sigma |Qv|_{[j]} \geq \sigma s_{\text{min,nonzeros}} \left(Q^\top\right) \frac{1}{\sqrt{n}}, \tag{28}$$

where we used the fact that any vector v in the span of non-zero eigenvectors satisfies $\|Qv\|_2 \geq s_{\text{min, nonzeros}}(Q)$ and the existence of a solution v for any right-hand side as in Equation 27. Taking the expectation on both sides, Corollary 5.35 of Vershynin (2010) yields that with a probability greater than $1 - 2e^{-\frac{t^2}{2}}, t \geq 0$ we have

$$\tilde{\gamma} \geq \sigma \left(\frac{\sqrt{d-1} - t}{\sqrt{n}} - 1 \right). \tag{29}$$

B BOUNDS ON THE SUSCEPTIBILITY SCORE

In Theorem 3.1, we give non-asymptotic bounds on the robust and standard error of a linear classifier trained with adversarial logistic regression. Moreover, we use the robust error decomposition in susceptibility and standard error to gain intuition about how adversarial training may hurt robust generalization. In this section, we complete the result of Theorem 3.1 by also deriving non-asymptotic bounds on the susceptibility score of the max ℓ_2-margin classifier.

Using the results in Appendix A, we can prove following Corollary B.1, which gives non asymptotic bounds on the susceptibility score.

Corollary B.1. *Assume $d - 1 > n$. For the ϵ_{te}-susceptibility on test samples from $\mathbb{P}_r$ with $2\epsilon_{te} < r$ and perturbation sets in Equation equation 3 and equation 9 the following holds:*

For $\epsilon_{tr} < \frac{r}{2} - \tilde{\gamma}_{\max}$, with probability at least $1 - 2\mathbb{E}^{-\frac{\alpha^2 (d-1)}{2}}$ for any $0 < \alpha < 1$, over the draw of a dataset D with n samples from $\mathbb{P}_r$, the ϵ_{te}-susceptibility is upper and lower bounded by

$$\begin{aligned} \textit{Susc}(\widehat{\theta}^{\epsilon_{tr}}; \epsilon_{te}) &\leq \Phi\left(\frac{(r - 2\epsilon_{tr})(\epsilon_{te} - \frac{r}{2})}{2\tilde{\gamma}_{\max}\sigma} \right) - \Phi\left(\frac{(r - 2\epsilon_{tr})(-\epsilon_{te} - \frac{r}{2})}{2\tilde{\gamma}_{\min}\sigma} \right) \\ \textit{Susc}(\widehat{\theta}^{\epsilon_{tr}}; \epsilon_{te}) &\geq \Phi\left(\frac{(r - 2\epsilon_{tr})(\epsilon_{te} - \frac{r}{2})}{2\tilde{\gamma}_{\min}\sigma} \right) - \Phi\left(\frac{(r - 2\epsilon_{tr})(-\epsilon_{te} - \frac{r}{2})}{2\tilde{\gamma}_{\max}\sigma} \right) \end{aligned} \tag{30}$$

We give the proof in Subsection B.1. Observe that the bounds on the susceptibility score in Corollary B.1 consist of two terms each, where the second term decreases with ϵ_{tr}, but the first term increases. We recognise following two regimes: the max ℓ_2-margin classifier is close to the ground truth e_1 or not. Clearly, the ground truth classifier has zero susceptibility and hence classifiers close to the ground truth also have low susceptibility. On the other hand, if the max l_2-margin classifier is not close to the ground truth, then putting less weight on the first coordinate increases invariance to the perturbations along the first direction. Recall that by Lemma A.1, increasing ϵ_{tr}, decreases the weight on the first coordinate of the max ℓ_2-margin classifier. Furthermore, in the low sample size regime, we are likely not close to the ground truth. Therefore, the regime where the susceptibility decreases with increasing ϵ_{tr} dominates in the low sample size regime.

To confirm the result of Corollary B.1, we plot the mean and standard deviation of the susceptibility score of 5 independent experiments. The results are depicted in Figure 7. We see that for low standard

error, when the classifier is reasonably close to the optimal classifier, the susceptibility increases slightly with increasing adversarial budget. However, increasing the adversarial training budget, ϵ_{tr}, further, causes the susceptibility score to drop greatly. Hence, we can recognize both regimes and validate that, indeed, the second regime dominates in the low sample size setting.

B.1 Proof of Corollary B.1

We proof the statement by bounding the robustness of a linear classifier. Recall that the robustness of a classifier is the probability that a classifier does not change its prediction under an adversarial attack. The susceptibility score is then given by

$$\text{Susc}(\widehat{\theta}^{\epsilon_{\text{tr}}}; \epsilon_{\text{te}}) = 1 - \text{Rob}(\widehat{\theta}^{\epsilon_{\text{tr}}}; \epsilon_{\text{te}}). \tag{31}$$

The proof idea is as follows: since the perturbations are along the first basis direction, e_1, we compute the distance from the robust l_2-max margin $\widehat{\theta}^{\epsilon_{\text{tr}}}$ to a point $(X, Y) \sim \mathbb{P}$. Then, we note that the robustness of $\widehat{\theta}^{\epsilon_{\text{tr}}}$ is given by the probability that the distance along e_1, from X to the decision plane induced by $\widehat{\theta}^{\epsilon_{\text{tr}}}$ is greater then ϵ_{te}. Lastly, we use the non-asymptotic bounds of Lemma A.2.

Recall, by Lemma A.1, the max l_2-margin classifier is of the form of

$$\widehat{\theta}^{\epsilon_{\text{tr}}} = \frac{1}{\sqrt{(r - 2\epsilon_{\text{tr}})^2 + 4\tilde{\gamma}^2}} \left[r - 2\epsilon_{\text{tr}}, 2\tilde{\gamma}\tilde{\theta} \right]. \tag{32}$$

Let $(X, Y) \sim \mathbb{P}$. The distance along e_1 from X to the decision plane induced by $\widehat{\theta}^{\epsilon_{\text{tr}}}$, $\mathcal{H}(\widehat{\theta}^{\epsilon_{\text{tr}}})$, is given by

$$d_{e_1}(X, \mathcal{H}(\widehat{\theta}^{\epsilon_{\text{tr}}})) = \left| X_{[1]} + \frac{1}{\widehat{\theta}^{\epsilon_{\text{tr}}}_{[0]}} \sum_{i=2}^{d} \widehat{\theta}^{\epsilon_{\text{tr}}}_{[i]} X_{[i]} \right|.$$

Substituting the expression of $\widehat{\theta}^{\epsilon_{\text{tr}}}$ in Equation 32 yields

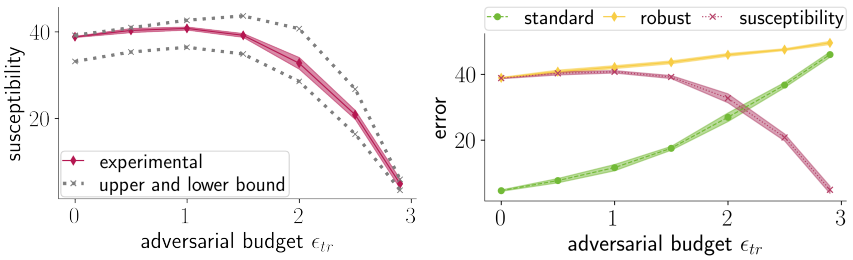
$$d_{e_1}(X, \mathcal{H}(\widehat{\theta}^{\epsilon_{\text{tr}}})) = \left| X_{[1]} + 2\tilde{\gamma} \frac{1}{(r - \epsilon_{\text{tr}})} \sum_{i=2}^{d} \tilde{\theta}_{[i]} X_{[i]} \right|.$$

Let N be a standard normal distributed random variable. By definition $\|\tilde{\theta}\|_2^2 = 1$ and using that a sum of Gaussian random variables is again a Gaussian random variable, we can write

$$d_{e_1}(X, \mathcal{H}(\widehat{\theta}^{\epsilon_{\text{tr}}})) = \left| X_{[1]} + 2\tilde{\gamma} \frac{\sigma}{(r - \epsilon_{\text{tr}})} N \right|.$$

The robustness of $\widehat{\theta}^{\epsilon_{\text{tr}}}$ is given by the probability that $d_{e_1}(X, \mathcal{H}(\widehat{\theta}^{\epsilon_{\text{tr}}})) > \epsilon_{\text{te}}$. Hence, using that $X_1 = \pm\frac{r}{2}$ with probability $\frac{1}{2}$, we get

$$\text{Rob}(\widehat{\theta}^{\epsilon_{\text{tr}}}; \epsilon_{\text{te}}) = P\left[\frac{r}{2} + 2\tilde{\gamma} \frac{\sigma}{(r - 2\epsilon_{\text{tr}})} N > \epsilon_{\text{te}} \right] + P\left[\frac{r}{2} + 2\tilde{\gamma} \frac{\sigma}{(r - \epsilon_{\text{tr}})} N < -\epsilon_{\text{te}} \right]. \tag{33}$$



(a) Susceptibility score decreases with ϵ_{tr} (b) Robust error decomposition

Figure 7: We set $r = 6$, $d = 1000$, $n = 50$ and $\epsilon_{\text{te}} = 2.5$. (a) The average susceptibility score and the standard deviation over 5 independent experiments. Note how the bounds closely predict the susceptibility score. (b) For comparison, we also plot the robust error decomposition in susceptibility and standard error. Even though the susceptibility decreases, the robust error increases with increasing adversarial budget ϵ_{tr}.

We can rewrite Equation 33 in the form

$$\text{Rob}(\widehat{\theta}^{\epsilon_{\text{tr}}}; \epsilon_{\text{te}}) = P\left[N > \frac{(r - 2\epsilon_{\text{tr}})(\epsilon_{\text{te}} - \frac{r}{2})}{2\tilde{\gamma}\sigma}\right] + P\left[N < \frac{(r - 2\epsilon_{\text{tr}})(-\epsilon_{\text{te}} - \frac{r}{2})}{2\tilde{\gamma}\sigma}\right].$$

Recall, that N is a standard normal distributed random variable and denote by Φ the cumulative standard normal density. By definition of the cumulative denisity function, we find that

$$\text{Rob}(\widehat{\theta}^{\epsilon_{\text{tr}}}; \epsilon_{\text{te}}) = 1 - \Phi\left(\frac{(r - 2\epsilon_{\text{tr}})(\epsilon_{\text{te}} - \frac{r}{2})}{2\tilde{\gamma}\sigma}\right) + \Phi\left(\frac{(r - 2\epsilon_{\text{tr}})(-\epsilon_{\text{te}} - \frac{r}{2})}{2\tilde{\gamma}\sigma}\right).$$

Substituting the bounds on $\tilde{\gamma}$ of Lemma A.2 gives us the non-asymptotic bounds on the robustness score and by Equation 31 also on the susceptibility score.

C Experimental details on the linear model

In this section, we provide detailed experimental details to Figures 3 and 4.

We implement adversarial logistic regression using stochastic gradient descent with a learning rate of 0.01. Note that logistic regression converges logarithmically to the robust max l_2-margin solution. As a consequence of the slow convergence, we train for up to 10^7 epochs. Both during training and test time we solve $\max_{x_i' \in T(x_i;\epsilon_{\text{tr}})} L(f_\theta(x_i')y_i)$ exactly. Hence, we exactly measure the robust error. Unless specified otherwise, we set $\sigma = 1$, $r = 12$ and $\epsilon_{\text{te}} = 4$.

Experimental details on Figure 3 (a) We draw 5 datasets with $n = 50$ samples and input dimension $d = 1000$ from the distribution $\mathbb{P}$. We then run adversarial logistic regression on all 5 datasets with adversarial training budgets, $\epsilon_{\text{tr}} = 1$ to 5. To compute the resulting robust error gap of all the obtained classifiers, we use a test set of size 10^6. Lastly, we compute the lower bound given in part 2. of Theorem 3.1. (b) We draw 5 datasets with different sizes n between 50 and 10^4. We take an input dimension of $d = 10^4$ and plot the mean and standard deviation of the robust error after adversarial and standard logistic regression over the 5 samples.(c) We again draw 5 datasets for each d/n constellation and compute the robust error gap for each dataset.

Experimental details on Figure 4 For both (a) and (b) we set $d = 1000$, $\epsilon_{\text{te}} = 4$, and vary the adversarial training budget (ϵ_{tr}) from 1 to 5. For every constellation of n and ϵ_{tr}, we draw 10 datasets and show the average and standard deviation of the resulting robust errors. In (b), we set $n = 50$.

D Experimental details on the Waterbirds dataset

In this section, we discuss the experimental details and construction of the Waterbirds dataset in more detail. We also provide ablation studies of attack parameters such as the size of the motion blur kernel, plots of the robust error decomposition with increasing n, and some experiments using early stopping.

D.1 The waterbirds dataset

To build the Waterbirds dataset, we use the CUB-200 dataset Welinder et al. (2010), which contains images and labels of 200 bird species, and 4 background classes (forest, jungle/bamboo, water ocean, water lake natural) of the Places dataset Zhou et al. (2017).The aim is to recognize whether or not the bird, in a given image, is a waterbird (e.g. an albatros) or a landbird (e.g. a woodpecker). To create the dataset, we randomly sample equally many water- as landbirds from the CUB-200 dataset. Thereafter, we sample for each bird image a random background image. Then, we use the segmentation provided in the CUB-200 dataset to segment the birds from their original images and paste them onto the randomly sampled backgrounds. The resulting images have a size of 256×256. Moreover, we also resize the segmentations such that we have the correct segmentation profiles of the birds in the new dataset as well. For the concrete implementation, we use the code provided by Sagawa et al. (2020).

(a) Original (b) $M = 5$ (c) $M = 10$ (d) $M = 15$ (e) $M = 20$

Figure 8: An ablation study of the motion blur kernel size, which corresponds to the severity level of the blur. For increasing M, the severity of the motion blur increases. In particular, note that for $M = 15$ and even $M = 20$, the bird remains recognizable: we do not semantically change the class, i.e. the perturbations are consistent.

D.2 Experimetal training details

Following the example of Sagawa et al. (2020), we use a ResNet50 or ResNet18 pretrained on the ImageNet dataset for all experiments in the main text, a weight-decay of 10^{-4}, and train for 300 epochs using the Adam optimizer. Extensive fine-tuning of the learning rate resulted in an optimal learning rate of 0.006 for all experiments using the adversarial illumination attack and a pretrained ResNet50. For the experiments considering the adversarial illumnination attack using a pretrained VGG19 or Densenet121 network, we found optimal learning rates of 0.001 and 0.002 respectively. Lastly, we found that for all experiments using the motion blur attack a learning rate of 0.0011 was optimal. Adversarial training is implemented as suggested in Madry et al. (2018): at each iteration we find the worst case perturbation with an exact or approximate method. In all our experiments, the resulting classifier interpolates the training set. We plot the mean over all runs and the standard deviation of the mean.

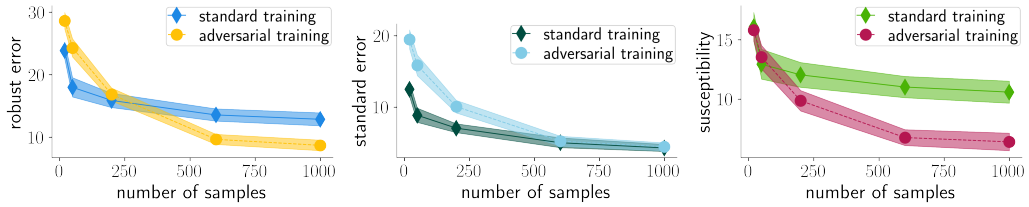
D.3 Specifics to the motion blur attack

Fast moving objects or animals are hard to photograph due to motion blur. Hence, when trying to classify or detect moving objects from images, it is imperative that the classifier is robust against reasonable levels of motion blur. We implement the attack as follows. First, we segment the bird from the original image, then use a blur filter and lastly, we paste the blurred bird back onto the background. We are able to apply more severe blur, by enlarging the kernel of the filter. See Figure 8 for an ablation study of the kernel size.

The motion blur filter is implemented as follows. We use a kernel of size $M \times M$ and build the filter as follows: we fill the row $(M-1)/2$ of the kernel with the value $1/M$. Thereafter, we use the 2D convolution implementation of OpenCV (filter2D) Bradski (2000) to convolve the kernel with the image. Note that applying a rotation before the convolution to the kernel, changes the direction of the resulting motion blur. Lastly, we find the most detrimental level of motion blur using a list-search over all levels up to M_{max}.

(a) $\epsilon = -0.3$ (b) $\epsilon = -0.2$ (c) $\epsilon = -0.1$ (d) Original (e) $\epsilon = 0.1$ (f) $\epsilon = 0.2$ (g) $\epsilon = 0.3$

Figure 9: An ablation study of the different lighting changes of the adversarial illumination attack. Even though the directed attack perturbs the signal component in the image, the bird remains recognizable in all cases.



(a) Robust error (b) Standard error (c) Susceptibility

Figure 10: The robust error decomposition of the experiments depicted in Figure 10a. The plots depict the mean and standard deviation of the mean over several independent experiments. We see that, in comparison to standard training, the reduction in susceptibility for adversarial training is minimal in the low sample size regime. Moreover, the increase in standard error of adversarial training is quite severe, leading to an overall increase in robust error in the low sample size regime.

D.4 Specifics to the adversarial illumination attack

An adversary can hide objects using poor lightning conditions, which can for example arise from shadows or bright spots. To model poor lighting conditions on the object only (or targeted to the object), we use the adversarial illumination attack. The attack is constructed as follows: First, we segment the bird from their background. Then we apply an additive constant ϵ to the bird, where the absolute size of the constant satisfies $|\epsilon| < \epsilon_{\text{te}} = 0.3$. Thereafter, we clip the values of the bird images to $[0, 1]$, and lastly, we paste the bird back onto the background. See Figure 9 for an ablation of the parameter ϵ of the attack. It is non-trivial how to (approximately) find the worst perturbation. We find an approximate solution by searching over all perturbations with increments of size $\epsilon_{\text{te}}/K_{\text{max}}$. Denote by seg, the segmentation profile of the image x. We consider all perturbed images in the form of

$$x_{pert} = (1 - seg)x + seg(x + \epsilon \frac{K}{K_{\text{max}}} 1_{255\times 255}), \quad K \in [-K_{max}, K_{max}].$$

During training time we set $K_{max} = 16$ and therefore search over 33 possible images. During test time we search over 65 images ($K_{max} = 32$).

D.5 Early stopping

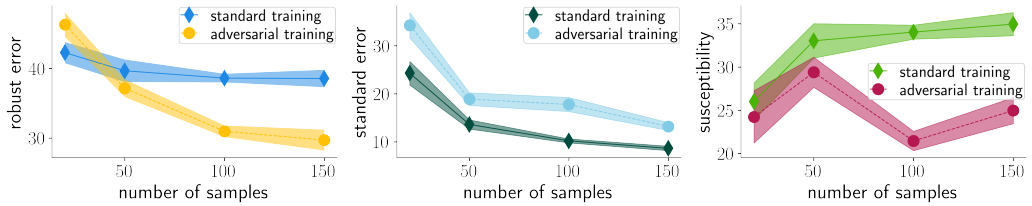
In all our experiments on the Waterbirds dataset, a parameter search lead to an optimal weight-decay and learning rate of 10^{-4} and 0.006 respectively. Another common regularization technique is early stopping, where one stops training on the epoch where the classifier achieves minimal robust error on a hold-out dataset. To understand if early stopping can mitigate the effect of adversarial training aggregating robust generalization in comparison to standard training, we perform the following experiment. On the Waterbirds dataset of size $n = 20$ and considering the adversarial illumination attack, we compare standard training with early stopping and adversarial training ($\epsilon_{\text{tr}} = \epsilon_{\text{te}} = 0.3$) with early stopping. Considering several independent experiments, early stopped adversarial training has an average robust error of 33.5 a early stopped standard training 29.1. Hence, early stopping does decrease the robust error gap, but does not close it.

D.6 Error decomposition with increasing n

In Figure 10a and 11a, we see that adversarial training hurts robust generalization in the small sample size regime. For completeness, we plot the robust error composition for adversarial and standard training in Figure 10. We see that in the low sample size regime, the drop in susceptibility that adversarial training achieves in comparison to standard training, is much lower than the increase in standard error. Conversely, in the high sample regime, the drop of susceptibility from adversarial training over standard training is much bigger than the increase in standard error.

D.7 Different architectures

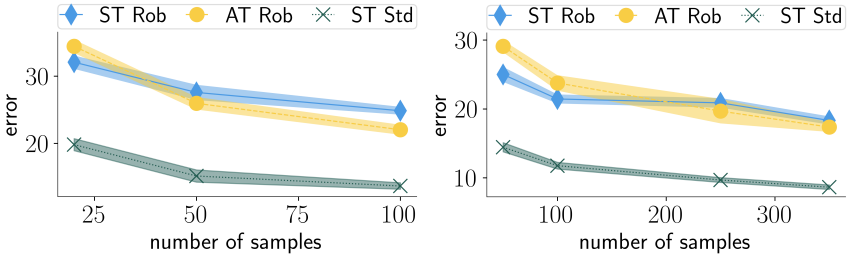
For completeness, we also performed similar experiments on the waterbirds dataset using the adversarial illumination attack with different network architectures as with the pretrained ResNet50



(a) Robust error (b) Standard error (c) Susceptibility

Figure 11: The robust error decomposition of the experiments depicted in Figure 6. The plots depict the mean and standard deviation of the mean over several independent experiments. We see that, in comparison to standard training, the reduction in susceptibility for adversarial training is minimal in the low sample size regime. Moreover, the increase in standard error of adversarial training is quite severe, leading to an overall increase in robust error in the low sample size regime.

network. In particular, we considered the following pretrained network architectures: VGG19 and Densenet121. See Figure 12 for the results. We observe that accros models, adversarial training hurts in the low sample size regime, but helps when enough data is available.



(a) VGG19 (b) Densenet121

Figure 12: The robust error of adversarial training and standard training with increasing sample size using the adversarial illumination attack with $\epsilon_{\text{te}} = 0.3$. We depict the mean and the standard deviation of the mean for multiple runs. Observe that across models, adversarial training hurts in the low sample size regime, but helps when enough samples are available.

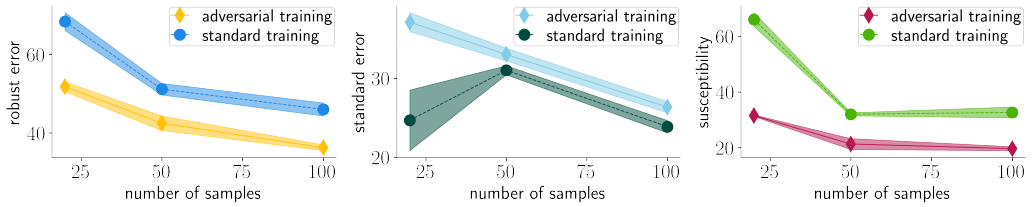
D.8 Undirected attacks on the Waterbirds dataset

In this section, we analyse adversarial training for ℓ_2-and ℓ_∞-ball perturbations in the small sample size regime. We observe that while adversarial training hurts standard generalization, it helps robust generalization.

Adversarial training with ℓ_2-balls We train and test with small ℓ_2-balls, $\epsilon_{\text{te}} = 0.2$, such that the networks trained with standard training achieve a non-zero robust accuracy and the networks trained with adversarial training achieve non-trivial standard accuracy. We see in Figure 13, that adversarial training with ℓ_2-balls hurts standard generalization while increasing robust generalization. Moreover, in Figure 14, we see that also in the very small sample size regime, adversarial training with increasing ϵ_{tr} increases the standard error, but reduces the susceptibility.

Adversarial training with ℓ_∞-balls We also consider ℓ_∞-ball perturbation. We see in Figure 15 that even the smallest perturbation budget $\epsilon_{\text{te}} = \frac{2}{255}$, standard training has robust error of 100 percent. On the other hand, adversarial training achieves low, but non-zero robust error.

Experimental details We use an ImageNet pretrained ResNet34 and train for 300 epochs. Moreover, for reliable robust error and susceptibility evaluation of the attacks we use AutoAttack Croce & Hein (2020). All networks were trained such that the network interpolates the training dataset and has low robust error with non-trivial standard error. For the networks trained using standard training we use a learning rate of 0.006 and for the networks trained with adversarial training we used a learning



(a) Robust error (b) Standard error (c) Susceptibility

Figure 13: The robust error decomposition of adversarial training with ℓ_2-balls of size $\epsilon_{\text{tr}} = 0.2$ and test adversaries with ℓ_2-balls of size $\epsilon_{\text{te}} = 0.2$. The plots depict the mean and standard deviation of the mean over several independent experiments. We see that even though adversarial training hurts standard generalization, it increases robust generalization as it decreases the susceptibility of the classifiers.

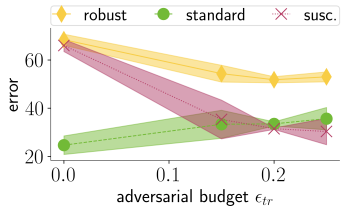
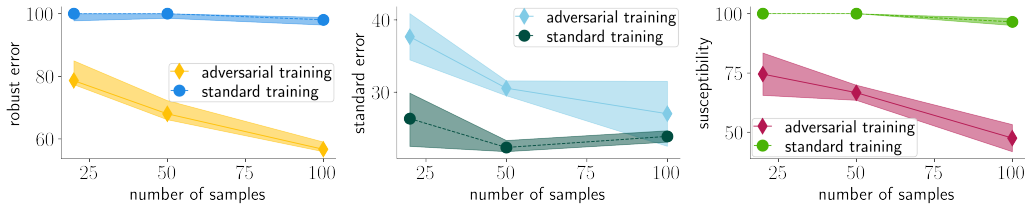


Figure 14: The robust error decomposition of adversarial training in function of ϵ_{tr} in the small sample size regime $n = 20$. We see that even though adversarial training hurts standard generalization, it increases robust generalization as it decreases the susceptibility of the classifiers with increasing ϵ_{tr}. We take $n = 20$ and consider test adversaries with ℓ_2-balls of size $\epsilon_{\text{te}} = 0.2$. The plots depict the mean and standard deviation of the mean over several independent experiments.



(a) Robust error (b) Standard error (c) Susceptibility

Figure 15: The robust error decomposition of adversarial training with ℓ_∞-balls of size $\epsilon_{\text{tr}} = \frac{2}{255}$ and test adversaries with ℓ_∞-balls of size $\epsilon_{\text{te}} = \frac{2}{255}$. The plots depict the mean and standard deviation of the mean over several independent experiments. We see that even though adversarial training hurts standard generalization, it increases robust generalization as it decreases the susceptibility of the classifiers.

rate of $5 \cdot 10^{-4}$. We also trained with a weight decay of 10^{-4}, a batch size of 8 and a momentum of 0.9 for all networks. We train at least 3 networks for all settings and report the mean and standard deviation of the mean of the standard error, robust error and susceptibility over the three runs.

E Experimental details on CIFAR-10

In this section, we give the experimental details on the CIFAR-10-based experiments shown in Figures 1 and 17.

Subsampling CIFAR-10 In all our experiments we subsample CIFAR-10 to simulate the low sample size regime. We ensure that for all subsampled versions the number of samples of each class are equal. Hence, if we subsample to 500 training images, then each class has exactly 50 images, which are drawn uniformly from the $5k$ training images of the respective class.

Mask perturbation on CIFAR-10 On CIFAR-10, we consider the square black mask attack where the adversary can mask a square in the image of size $\epsilon_{\text{te}} \times \epsilon_{\text{te}}$ by setting the pixel values zero. To ensure that the mask cannot cover all the information about the true class in the image, we restrict the size of the masks to be at most 2×2, while allowing for all possible locations of the mask in the targeted image. For exact robust error evaluation, we perform a full grid search over all possible locations during test time. We show an example of a black-mask attack on each of the classes in CIFAR-10 in Figure 16.

During training, a full grid search is computationally intractable so that we use an approximate attack similar to Wu et al. (2020) during training time: by identifying the $K = 16$ most promising mask locations with a heuristic as follows. First, we identify promising mask locations by analyzing the gradient, $\nabla_x L(f_\theta(x), y)$, of the cross-entropy loss with respect to the input. Masks that cover part of the image where the gradient is large, are more likely to increase the loss. Hence, we compute the K mask locations (i, j), where $\|\nabla_x L(f_\theta(x), y)_{[i:i+2, j:j+2]}\|_1$ is the largest and take using a full list-search the mask that incurs the highest loss. Our intuition from the theory predicts that higher K, and hence a more exact "defense", only increases the robust error of adversarial training, since the mask could then more efficiently cover important information about the class.

Figure 16: We show an example of a mask perturbation for all 10 classes of CIFAR-10. Even though the attack occludes part of the images, a human can still easily classify all images correctly.

Experimental training details For all our experiments on CIFAR-10, we adjusted the code provided by Phan (2021). As typically done for CIFAR-10, we augment the data with random cropping and horizontal flipping. For the experiments with results depicted in Figures 1 and 17, we use a ResNet18 network and train for 100 epochs. We tune the parameters learning rate and weight decay for low robust error. For standard standard training, we use a learning rate of 0.01 with equal weight decay. For adversarial training, we use a learning rate of 0.015 and a weight decay of 10^{-4}. We run each experiment three times for every dataset with different initialization seeds, and plot the average and standard deviation over the runs.

For the experiments in Figure 1 and 18 we use an attack strength of $K = 4$. Recall that we perform a full grid search at test time and hence have a good approximation of the robust accuracy and susceptibility score.

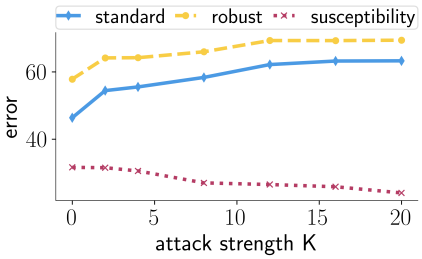


Figure 17: The robust error decomposition in standard error and susceptibility for varying attack strengths K. We see that the larger K, the lower the susceptibility, but the higher the standard error.

Increasing training attack strength We investigate the influence of the attack strength K on the robust error for adversarial training. We take $\epsilon_{\text{tr}} = 2$ and $n = 500$ and vary K. The results are depicted in Figure 17. We see that for increasing K, the susceptibility decreases, but the standard error increases more severely, resulting in an increasing robust error.

Robust error decomposition In Figure 1, we see that the robust error increases for adversarial training compared to standard training in the low sample size regime, but the opposite holds when enough samples are available. For completeness, we provide a full decomposition of the robust error in standard error and susceptibility for standard and adversarial training. We plot the decomposition in Figure 18.

F STATIC HAND GESTURE RECOGNITION

The goal of static hand gesture or posture recognition is to recognize hand gestures such as a pointing index finger or the okay-sign based on static data such as images Oudah et al. (2020); Yang et al.

(2013). The current use of hand gesture recognition is primarily in the interaction between computers and humans Oudah et al. (2020). More specifically, typical practical applications can be found in the environment of games, assisted living, and virtual reality Mujahid et al. (2021). In the following, we conduct experiments on a hand gesture recognition dataset constructed by Mantecón et al. (2019), which consists of near-infrared stereo images obtained using the Leap Motion device. First, we crop or segment the images after which we use logistic regression for classification. We see that adversarial logistic regression deteriorates robust generalization with increasing ϵ_{tr}.

Static hand-gesture dataset We use the dataset made available by Mantecón et al. (2019). This dataset consists of near-infrared stereo images taken with the Leap Motion device and provides detailed skeleton data. We base our analysis on the images only. The size of the images is 640×240 pixels. The dataset consists of 16 classes of hand poses taken by 25 different people. We note that the variety between the different people is relatively wide; there are men and women with different posture and hand sizes. However, the different samples taken by the same person are alike.

We consider binary classification between the index-pose and L-pose, and take as a training set 30 images of the users 16 to 25. This results in a training dataset of 300 samples. We show two examples of the training dataset in Figure 19, each corresponding to a different class. Observe that the near-infrared images darken the background and successfully highlight the hand-pose. As a test dataset, we take 10 images of each of the two classes from the users 1 to 10 resulting in a test dataset of size 200.

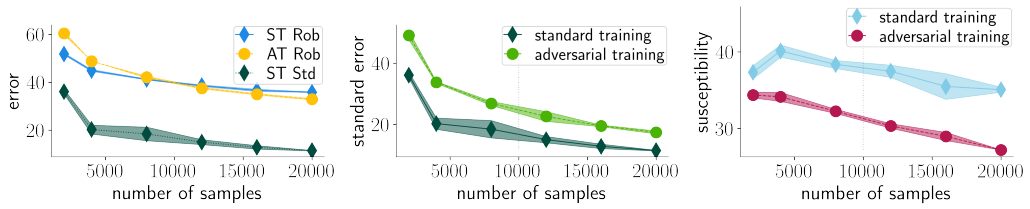
Cropping the dataset To speed up training and ease the classification problem, we crop the images from a size of 640×240 to a size of 200×200. We crop the images using a basic image segmentation technique to stay as close as possible to real-world applications. The aim is to crop the images such that the hand gesture is centered within the cropped image.

For every user in the training set, we crop an image of the L-pose and the index pose by hand. We call these images the training masks $\{\text{masks}_i\}_{i=1}^{20}$. We note that the more a particular window of an image resembles a mask, the more likely that the window captures the hand gesture correctly. Moreover, the near-infrared images are such that the hands of a person are brighter than the surroundings of the person itself. Based on these two observations, we define the best segment or window, defined by the upper left coordinates (i, j), for an image x as the solution to the following optimization problem:

$$\underset{i\in[440],\ \ j\in[40]}{\arg\min} \sum_{l=1}^{20} \|\text{masks}_l - x_{\{i:i+200,j:j+200\}}\|_2^2 - \frac{1}{2}\|x_{\{i+w,j+h\}}\|_1. \tag{34}$$

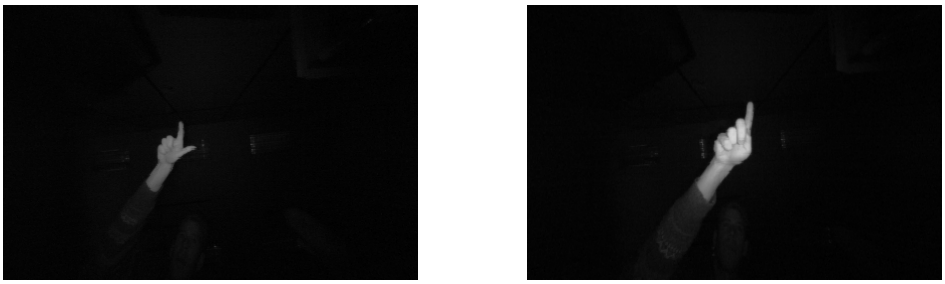
Equation 34 is solved using a full grid search. We use the result to crop both training and test images. Upon manual inspection of the cropped images, close to all images were perfectly cropped. We replace the handful poorly cropped training images with hand-cropped counterparts.

Square-mask perturbations Since we use logistic regression, we perform a full grid search to find the best adversarial perturbation at training and test time. For completeness, the upper left coordinates



(a) Robust error (b) Standard error (c) Susceptibility

Figure 18: The robust error decomposition in standard error and susceptibility of the subsampled datasets of CIFAR-10 after adversarial and standard training. For small sample size, adversarial training has higher robust error then standard training.



(a) L pose (b) Index pose

Figure 19: Examples of the original images of the considered hand-gestures. We recognize the "L"-sign in Figure 19a and the index sign in Figure 19b. Observe that the near-infrared images highlight the hand pose well and blends out much of the non-useful or noisy background.

(a) Cropped L pose (b) Cropped index pose (c) Black-mask perturbation

Figure 20: Examples of the cropped hand-gesture images. We see that the hands are centered and the images have a size of 200×200. In Figure 20c we show an example of the square black-mask perturbation.

of the optimal black-mask perturbation of size $\epsilon_{\text{tr}} \times \epsilon_{\text{tr}}$ can be found as the solution to

$$\arg \max_{i \in [200-\epsilon_{\text{tr}}],\ j \in [200-\epsilon_{\text{tr}}]} \sum_{l,m \in [\epsilon_{\text{tr}}]} \theta_{[i:i+l, j:j+m]}. \tag{35}$$

The algorithm is rather slow as we iterate over all possible windows. We show a black-mask perturbation on an L-pose image in Figure 20c.

Results We run adversarial logistic regression with square-mask perturbations on the cropped dataset and vary the adversarial training budget and plot the result in Figure 21. We observe attack that adversarial logistic regression deteriorates robust generalization.

Because we use adversarial logistic regression, we are able to visualize the classifier. Given the classifier induced by θ, we can visualize how it classifies the images by plotting $\frac{\theta - \min_{i \in [d]} \theta_{[i]}}{\max_{i \in [d]} \theta_{[i]}} \in [0,1]^d$. Recall that the class-prediction of our predictor for a data point (x, y) is given by $\text{sign}(\theta^\top x) \in \{\pm 1\}$. The lighter parts of the resulting image correspond to the class with label 1 and the darker patches with the class corresponding to label -1.

We plot the classifiers obtained by standard logistic regression and adversarial logistic regression with training adversarial budgets ϵ_{tr} of 10 and 25 in Figure 22. The darker parts in the classifier correspond to patches that are typically bright for the L-pose. Complementary, the lighter patches in the classifier correspond to patches that are typically bright for the index pose. We see that in the case of adversarial logistic regression, the background noise is much higher than for standard logistic regression. In other words, adversarial logistic regression puts more weight on non-signal parts in the images to classify the

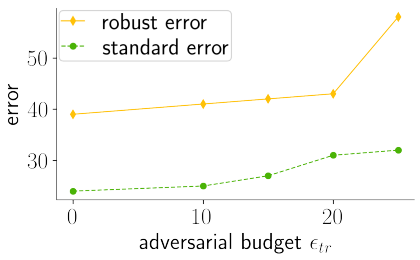
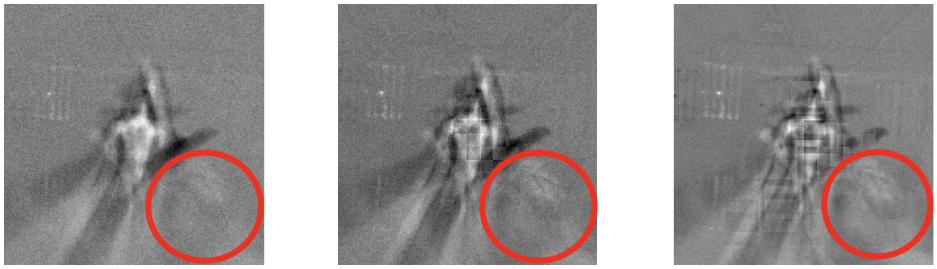


Figure 21: The standard error and robust error for varying adversarial training budget ϵ_{tr}. We see that the larger ϵ_{tr} the higher the robust error.

training dataset and hence exhibits worse performance on the test dataset.



(a) $\epsilon_{\text{tr}} = 0$ (b) $\epsilon_{\text{tr}} = 10$ (c) $\epsilon_{\text{tr}} = 25$

Figure 22: We visualize the logistic regression solutions. In Figure 22a we plot the vector that induces the classifier obtained after standard training. In Figure 22b and Figure 22c we plot the vector obtained after training with square-mask perturbations of size 10 and 25, respectively. We note the non-signal enhanced background correlations at the parts highlighted with the red circles in the image projection of the adversarially trained classifiers.