Sparsity-based Defense against Adversarial Attacks on Linear Classifiers

Zhinus Marzi*, Soorya Gopalakrishnan*, Upamanyu Madhow, Ramtin Pedarsani
University of California, Santa Barbara
Email: {zhinus_marzi, soorya, madhow, ramtin}@ucsb.edu

***Abstract*—Deep neural networks represent the state of the art in machine learning in a growing number of fields, including vision, speech and natural language processing. However, recent work raises important questions about the robustness of such architectures, by showing that it is possible to induce classification errors through tiny, almost imperceptible, perturbations. Vulnerability to such "adversarial attacks", or "adversarial examples", has been conjectured to be due to the excessive linearity of deep networks. In this paper, we study this phenomenon in the setting of a linear classifier, and show that it is possible to exploit sparsity in natural data to combat ℓ_∞-bounded adversarial perturbations. Specifically, we demonstrate the efficacy of a sparsifying front end via an ensemble averaged analysis, and experimental results for the MNIST handwritten digit database. To the best of our knowledge, this is the first work to show that sparsity provides a theoretically rigorous framework for defense against adversarial attacks.**

I. Introduction

Recent work in machine learning security points out the vulnerability of deep neural networks to adversarial perturbations [1]–[4]. These perturbations can be designed to be barely noticeable to the human eye, but can cause large classification errors in state of the art deep networks. While it is tempting to speculate that this vulnerability arises from the complex, nonlinear nature of deep networks, a more plausible explanation is that it is due to the excessive linearity of such networks [3]–[6]. When we take a linear combination of the components of a high-dimensional input, small, adversarially chosen, perturbations of each component can add up to a large perturbation at the output. Complex operations such as a rectified linear unit (ReLU) operating beyond its bias, or a sigmoid in its linear region, together with operations such as max pooling or average pooling, when cascaded through multiple stages, still amount to an approximately linear combination of the input. Of course, the coefficients of the linear combination exhibit some dependence on the input, but these can be viewed as on-off switches rather than a change in the value of the coefficients: for example, whether the input is such that a ReLU unit is operating in its linear region, or the identity of the argument of the maximum in a max pooling unit. This motivates us to take a step back in this paper, and study adversarial perturbations in the simplest possible setting: a linear classifier.

Sparsity is an intuitively plausible mechanism: we understand that humans reject small perturbations by focusing on the key features that stand out. Our proposed approach is based on this intuition. In this paper we show via both theoretical results and experiments that a sparsity-based defense is effective against ℓ_∞-bounded perturbations.

We consider a system consisting of a linear classifier and two participants: the adversary and the defender. The adversary perturbs the input data, with the goal of causing misclassification. The defender inserts a pre-processing function in order to attenuate the impact of the adversary. We propose a sparsifying front end as the preprocessing function and evaluate its performance in two scenarios: a "semi-white box" setting where the adversary designs the perturbation based on the linear model, but without accounting for the pre-processing, and a "white box" setting where the attack accounts for both the pre-processing function and the classifier.

Contributions: We develop a theoretical framework to assess and demonstrate the effectiveness of a sparsity-based defense against adversarial attacks. To the best of our knowledge, this is the first work to show that sparsity provides a rigorous foundation for defense against adversarial perturbations. Our main contributions in this paper are as follows:

- We quantify the achievable gain of the sparsity-based defense via an ensemble-averaged analysis based on a stochastic model for the linear classifier. As the main theoretical contribution of the paper, in Theorems 1 and 2 we show that with high probability, sparsity-based defense reduces the adversarial impact by a factor of K/N in the semi-white box setting, and by $\mathcal{O}(K\,\mathrm{polylog}(N)/N)$ in the white box setting, where K is the sparsity of the signal, and N is the signal's dimension.
- We demonstrate the robustness of our proposed defense through experimental results for binary classification using a linear SVM on the MNIST handwritten digit database. Small adversarial perturbations can render such a classifier useless (0% accuracy), but our sparsity-based defense limits the damage to 1-4% degradation in accuracy for the semi-white and white box attacks, respectively.

II. Related Work

The existence of "blind spots" in deep neural networks [1] has been the subject of extensive recent study in machine learning literature [2]. It was initially hypothesized that this phenomenon is due to the high complexity of neural networks, but work on linearization-based attacks [3], [4] and decision

*Joint first authors.

978-1-5386-4780-6/18/$31.00©2018 IEEE

boundaries of deep networks [5], [6] indicates that it is instead due to their excessive linearity. A variety of defenses have been proposed to combat adversarial attacks, including some that implicitly make use of sparsity-related techniques [7], [8]. The evaluations in such prior work have been purely empirical. Our analytical framework supplements these by providing a theoretical justification for systematic and explicit pursuit of sparsity-based defenses. It is worth noting that sparsity has also been suggested purely as a means of improving classification performance [9], which indicates that the performance penalty for appropriately designed sparsity-based defenses could be minimal.

III. Problem Formulation

A. Preliminaries

We denote by $\boldsymbol{x} \in \mathbb{R}^N$ a data sample with K-sparse representation in orthonormal basis Ψ $(= [\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \ldots, \boldsymbol{\psi}_N])$:

$$\left\|\Psi^T \boldsymbol{x}\right\|_0 \leq K \qquad (K \ll N).$$

Given a linear model $\boldsymbol{w} \in \mathbb{R}^N$, and denoting by $\hat{\boldsymbol{x}}$ a *modified* data sample, we define performance measure Δ:

$$\Delta(\boldsymbol{x}, \hat{\boldsymbol{x}}) = |\boldsymbol{w}^T \hat{\boldsymbol{x}} - \boldsymbol{w}^T \boldsymbol{x}|.$$

B. System Model

Now we describe our system (depicted in Fig. 1) composed of two blocks, the adversary and the defense:

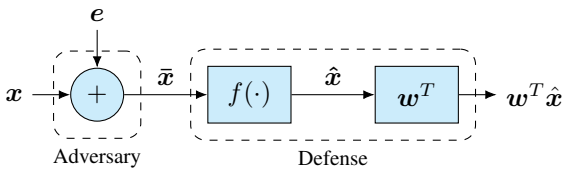

Fig. 1. Block diagram of the system.

– The *adversary* induces an ℓ_∞-bounded additive perturbation $\boldsymbol{e} \in \mathbb{R}^N$ to data $\boldsymbol{x}$, with the goal of maximizing Δ:

$$\begin{aligned} \max_{\boldsymbol{e}} \quad & \Delta(\boldsymbol{x}, \hat{\boldsymbol{x}}) \\ \text{s.t.} \quad & \|\boldsymbol{e}\|_\infty < \epsilon. \end{aligned}$$

– The *defense* adds a pre-processing function $f : \mathbb{R}^N \to \mathbb{R}^N$ to the linear model $\boldsymbol{w}$, with the goal of minimizing Δ.

IV. Sparsity-based defense

A. Pre-processing Function

Given a linear classifier, we propose a pre-processing function via a *sparsifying front end* to combat adversarial attacks. Figure 2 shows a block diagram of our model, composed of sparsity-based preprocessing and a linear machine learning model $\boldsymbol{w}^T$. Function $\mathcal{H}_{\mathrm{K}}(\cdot)$ enforces sparsity by retaining the K coefficients largest in magnitude and zeroing out the rest. Since $\boldsymbol{x}$ is K-sparse in orthonormal basis Ψ, we note that $\hat{\boldsymbol{x}} = \boldsymbol{x}$ when there is no attack ($\boldsymbol{e} = \boldsymbol{0}$).

We define the following quantities:

$$\begin{aligned} \mathcal{S}_{\mathrm{K}}(\boldsymbol{x}) &\triangleq \operatorname{supp}\big(\mathcal{H}_{\mathrm{K}}\big(\Psi^T \boldsymbol{x}\big)\big), \\ \mathcal{P}_{\mathrm{K}}(\boldsymbol{e}, \boldsymbol{x}) &\triangleq \sum_{k \in \mathcal{S}_{\mathrm{K}}(\boldsymbol{x})} \boldsymbol{\psi}_k \boldsymbol{\psi}_k^T \boldsymbol{e}, \end{aligned}$$

where $\mathcal{S}_{\mathrm{K}}(\boldsymbol{x})$ is the support of the K-sparse representation of $\boldsymbol{x}$, and $\mathcal{P}_{\mathrm{K}}(\boldsymbol{e}, \boldsymbol{x})$ is the projection of $\boldsymbol{e}$ on the subspace spanned by $\mathcal{S}_{\mathrm{K}}(\boldsymbol{x})$.

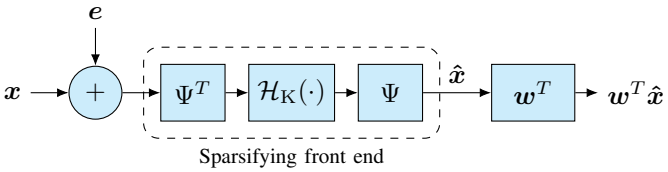

Fig. 2. Block diagram of sparsity-based defense

We also define the *high SNR regime* as the operating region where the additive perturbation does not shift the K-dimensional subspace of $\boldsymbol{x}$:

$$\mathcal{S}_{\mathrm{K}}(\boldsymbol{x}) = \mathcal{S}_{\mathrm{K}}(\boldsymbol{x} + \boldsymbol{e}). \tag{1}$$

In Section V, Proposition 1, we characterize the conditions that guarantee (1). Now assuming that we operate in the high SNR regime, we get

$$\mathcal{H}_{\mathrm{K}}\big(\Psi^T(\boldsymbol{x} + \boldsymbol{e})\big) = \mathcal{H}_{\mathrm{K}}\big(\Psi^T \boldsymbol{x}\big) + \bar{\boldsymbol{e}} = \Psi^T \boldsymbol{x} + \bar{\boldsymbol{e}},$$

where

$$\bar{e}_k = \begin{cases} \boldsymbol{\psi}_k^T \boldsymbol{e}, & \text{if } k \in \mathcal{S}_{\mathrm{K}}(\boldsymbol{x}) \\ 0, & \text{otherwise.} \end{cases}$$

The output of the pre-processing function thus becomes

$$\hat{\boldsymbol{x}} = \boldsymbol{x} + \sum_{k \in \mathcal{S}_{\mathrm{K}}(\boldsymbol{x})} \boldsymbol{\psi}_k \boldsymbol{\psi}_k^T \boldsymbol{e} \;=\; \boldsymbol{x} + \mathcal{P}_{\mathrm{K}}(\boldsymbol{e}, \boldsymbol{x}).$$

Therefore, the performance measure or adversarial attack's impact will be

$$\begin{aligned} \Delta = \left|\boldsymbol{w}^T \hat{\boldsymbol{x}} - \boldsymbol{w}^T \boldsymbol{x}\right| &= \left|\boldsymbol{w}^T \mathcal{P}_{\mathrm{K}}(\boldsymbol{e}, \boldsymbol{x})\right| \\ &= \left|\boldsymbol{e}^T \mathcal{P}_{\mathrm{K}}(\boldsymbol{w}, \boldsymbol{x})\right|, \end{aligned} \tag{2}$$

where (2) follows directly from the definition of $\mathcal{P}_{\mathrm{K}}(\boldsymbol{e}, \boldsymbol{x})$.

B. Attacks and defenses

We now compare the robustness of both the plain classifier and our proposed model against various attacks designed based on partial/full knowledge of the defense.

1. **No front end:** Here the perturbed data is directly input to the ML classifier, i.e, $\Delta_0 = \left|\boldsymbol{w}^T \boldsymbol{e}\right|$. We use this scenario as a baseline to assess the efficacy of our defense.
   Assuming the adversary has knowledge of $\boldsymbol{w}$, the most effective attack would be in the direction orthogonal to the classifier's decision boundary, subject to the ℓ_∞ constraint:

$$\boldsymbol{e} = \epsilon \operatorname{sgn}(\boldsymbol{w}).$$

   This yields

$$\Delta_0 = \epsilon \left\|\boldsymbol{w}\right\|_1.$$

2. **Semi-white box attack:** In this scenario the defender employs the sparsifying front end, but the adversary designs the perturbation based on knowledge of $\boldsymbol{w}$ alone. Hence the perturbation remains

$$\boldsymbol{e}_{\mathrm{SW}} = \epsilon \operatorname{sgn}(\boldsymbol{w}).$$

Using (2), we get the impact of the attack as follows:

$$\Delta_{\mathrm{SW}} = \epsilon \big| \operatorname{sgn}(\boldsymbol{w}^T)\, \mathcal{P}_{\mathrm{K}}(\boldsymbol{w}, \boldsymbol{x}) \big|.$$

3. **White box attack:** Here the adversary has knowledge of both $\boldsymbol{w}$ and the front end, and designs perturbations accordingly. This results in the following optimization problem:

$$\begin{aligned} \max_{\boldsymbol{e}} \quad & \big| \boldsymbol{e}^T\, \mathcal{P}_{\mathrm{K}}(\boldsymbol{w}, \boldsymbol{x}) \big| \\ \text{s.t.} \quad & \|\boldsymbol{e}\|_\infty < \epsilon. \end{aligned}$$

The optimal perturbation is

$$\boldsymbol{e}_{\mathrm{W}} = \epsilon \operatorname{sgn}(\mathcal{P}_{\mathrm{K}}(\boldsymbol{w}, \boldsymbol{x})),$$

and its impact becomes

$$\Delta_{\mathrm{W}} = \epsilon \, \|\mathcal{P}_{\mathrm{K}}(\boldsymbol{w}, \boldsymbol{x})\|_1.$$

Thus, instead of aligning with $\boldsymbol{w}$, $\boldsymbol{e}_{\mathrm{W}}$ is aligned to the projection of $\boldsymbol{w}$ on the subspace that $\boldsymbol{x}$ lies in.

V. Analytical results

A. Characterizing the High SNR Regime

Proposition 1. *For sparsity level K, the sparsifying front end preserves the input coefficients if the following SNR condition holds:*

$$\mathrm{SNR} \triangleq \frac{\lambda}{\epsilon} > 2M,$$

where λ is the magnitude of the smallest non-zero entry of $\mathcal{H}_{\mathrm{K}}(\Psi^T x)$ and $M = \max_l \|\boldsymbol{\psi}_l\|_1$.

Proof. It is easy to see that (1) is equivalent to

$$\min_{i \in \mathcal{S}_{\mathrm{K}}(\boldsymbol{x})} \big| \boldsymbol{\psi}_i^T (\boldsymbol{x} + \boldsymbol{e}) \big| > \max_{j \notin \mathcal{S}_{\mathrm{K}}(\boldsymbol{x})} \big| \boldsymbol{\psi}_j^T (\boldsymbol{x} + \boldsymbol{e}) \big| = \max_{j \notin \mathcal{S}_{\mathrm{K}}(\boldsymbol{x})} \big| \boldsymbol{\psi}_j^T \boldsymbol{e} \big|,$$

where the equality follows from the definition of $\mathcal{S}_{\mathrm{K}}(\cdot)$. Denoting the optimal indices by i_0 and j_0, we use triangle inequality to obtain $|\boldsymbol{\psi}_{i_0}^T \boldsymbol{x}| > |\boldsymbol{\psi}_{i_0}^T \boldsymbol{e}| + |\boldsymbol{\psi}_{j_0}^T \boldsymbol{e}|$. The proposition follows by applying Hölder's inequality and using the ℓ_∞-bound on $\boldsymbol{e}$. □

Remarks.

1. The SNR condition is easier to satisfy for bases with sparser, or more localized, basis functions (smaller M). For example, we expect a wavelet basis to be better than a DCT basis.
2. When $\boldsymbol{x}$ is approximately K-sparse, choosing smaller K allows the SNR condition to hold for larger perturbations, but at the expense of higher signal perturbation. These must be traded off to optimize classification performance.

All of our subsequent analysis in this section is based on the assumption that the SNR condition in Proposition 1 holds. In this case, the sparsifying front end is signal-preserving, hence the output distortion can be quantified solely by analyzing its effect on the adversarial perturbation. In our experiments with MNIST data, we find that the SNR condition is approximately satisfied for the range of K that works most effectively (1-5% of the coefficients in a wavelet basis).

B. Ensemble Averaged Performance

We now provide an analysis that quantifies the robustness provided by sparsification over an ensemble of linear classifiers, by imposing a stochastic model for $\boldsymbol{w}$.

Assumption. For $\boldsymbol{w} = (w_1, \ldots, w_N)^T$, we model the $\{w_i,\, i = 1, \ldots, N\}$ as i.i.d., with zero mean and median: $\mathrm{E}[w_1] = 0$ and $\mathrm{E}[\operatorname{sgn}(w_1)] = 0$. Let $\mathrm{E}[|w_1|] = \mu$ and $\mathrm{E}\big[w_1^2\big] = \sigma^2$.

1) Semi-White Box Attack

Theorem 1. *As K approaches infinity, Δ_{SW}/K converges to μ in probability, i.e.*

$$\lim_{K \to \infty} \Pr\left(\left| \frac{\Delta_{\mathrm{SW}}}{K} - \mu \right| \le \delta \right) = 1 \quad \forall\, \delta > 0.$$

Remark. After sparsification, the impact Δ_{SW} of the adversarial perturbation scales linearly with the sparsity level K. Thus, the sparsifying front end provides an attenuation of K/N on the effect of the semi-white box adversarial attack.

Proof. Assuming without loss of generality that $\mathcal{S}_{\mathrm{K}}(\boldsymbol{x}) = \{1, 2, \ldots, K\}$, the output distortion can be written as $\Delta_{\mathrm{SW}} = |Z_K|,\ Z_K = \sum_{i=1}^K U_i V_i$, where

$$U_i = \sum_{m=1}^N \psi_i[m]\, w_m,\ V_i = \sum_{m=1}^N \psi_i[m]\, \operatorname{sgn}(w_m),\ \ i = 1, \ldots, K.$$

We now state the following lemma:

Lemma 1. *The mean and variance of Z_K are bounded by linear functions of K:*

$$\mathrm{E}(Z_K) = K\mu, \quad \operatorname{var}(Z_K) \le K\big(\sigma^2 + \mu^2\big).$$

Proof. We observe that for $i, j \in \{1, \ldots, K\}$, $\mathrm{E}[U_i V_i] = \mu$,

$$\begin{aligned} \operatorname{var}(U_i V_i) &= \sigma^2 + \mu^2 - 2\mu^2 \sum_{m=1}^N \psi_i^4[m], \\ \operatorname{cov}(U_i V_i, U_j V_j) &= -2\mu^2 \sum_{m=1}^N \psi_i^2[m]\, \psi_j^2[m], \quad i \ne j. \end{aligned}$$

Hence we get $\mathrm{E}[Z_K] = K\mu$, and

$$\begin{aligned} \operatorname{var}(Z_K) &= \sum_{i=1}^K \operatorname{var}(U_i V_i) - \sum_{\substack{i,j=1 \\ i \ne j}}^K \operatorname{cov}(U_i V_i, U_j V_j) \\ &= K\big(\sigma^2 + \mu^2\big) - 2\mu^2 \sum_{m=1}^N \sum_{i,j=1}^K \psi_i^2[m]\, \psi_j^2[m] \\ &\le K\big(\sigma^2 + \mu^2\big). \end{aligned}$$

□

We now apply Chebyshev's inequality to $Y_K = Z_K/K$, noting that $\mathrm{E}[Y_K] = \mu$ and $\mathrm{var}(Y_K) \leq \left(\sigma^2 + \mu^2\right)/K$:

$$\Pr(|Y_K - \mu| \leq \delta) \geq 1 - \frac{1}{K}\left(\frac{\sigma^2 + \mu^2}{\delta^2}\right) \quad \forall\, \delta \geq 0.$$

The theorem follows by applying the sandwich theorem to the above inequality as $K \to \infty$, observing that $|\Delta_{\mathrm{SW}}/K - \mu| = ||Y_K| - \mu| \leq |Y_K - \mu|$. □

2) White Box Attack

Lemma 2. *An upper bound on the white box attack distortion is given by*

$$\Delta_{\mathrm{W}} \leq \sum_{k=1}^{K} \left|\boldsymbol{\psi}_k^T \boldsymbol{w}\right| \|\boldsymbol{\psi}_k\|_1.$$

Proof.

$$\begin{aligned}\Delta_{\mathrm{W}} &= \sum_{i=1}^{N} \left| \sum_{k=1}^{K} \left(\boldsymbol{\psi}_k^T \boldsymbol{w}\right) \psi_k[i] \right| \\ &\leq \sum_{i=1}^{N} \sum_{k=1}^{K} \left|\boldsymbol{\psi}_k^T \boldsymbol{w}\right| |\psi_k[i]| = \sum_{k=1}^{K} \left|\boldsymbol{\psi}_k^T \boldsymbol{w}\right| \|\boldsymbol{\psi}_k\|_1.\end{aligned}$$

□

Remarks.

1. The upper bound is exact if the supports of the K selected basis functions do not overlap. In our MNIST experiments, this is approximately satisfied for the range of K that works most effectively (1-5% of the coefficients in a wavelet basis).
2. Since the upper bound has K terms, the distortion cannot grow slower than K. As stated in the following theorem, however, if the basis functions are "localized" with ℓ_1 norms that do not scale too fast with N, then the output distortion scales as $\mathcal{O}(K\,\mathrm{polylog}(N))$.

Theorem 2. *With high probability,*

$$\Delta_{\mathrm{W}} \leq \mathcal{O}(K\,\mathrm{polylog}(N)),$$

under the assumptions $\|\boldsymbol{\psi}_k\|_1 = \mathcal{O}(\log N)$, $\|\boldsymbol{\psi}_k\|_\infty = o(1)$ $\forall\, k \in \{1, 2, \ldots, K\}$, *and* $\|\boldsymbol{w}\|_\infty = \mathcal{O}(1)$. *Equivalently,*

$$\lim_{N\to\infty} \Pr(\Delta_{\mathrm{W}} \leq \mathcal{O}(K\,\mathrm{polylog}(N))) = 1.$$

Proof. Letting $Z_k = \boldsymbol{\psi}_k^T \boldsymbol{w}$, we first state the following lemma:

Lemma 3. $Z_k \to \mathcal{N}(0,\, \sigma^2)$ *in distribution.*

Proof. We show that we can apply Lindeberg's version of the central limit theorem, noting that $Z_k = \sum_{i=1}^{N} Y_i$, where $Y_i = \psi_k[i] w_i$ are independent random variables with $\mathrm{E}[Y_i] = 0$ and $\mathrm{var}(Y_i) = \sigma_i^2$, with $\sum_{i=1}^{N} \sigma_i^2 = \sigma^2$.

Now, given $\delta > 0$, we investigate the following quantity in order to check Lindeberg's condition:

$$L(\delta, N) = \frac{1}{\sigma^2} \sum_{i=1}^{N} \mathrm{E}\left[Y_i^2 \mathbb{1}_{\{|Y_i| > \delta\sigma\}}\right].$$

From the ℓ_∞ assumptions on $\boldsymbol{\psi}_k$ and $\boldsymbol{w}$, we observe that

$$\begin{aligned}\mathrm{E}\left[\psi_k^2[i] w_i^2 \mathbb{1}_{\{|Y_i| > \delta\sigma\}}\right] &\leq o^2(1)\, \mathcal{O}^2(1) \Pr\left(|Y_i| > \delta\sigma\right) \\ &= o^2(1)\, \mathcal{O}^2(1) \Pr\left(|w_i| > \frac{\delta\sigma}{o(1)}\right).\end{aligned}$$

Also note that $\forall\, \delta > 0,\ \exists\, M$ s.t. $\forall N > M$, $|w_i| < \delta\sigma/\, o(1)$ $\forall\, i \in \{1, \ldots, N\}$. Hence we get $\lim_{N\to\infty} L(\delta, N) = 0$, which is Lindeberg's condition. □

From Lemmas 2 and 3, we get

$$\begin{aligned}&\Pr(\Delta_{\mathrm{W}} > \delta) \leq \Pr\left(\sum_{k=1}^{K} |Z_k| \|\boldsymbol{\psi}_k\|_1 > \delta\right) \\ &\leq \Pr\left(\bigcup_{k=1}^{K} \left\{|Z_k| > \frac{\delta}{K\|\boldsymbol{\psi}_k\|_1}\right\}\right) \leq \sum_{k=1}^{K} \Pr\left(|Z_k| > \frac{\delta}{K\|\boldsymbol{\psi}_k\|_1}\right) \\ &= \sum_{k=1}^{K} 2Q\left(\frac{\delta}{\sigma K \|\boldsymbol{\psi}_k\|_1}\right) = 2KQ\left(\frac{\delta}{\sigma}\, \mathcal{O}\left(\frac{1}{K \log N}\right)\right),\end{aligned}$$

where $Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-t^2/2} dt$, and we have used the ℓ_1 assumption on $\boldsymbol{\psi}_k$ in the last step. The theorem follows by setting $\delta = \mathcal{O}(K\,\mathrm{polylog}(N))$ and applying the sandwich theorem as $N \to \infty$. □

VI. Experimental results

In this section we demonstrate the efficacy of sparsifying front ends on an inference task where our analysis directly applies: classification of digit pairs from the MNIST handwritten digit database [10] via linear SVM.[1]

A. Setup

We consider the task of discriminating between digits d_1 and d_2, where $d_1 \neq d_2 \in \{0, 1, \ldots 9\}$. The dataset of interest is $\mathcal{X} = \{\boldsymbol{x} : \mathcal{L}(\boldsymbol{x}) \in \{d_1, d_2\}\}$, where $\boldsymbol{x}$ denotes the images normalized to $[-1, 1]$ and $\mathcal{L}(\boldsymbol{x})$ the true labels. We divide $\mathcal{X}$ into training and test sets $\mathcal{X}_{tr}, \mathcal{X}_{te}$ in a 3:1 ratio.

We train a linear SVM classifier $f(\cdot)$ on $\mathcal{X}_{tr}$ and obtain class predictions $\mathcal{C}(\cdot)$ as follows:

$$f(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x} + b, \qquad \mathcal{C}(\boldsymbol{x}) = \begin{cases} d_1, & f(\boldsymbol{x}) < 0 \\ d_2, & f(\boldsymbol{x}) > 0. \end{cases}$$

In the scenario without front end, we consider the adversarial perturbation $\boldsymbol{e} = \epsilon\, \mathrm{sgn}(\boldsymbol{w})$ on $\mathcal{X}_{te}$, where the "direction" of the attack is opposite that of the correct class:

$$\bar{\boldsymbol{x}} = \begin{cases} \boldsymbol{x} + \boldsymbol{e}, & \mathcal{L}(\boldsymbol{x}) = d_1 \\ \boldsymbol{x} - \boldsymbol{e}, & \mathcal{L}(\boldsymbol{x}) = d_2 \end{cases} \qquad \forall\, \boldsymbol{x} \in \mathcal{X}_{te}.$$

In practice, the adversary usually only has access to $\mathcal{C}(\cdot)$ and not $\mathcal{L}(\cdot)$ for the test set. Hence this provides an upper bound for the classification error.

For the sparsifying front end, we use the CohenDaubechies-Feauveau 9/7 wavelet [11] and impose sparsity in the wavelet domain. We retrain the SVM with the sparsified $\mathcal{X}_{tr}$ for various values of $\rho = K/N$, and evaluate the impact of semi-white box and white box attacks on $\mathcal{X}_{te}$.

[1] Code is available at https://github.com/soorya19/sparsity-based-defenses/.

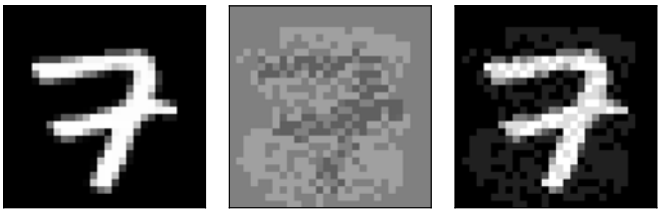

Fig. 3. Sample image before and after attack ($\epsilon = 0.25$). The adversarial perturbation causes digit 7 to be misclassified as 3.

B. Results

We begin with 3 vs. 7 classification. Without the front end, an attack with $\epsilon = 0.25$ completely overwhelms the classifier, reducing accuracy from 98.20% to 0%. Fig. 3 shows a sample image before and after attack.

Insertion of the sparsifying front end confers resiliency to attacks: at low values of ρ, accuracy is restored to near-baseline levels. The optimal value of ρ must trade off signal distortion versus perturbation attenuation. We find $\rho = 2\%$ to be the best choice for the 3 versus 7 scenario, and report on the accuracies obtained in Table I. Results for other digit pairs show a similar trend. Insertion of the front end greatly improves resilience to adversarial attacks. The optimal value of ρ lies between $1-5\%$, with $\rho = 2\%$ working well for all scenarios.

To give a concrete feel of the front end at work, Fig. 4 shows an example image, the attacked image, and the attacked image after sparsification.

Fig. 5 reports on accuracy as a function of ρ. At the low values of ρ that we are interested in, the white box attack is more damaging than the semi-white box attack. At higher ρ, a white box attack performs worse than the semi-white box attack: the high SNR condition in Proposition 1 is no longer satisfied, hence the white box attack is attacking the "wrong subspace." It is easy to devise iterative white box attacks that do better, but we do not discuss them here because the scenario of large ρ is not of practical interest, since it does not provide enough attenuation of the adversarial perturbation.

Acknowledgment

This work was supported in part by the National Science Foundation under grants CNS-1518812 and CCF-1755808, by Systems on Nanoscale Information fabriCs (SONIC), one of the six SRC STARnet Centers, sponsored by MARCO and DARPA, and by the UC Office of the President under grant No. LFR-18-548175.

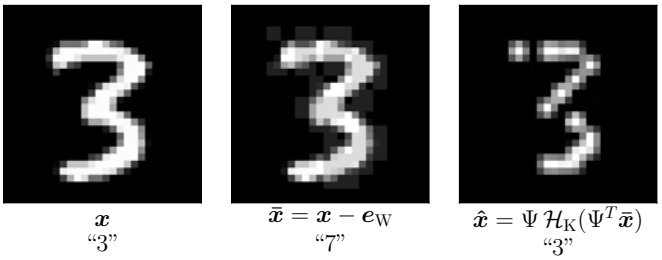

Fig. 4. Sample image after white box attack ($\epsilon = 0.25$) and front end (2%-sparse). The sparsified version of the attacked image resists misclassification.

TABLE I
BINARY CLASSIFICATION ACCURACIES (3 VS. 7)

	No front end	Sparsifying front end ($\rho = 2\%$)
No attack	98.20%	98.59%
Semi-white box attack	0%	97.31%
White box attack	0%	94.62%

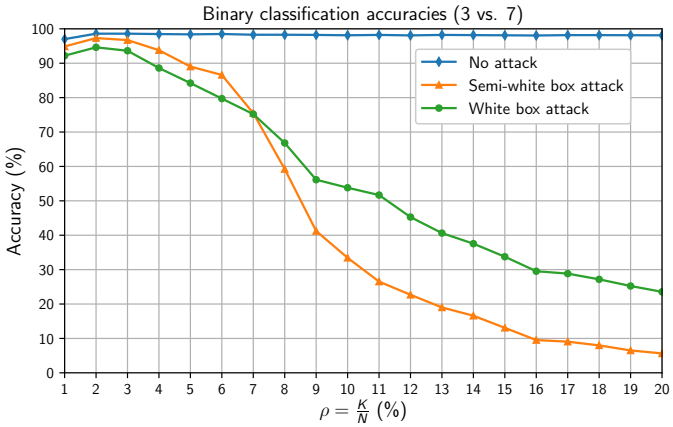

Fig. 5. Binary classification accuracies as a function of front end sparsity. All attacks use $\epsilon = 0.25$. Effectiveness of the front end decreases with increase in ρ.

References

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

[2] A. Fawzi, S.-M. Moosavi-Dezfooli, and P. Frossard, "The robustness of deep networks: A geometrical perspective," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 50–62, 2017.

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

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

[5] A. Fawzi, S.-M. Moosavi-Dezfooli, P. Frossard, and S. Soatto, "Classification regions of deep neural networks," *arXiv preprint arXiv:1705.09552*, 2017.

[6] B. Poole, S. Lahiri, M. Raghu, J. Sohl-Dickstein, and S. Ganguli, "Exponential expressivity in deep neural networks through transient chaos," in *Advances in Neural Information Processing Systems (NIPS)*, pp. 3360–3368, 2016.

[7] A. N. Bhagoji, D. Cullina, and P. Mittal, "Dimensionality reduction as a defense against evasion attacks on machine learning classifiers," *arXiv preprint arXiv:1704.02654*, 2017.

[8] N. Das, M. Shanbhogue, S.-T. Chen, F. Hohman, L. Chen, M. E. Kounavis, and D. H. Chau, "Keeping the bad guys out: Protecting and vaccinating deep learning with JPEG compression," *arXiv preprint arXiv:1705.02900*, 2017.

[9] A. Makhzani and B. Frey, "k-Sparse autoencoders," in *International Conference on Learning Representations (ICLR)*, 2014.

[10] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[11] A. Cohen, I. Daubechies, and J.-C. Feauveau, "Biorthogonal bases of compactly supported wavelets," *Communications on Pure and Applied Mathematics*, vol. 45, no. 5, pp. 485–560, 1992.