Compressive Recovery Defense: Defending Neural Networks Against ℓ_2, ℓ_∞, and ℓ_0 Norm Attacks

Jasjeet Dhaliwal
San Jose State University
jasjeet.dhaliwal@sjsu.edu

Kyle Hambrook
San Jose State University
kyle.hambrook@sjsu.edu

***Abstract*— We consider the problem of defending neural networks against adversarial inputs. In particular, we extend the framework introduced in [1] to defend neural networks against ℓ_2, ℓ_∞, and ℓ_0 norm attacks. We call this defense framework Compressive Recovery Defense (CRD) as it utilizes recovery algorithms from the theory of compressive sensing. For defending against ℓ_2-norm and ℓ_0-norm attacks, we use Basis Pursuit (BP) as the recovery algorithm and for the case of ℓ_∞-norm attacks, we utilize the Dantzig Selector (DS) with a novel constraint. For each recovery algorithm used, we provide rigorous recovery guarantees that do not depend on the noise generating mechanism and can therefore be utilized by CRD against any ℓ_2, ℓ_∞, or ℓ_0 norm attacks. Finally, we experimentally demonstrate that CRD is effective in defending neural networks against state of the art ℓ_2, ℓ_∞, and ℓ_0-norm attacks.**

I. Introduction

Signal measurements are often corrupted by noise. The theory of compressive sensing ([2]) allows us to retrieve the original signal from a corrupted measurement, under some structural assumptions on the measurement mechanism and the signal. Let us consider the class of machine learning problems where the inputs are compressible (i.e., approximately sparse) in some domain. For instance, images and audio signals are known to be compressible in their frequency domain and machine learning algorithms have been shown to perform exceedingly well on classification tasks that take such signals as input ([3], [4]). However, it was found in [5] that neural networks can be easily forced into making incorrect predictions by adding adversarial perturbations to their inputs; see also [6]–[9]. Further, the adversarial perturbations that led to incorrect predictions were shown to be imperceptible to human beings. For this class of machine learning tasks, we show how to approximately recover original inputs from adversarial inputs and defend neural networks.

We explain the idea behind the CRD framework in the context of an image classifier. Let $x \in \mathbb{C}^n$ be a (flattened) image vector we wish to classify. However, suppose an adversary perturbs x with a noise vector $e \in \mathbb{C}^n$ such that we observe $y = x + e$, where x and e are unknown to us. Let $F \in \mathbb{C}^{n\times n}$ be the Discrete Fourier Transform (DFT) matrix. The Fourier coefficients of x are $\hat{x} = Fx$. We can therefore write the observed input y as:

$$y = F^{-1}\hat{x} + e \tag{1}$$

It is well-known that natural images are approximately sparse in the frequency domain, so we expect that $\hat{x}$ is approximately sparse (meaning roughly that most of the entries of $\hat{x}$ are very small). If $\|e\|_p \leq \eta$ with η small (as in a ℓ_p-attack), then we can apply an appropriate sparse recovery algorithm, with y and F^{-1} as input, to recover a good approximation $x^\#$ to $\hat{x}$. Since F is unitary, $F^{-1}x^\#$ will be a good approximation (i.e., reconstruction) of $x = F^{-1}\hat{x}$ as long as $x^\#$ is a good approximation to $\hat{x}$. If $x^\#$ is indeed a good approximation to $\hat{x}$, we can feed $F^{-1}x^\#$ into the classifier and expect to get the same classification as we would have for x.

Note that the same framework can be applied with audio signals or other types of data instead of images. Moreover, the DFT can be replaced by any unitary transformation F for which $\hat{x} = Fx$ is approximately sparse. For example, F may be the Cosine Transform, Sine Transform, Hadamard Transform, or another wavelet transform.

Novel Contributions. The authors of [1] introduced this defense framework for the case of ℓ_0-attacks with Iterative Hard Thresholding (IHT) as the recovery algorithm. We make the following novel extensions to this framework:

- We extend the framework to handle ℓ_2 and ℓ_∞ norm attacks. introduce the Modified Dantzig Selector (MDS) which uses a novel constraint to provide better recovery for ℓ_∞ norm attacks.
- We extend the ℓ_0-norm defense introduced in [1] to defend against a larger adversarial noise budget.

Structure. We cover related work in Section II and then introduce notation, describe the recovery algorithms used, and state formal recovery guarantees in Section III. In Section IV, we experimentally demonstrate the performance of the CRD framework on CIFAR-10, MNIST, and Fashion-MNIST datasets against state of the art ℓ_2, ℓ_∞, and ℓ_0-norm attacks. We conclude the paper in Section V and provide proofs for the theorems in Section VI.

II. Related Work

The authors of [1] introduced the CRD framework which inspired this work. They utilized Iterative Hard Thresholding (IHT) as a recovery algorithm and provided guarantees for ℓ_0 norm attacks. We extend this framework to handle ℓ_2 and ℓ_∞ attacks and a larger attack budget for ℓ_0 norm attacks. We explain this in more detail in Section III-B.

Other works that provide guarantees against adversarial inputs include ([10]) and ([11]) where the authors regularize the Lipschitz constant of a network and lower bound the

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

perturbation required to change the classifier decision. The authors of [12] use robust optimization to train the network on adversarial inputs. A similar approach to ([12]) is ([13]) in which the authors use robust optimization to lower bound the adversarial perturbations on the training data required to cause misclassification. In [14], the authors use techniques from Differential Privacy ([15]) and augment the training procedure to improve robustness to adversarial inputs. The authors of [16] add i.i.d. Gaussian noise to the input and provide guarantees on classifier predictions for ℓ_2-norm bounded attack vectors.

Most defenses against adversarial inputs do not come with theoretical guarantees. Instead, a large body of research has focused on finding practical ways to improve robustness by either augmenting the training data ([7]), using adversarial inputs from various networks ([17]), or by transforming the input ([18]). For instance, [19] introduces the Projected Gradient Descent (PGD) defense that uses robust optimization based adversarial training. However, the effectiveness of their approach is determined by the amount and quality of training data available and its similarity to the distribution of the test data. A transformation based approach is ([20]). Here, the authors use Generative Adversarial Networks (GANs) to estimate the distribution of the training data and, during inference, use a GAN to reconstruct the input while removing adversarial noise. Other input transformation approaches include [21], where the authors randomly replace coordinates of the input vector with neighboring coordinates. Similarly, [22] use random resizing and padding to remove the effects of adversarial noise.

The field of compressive sensing was essentially initiated with the work of [23] and [24] in which the authors show rigorously how to recover sparse signals using only a small number of measurements with the choice of a random matrix. Some of the earlier work in extending compressive sensing to perform stable recovery with deterministic matrices was done by [25] and [2], where a sufficient condition for recovery was satisfaction of a restricted isometry hypothesis. [26] introduced IHT as an algorithm to recover sparse signals which was later modified in [27] to reduce the search space as long as the sparsity was structured. The standard DS algorithm was introduced in [28] in order to perform stable recovery in the presence of ℓ_∞ noise.

III. Compressive Recovery Defense

Notation. The ℓ_0-quasinorm of $x \in \mathbb{C}^N$, denoted $\|x\|_0$, is defined to be the number of non-zero entries of x, i.e. $\|x\|_0 = \text{card}(\text{supp}(x))$. We say that x is k-sparse if $\|x\|_0 \leq k$. We use $x_{h(k)}$ to denote a k-sparse vector in $\mathbb{C}^N$ consisting of the k largest (in absolute value) entries of x with all other entries zero. For example, if $x = [4, 5, -9, 1]^T$ then $x_{h(2)} = [0, 5, -9, 0]^T$. Note that $x_{h(k)}$ may not be uniquely defined. In contexts where a unique meaning for $x_{h(k)}$ is needed, we can choose $x_{h(k)}$ out of all possible candidates according to a predefined rule (such as the lexicographic order). We also define $x_{t(k)} = x - x_{h(k)}$. If $x = [x_1, x_2]^T \in \mathbb{C}^{2n}$ with $x_1, x_2 \in \mathbb{C}^n$, and if x_1 is k-sparse and x_2 is t-sparse, then x is called (k, t)-sparse. We define $x_{h(k,t)} = [(x_1)_{h(k)}, (x_2)_{h(t)}]^T$, which is a (k, t)-sparse vector in $\mathbb{C}^{2n}$.

A. Recovery Algorithms

Since the success of CRD depends on recovering a good approximation to $\hat{x}$ in (1), we select the recovery algorithm that provides the best recovery guarantees based on the type of noise used [1]. For ℓ_2 or ℓ_0 noise, we use Basis Pursuit (Algorithm 1) and for ℓ_∞ noise we use Dantzig Selector with a novel constraint (Algorithm 2).

To motivate the following algorithms, note sparse recovery (without noise) is the optimization $\arg\min_{z \in \mathbb{C}^N} \|z\|_0$ subject to $Az = y$. This is non-convex and NP-hard, and it remains so if $\|z\|_0$ is approximated by $\|z\|_q$ for $0 < q < 1$. Taking $q > 1$ gives a convex, P-time problem, but the solution need not be sparse. However, if $q = 1$, the problem is convex, P-time, and gives sparse solutions. (Details: [29, p.55-62],[30]–[32]).

Algorithm 1 Basis Pursuit: BP(y, A, η)

Input: $y \in \mathbb{C}^m$, where $y = A\hat{x} + e$, $A \in \mathbb{C}^{m \times N}$, and η such that $||e||_2 \leq \eta$

Output: $x^\# \leftarrow \arg\min_{z \in \mathbb{C}^N} ||z||_1$ subject to$||Az - y||_2 \leq \eta$

For details, see p.55-62,77 in [29]). The problem becomes convex and tractable for $q = 1$ and has been shown to provide sparse solutions; see [30]–[32].

For ℓ_2-norm noise, BP is applied with $A = F^{-1}$, a unitary matrix. As unitary matrices are isometries in ℓ_2 norm, BP provides good recovery guarantees for such matrices since they satisfy the robust null space property (Definition 4). Also, since the noise is bounded in ℓ_2 norm and since the solution to BP minimizes the error in ℓ_2 norm, BP proves to be a very good candidate for recovery.

For ℓ_0-norm noise, where e is t-sparse, the approach is only slightly different. We set $A = [F^{-1}, I]$ and write

$$y = F^{-1}\hat{x}_{h(k)} + e + F^{-1}\hat{x}_{t(k)} = A[\hat{x}_{h(k)}, e]^T + F^{-1}\hat{x}_{t(k)}$$

so that $[\hat{x}_{h(k)}, e]^T$ is a (k, t)-sparse vector that we can recover using Algorithm 1. We utilize BP for ℓ_0 attacks because it provides recovery guarantees for larger values of k and t than IHT which is used in [1]. For instance, in the case of MNIST and Fashion-MNIST, using IHT would allow us to set $k = 4$ and $t = 3$, whereas BP (Theorem 3) allows us to set $k = 8$ and $t = 8$.

Algorithm 2 Modified Dantzig Selector: MDS(y, A, η)

Input: $y \in \mathbb{C}^m$, where $y = A\hat{x} + e$, $A \in \mathbb{C}^{m \times N}$, and η such that $||e||_\infty \leq \eta$

Output: $x^\# \leftarrow \arg\min_{z \in \mathbb{C}^N} \|z\|_1$ subject to $\|A^*(Az - y)\|_\infty \leq \sqrt{n}\eta, \quad \|Az - y\|_\infty \leq \eta$

[1] An interesting follow up problem is choosing a recovery algorithm when the type of noise is not known a priori. In practice, inputs are normalized to lie within some range $[a, b]$ (for instance $[0, 1]$), thus the the attacker is still bounded in ℓ_2 norm. Thus, Algorithm 1 is a viable candidate for recovery. We leave a deeper analysis for future investigation.

We utilize MDS for ℓ_∞ norm attacks. The standard Dantzig Selector algorithm does not have the additional constraint $\|Az - y\|_\infty \leq \eta$. MDS includes this constraint for the following reason. In our application, we set $A = F^{-1}$ and we want the reconstruction $Ax^\#$ to be close to the original image x, so that they are classified identically. Thus, we want to the search space for $x^\#$ to be restricted to those $z \in \mathbb{C}^N$ such that $\|Az - x\|_\infty$ is small. Note, for any $z \in \mathbb{C}^N$, $\|Az - x\|_\infty \leq \|Az - y\|_\infty + \|x - y\|_\infty$. In an ℓ_∞-attack, $\|x - y\|_\infty = \|e\|_\infty$ is already small. Thus it suffices to require $\|Az - y\|_\infty$ is small. We experimentally illustrate the improvement in reconstruction due to the additional constraint in Section IV-B (Figure 2, Table II).

Remarks on Reverse-Engineered Attacks. In the case of Algorithm 1 and Algorithm 2, the minimization problems can be posed as semi-definite programming problems. If solved with interior point methods, one can use random initialization of the central path parameter and add randomness to the stopping criterion. Therefore, in addition to being non-differentiable, recovery is also non-deterministic and we expect that it would be non-trivial to create a successful reverse-engineered attack. However, we are aware that there are powerful reverse-engineered attacks designed for black-box settings ([33]) and for defenses relying on non-differentiability and randomness ([34]). We do not investigate reverse-engineered attacks against CRD in this work, but intend to do so in future work.

B. Recovery Guarantees

We now state the formal recovery guarantees based on the type of noise used by an attacker, i.e. ℓ_2, ℓ_∞, and ℓ_0 norm bounded noise.

Theorem 1 (ℓ_2-norm noise)**.** *If $\|e\|_2 \leq \eta$, then for $x^\# = BP(y, F^{-1}, \eta)$, we have the error bounds*

$$\|x^\# - \hat{x}\|_1 \leq 2\left(\|\hat{x}_{t(k)}\|_1 + 2\sqrt{k}\eta\right) \tag{2}$$

$$\|x^\# - \hat{x}\|_2 \leq \frac{2}{\sqrt{k}}\|\hat{x}_{t(k)}\|_1 + 6\eta \tag{3}$$

Theorem 2 (ℓ_∞-norm noise)**.** *If $\|e\|_\infty \leq \eta$, then for $x^\# = MDS(y, F^{-1}, \eta)$, we have the error bounds*

$$\|x^\# - \hat{x}\|_1 \leq 2\left(\|\hat{x}_{t(k)}\|_1 + 2k\sqrt{n}\eta\right) \tag{4}$$

$$\|x^\# - \hat{x}\|_2 \leq \frac{2}{\sqrt{k}}\|\hat{x}_{t(k)}\|_1 + 6\sqrt{kn}\eta \tag{5}$$

To interpret these results, first note that since F is an isometry, $\|x^\# - \hat{x}\|_2 = \|Fx^\# - F\hat{x}\|_2$. Thus the results of Theorem 1 and Theorem 2 also bound the norm difference of the original image $x = F\hat{x}$ and the reconstructed image $Fx^\#$, where $x^\#$ has no sparsity guarantees. Therefore, the inequalities indicate how confident we should be that the CRD scheme will be able to recover the correct class of the original image, and thus defend the classifier from the adversarial attack.

Note that the recovery guarantees decay with the sparsity of the vector $\hat{x}$. Theorem 1 allows us to recover sparse vectors with error that depends on the magnitude of its smallest coefficients $\hat{x}_{t(k)}$. Thus for approximately sparse vectors, CRD provides good recovery guarantees which consequently lead to better classifier performance on the recovered images. Theorem 2 provides similar guarantees when the noise is bounded in ℓ_∞-norm. Observe also that the results of Theorem 2 incur a factor of $\sqrt{n}$ in the error bounds due to the constraint $\|A^*(Az - y)\|_\infty \leq \sqrt{n}\eta$ in Algorithm 2 which is required to prove the robust null space property (refer to Section VI for details). This weaker guarantee can also be expected as bounding the ℓ_∞ norm of the noise vector is a very weak constraint.

Finally, we provide a novel result that extends the work of [1] for the case of ℓ_0 norm attacks by providing guarantees for a larger attack budget (i.e. larger values of k and t) than the main theorem of [1].

Theorem 3 (ℓ_0-norm noise)**.** *Assume $|F_{ij}|^2 \leq \frac{c}{n}$. Define $\delta_{k,t} = \sqrt{\frac{ckt}{n}}, \beta = \sqrt{\frac{\max\{k,t\}c}{n}}, \theta = \frac{\sqrt{k+t}}{(1-\delta_{k,t})}\beta, \tau = \frac{\sqrt{1+\delta_{k,t}}}{1-\delta_{k,t}}$.*

If $0 < \delta_{k,t} < 1$ and $0 < \theta < 1$, then for $x^\# = BP(y, [F^{-1}, I], \|\hat{x}_{t(k)}\|_2)$, we have the error bound

$$\|\hat{x}^\# - \hat{x}_{h(k)}\|_2 \leq \left(\frac{2\tau\sqrt{k+t}}{1-\theta}\left(1 + \frac{\beta}{1-\delta_{k,t}}\right) + 2\tau\right)\|\hat{x}_{t(k)}\|_2 \tag{6}$$

where we write $x^\# = [\hat{x}^\#, e^\#]^T \in \mathbb{C}^{2n}$ with $\hat{x}^\#, e^\# \in \mathbb{C}^n$.

IV. Experiments

All of our experiments are conducted on CIFAR-10 ([35]), MNIST ([36]), and Fashion-MNIST ([37]) datasets with pixel values of each image normalized to lie in $[0, 1]$. Each experiment is conducted on a set of 1000 points sampled uniformly at random from the test set of the respective dataset.

For every experiment, we use the Discrete Cosine Transform (DCT) and the Inverse Discrete Cosine Transform (IDCT) denoted by the matrices $F \in \mathbb{R}^{n\times n}$ and $F^T \in \mathbb{R}^{n\times n}$ respectively. That is, for an adversarial image $y \in \mathbb{R}^{\sqrt{n}\times\sqrt{n}}$, such that, $y = x + e$, we let $\hat{x} = Fx$, and $x = F^T\hat{x}$, where $x, \hat{x} \in \mathbb{R}^n$ and $e \in \mathbb{R}^n$ is the noise vector. For an adversarial image $y \in \mathbb{R}^{\sqrt{n}\times\sqrt{n}\times c}$, that contains c channels, we perform recovery on each channel independently by considering $y_m = x_m + e_m$, where $\hat{x}_m = Fx_m, x_m = F^T\hat{x}_m$ for $m = 1, \ldots, c$. The value k denotes the number of largest (in absolute value) DCT coefficients used for reconstruction of each channel. We set $k = 40$ for MNIST and F-MNIST and $k = 500$ for CIFAR-10. We implement Algorithm 1 and Algorithm 2 using the open source library CVXPY ([38]).

For CIFAR-10, we use the network architecture of [39] while the network architecture for MNIST and Fashion-MNIST datasets is provided in Table IV of the Appendix. We train our networks using the Adam optimizer for CIFAR-10 and the AdaDelta optimizer for MNIST and Fashion-MNIST. In both cases, we use a cross-entropy loss function.

We now describe the training and testing procedure for CRD. For each training image x, we compute $\hat{x}_{h(k)} = (Fx)_{h(k)}$, and then compute the compressed the image $x' =$

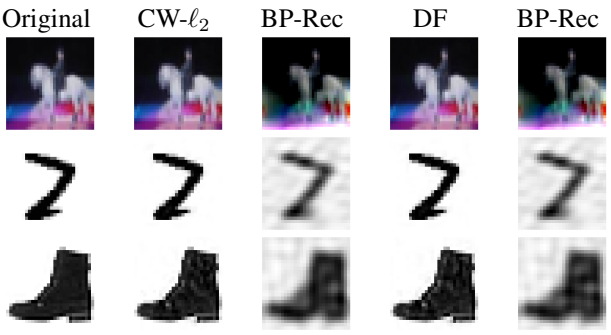

Fig. 1. Reconstruction quality of images against ℓ_2 attacks using Algorithm 1. The first column shows the original images, while the adversarial images are shown in the second and fourth column. The reconstructions are shown in columns three and five.

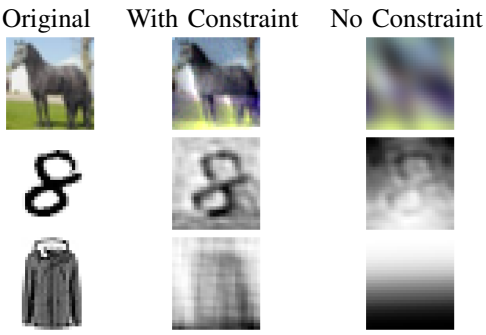

Fig. 2. Comparison of images reconstructed using Algorithm 2 (With Constraint) with images reconstructed using DS without the additional constraint (No Constraint).

$F^{-1}\hat{x}_{h(k)}$. We then add both x and x' to the training set and train the network in the usual way. Given a (potentially adversarial) test image y, we first use a sparse recovery algorithm to compute an approximation $x^{\#}$ to $\hat{x}$, then we compute the reconstructed image $y' = F^{-1}x^{\#}$ and feed it into the network for classification. The code to reproduce our experiments is available here: https://github.com/ijcnnanonymous2020/compressive_recovery_defense.

A. Defense against ℓ_2-norm attacks

We use the CW ℓ_2-norm attack ([9]) and the Deepfool attack ([40]) as they are widely considered state of the art. We note that Theorem 1 does not impose any restrictions on k and therefore the guarantees of equations (2) and (3) are applicable for recovery in all experiments of this section.

We test the performance of CRD in two ways: a) reconstruction quality, and b) network performance on reconstructed images. To analyze reconstruction quality of Algorithm 1, for each test image, we first create an adversarial image and then use Algorithm 1 to recover its largest k coefficients. We then perform the IDCT on these recovered co-efficients to generate reconstructed images. We illustrate reconstruction on a randomly selected image from the test set in Figure 1.

In order to check whether this high quality reconstruction also leads to improved performance in network accuracy, we test each network on reconstructed images using Algorithm 1. We report the results in Table I and note that Algorithm 1 provides a substantial improvement in network accuracy for each dataset and each attack method used.

For comparison, we implement the PGD defense framework of [19] for MNIST and Fashion-MNIST and see that CRD outperforms PGD. Due to time and resource constraints we do not report PGD results for CIFAR-10 as we were unable to get the network to converge to an accuracy over 70% for non-adversarial test samples. Since PGD is computationally very expensive [41], [42] and minimizes network loss on adversarial samples, training a network that performs well on adversarial and non-adversarial samples is highly non-trivial. We note that CRD does not suffer from either of these drawbacks as network training is decoupled from the CRD defense.

B. Defense against ℓ_∞-norm attacks

For ℓ_∞-norm bounded attacks, we use the BIM attack ([43]) as it is state of the art and allows us to control the ℓ_∞-norm of the attack vector explicitly [2]. Therefore, we limit our experimental analysis to the BIM attack. Note that for any attack vector e, $\|e\|_2 \leq \sqrt{n}\|e\|_\infty$ hence allowing ℓ_∞-norm attacks to create attack vectors with large ℓ_2-norm. Therefore, we could expect reconstruction quality and network accuracy to be lower when compared to ℓ_2-norm attacks.

In Figure 2, we compare the reconstruction quality of images reconstructed with Algorithm 2 to those reconstructed using DS without the additional constraint. It can bee seen that images reconstructed using DS without the additional constraint may not produce meaningful images. This is also reflected in Table II, which shows that the accuracy of the network is roughly random on images reconstructed without the additional constraint. We show examples of original images, adversarial images, and their reconstructions using Algorithm 2 in Figure 3. Finally, we report the network performance on reconstructed inputs using Algorithm 2 in Table II and also compare this to the performance on inputs reconstructed using DS without the additional constraint. We also report the results of the PGD defense of [19] and note that PGD outperforms CRD against the BIM attack. This can be expected as PGD training uses a very similar method to BIM in construction adversarial examples used for training.

C. Defense against ℓ_0-norm attacks

We test CRD against the CW ℓ_0-norm attack and JSMA. We find that even when t is much larger than the hypotheses of Theorem 3, we find that Algorithm 1 is still able to defend the network. We hypothesize that this may be related to the behavior of the RIP of a matrix for "most" vectors as opposed to the RIP for all vectors, and leave a more rigorous analysis for a follow up work.

Fig 4 shows the reconstruction quality of the images and the improvement in network performance on reconstructed

[2]We note that while the CW ℓ_∞-norm attack ([9]) has the ability to create attack vectors with ℓ_∞-norm less than or equal to BIM, it is computationally expensive and also does not allow one to pre-specify a value for the ℓ_∞-norm of an attack vector.

Dataset	Orig.	C&W ℓ_2				Deepfool			
	Acc.	$\ell_{2_{\text{avg}}}$	Acc.	BP Acc.	PGD Acc.	$\ell_{2_{\text{avg}}}$	Acc.	BP Acc.	PGD Acc.
CIFAR-10	84.9%	0.12	8.7%	72.3%	-	0.11	7.7%	71.6%	-
MNIST	99.17%	1.35	0.9%	92.4%	83.7%	1.72	1.1	90.7%	6.4%
Fashion-MNIST	90.3%	0.61	5.4%	78.3%	75.9%	0.63	5.5 %	76.4%	25.6%

TABLE I

THE $\ell_{2_{\text{AVG}}}$ COLUMN LISTS THE AVERAGE ℓ_2-NORM OF THE ATTACK VECTOR. THE ORIG. ACC COLUMN LISTS THE ACCURACY OF THE NETWORK ON ORIGINAL TEST INPUTS, WHILE THE ACC. COLUMNS UNDER C&W ℓ_2 AND DF COLUMNS REPORT NETWORK ACCURACY ON ADVERSARIAL INPUTS. BP ACC. COLUMNS LISTS THE ACCURACY OF THE NETWORK ON INPUTS RECONSTRUCTED USING ALGORITHM 1.PGD ACC. SHOWS ACCURACY OF THE DEFENSE IN [19].

Dataset	Orig.	BIM				
	Acc.	$\ell_{\infty_{\text{avg}}}$	Acc.	MDS Acc.	DS Acc.	PGD Acc.
CIFAR-10	84.9%	0.015	7.4%	49.4%	17.6%	-
MNIST	99.17%	0.15	4.9%	74.7%	10%	95.8%
Fashion-MNIST	90.3%	0.15	5.3%	57.5%	11.1%	77.3%

TABLE II

THE $\ell_{\infty_{\text{AVG}}}$ COLUMN LISTS THE ℓ_∞-NORM OF EACH ATTACK VECTOR, ORIG. ACC. AND BIM ACC. COLUMNS LIST THE ACCURACY OF THE NETWORK ON THE ORIGINAL AND ADVERSARIAL INPUTS RESPECTIVELY, AND THE MDS ACC. COLUMN LISTS THE ACCURACY OF THE NETWORK ON INPUTS RECONSTRUCTED USING ALGORITHM 2. WE ALSO SHOW ACCURACY OF THE NETWORK ON IMAGES RECONSTRUCTED WITH DS (WITHOUT THE ADDITIONAL CONSTRAINT) IN THE DS ACC. COLUMN. PGD ACC. SHOWS ACCURACY OF THE DEFENSE IN [19].

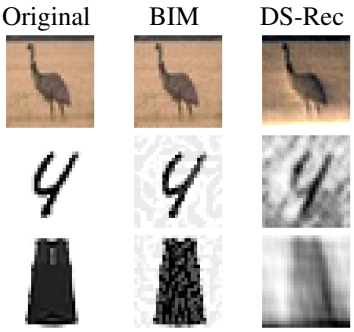

Fig. 3. Reconstruction quality of images using Algorithm 2. The first column shows the original images, while the second columns shows adversarial images and the third columns shows reconstructions using Algorithm 2 respectively.

adversarial images using CRD is reported in Table III. It can also be seen that CRD outperforms PGD for both ℓ_0 attacks.

V. CONCLUSION

We provided recovery guarantees for corrupted signals in the case of ℓ_2, ℓ_∞ and ℓ_0-norm bounded noise. We were able to utilize these results in CRD and improve the performance of neural networks substantially in the case of ℓ_2, ℓ_∞ and ℓ_0 norm bounded noise. In particular, we utilized the guarantees of Theorem 1 and Theorem 2 to extend the defense framework of [1] to defend neural networks against ℓ_2, ℓ_∞ and ℓ_0 norm attacks.

VI. APPENDIX

Notation. Let x be a vector in $\mathbb{C}^N$. Let $S \subseteq \{1, \dots, N\}$ and $\overline{S} = \{1, \dots, N\} \backslash S$. The cardinality of S is $|S|$. If $A \in \mathbb{C}^{m \times N}$ is a matrix, then $A_S \in \mathbb{C}^{m \times |S|}$ is the column submatrix of A consisting of the columns indexed by S. We denote by x_S either the sub-vector in $\mathbb{C}^S$ consisting of the entries indexed by S or the vector in $\mathbb{C}^N$ that is formed by starting with x and setting the entries indexed by $\overline{S}$ to zero. For example, if $x = [4, 5, -9, 1]^T$ and $S = \{1, 3\}$, then x_S is either $[4, -9]^T$ or $[4, 0, -9, 0]^T$. It will always be clear from context which meaning is intended. Note that, under the second meaning, $x_{\overline{S}} = x - x_S$. The support of x, denoted by $\text{supp}(x)$, is the set of indices of the non-zero entries of x, i.e., $\text{supp}(x) = \{i \in \{1, \dots, N\} : x_i \neq 0\}$.

Definition 4. *The matrix $A \in C^{m \times N}$ satisfies the ℓ_q robust null space property of order s with constants $0 < \rho < 1$, $\tau > 0$ and norm $\|\cdot\|$ if for every set $S \subseteq [N]$ with card$(S) \leq s$ and for every $v \in \mathbb{C}^N$ we have*

$$\|v_S\|_q \leq \frac{1}{s^{1-1/q}} \rho \|v_{\overline{S}}\|_1 + \tau \|Av\|$$

Note that if $q = 1$ then this is simply the robust null space property.

We now focus on proving Theorem 1. In order to do so, we will need some lemmas that will be used in the main proof.

Lemma 5. *If a matrix $A \in \mathbb{C}^{m \times N}$ satisfies the ℓ_2 robust null space property for $S \subset [N]$, with card$(S) = s$, then it satisfies the ℓ_1 robust null space property for S with constants $0 < \rho < 1, \tau' := \tau\sqrt{s} > 0$.*

Proof. For any $v \in \mathbb{C}^N$, $\|v_S\|_2 \leq \frac{\rho}{\sqrt{s}} \|v_{\bar{S}}\|_1 + \tau \|Av\|$. Then, using the fact that $\|v_S\|_1 \leq \sqrt{s} \|v_S\|_2$, we get:$\|v_S\|_1 \leq \rho \|v_{\bar{S}}\|_1 + \tau\sqrt{s} \|Av\|$.

□

Lemma 6 (Theorem 4.20 in [29])**.** *If a matrix $A \in \mathbb{C}^{m \times N}$ satisfies the ℓ_1 robust null space property (with respect to $\|.\|$) and for $0 < \rho < 1$ and $\tau > 0$ for $S \subset [N]$, then:*

$$\|z - x\|_1 \leq \frac{1 + \rho}{1 - \rho} (\|z\|_1 - \|x\|_1 + 2\|x_{\bar{S}}\|_1) + \frac{2\tau}{1 - \rho} \|A(z - x)\|$$

for all $z, x \in \mathbb{C}^N$.

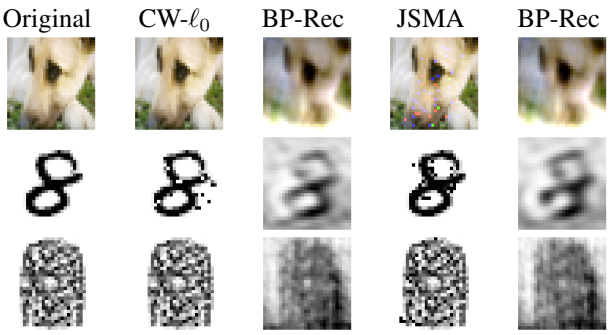

Fig. 4. Reconstruction quality of images using Algorithm 1 against ℓ_0 attacks. The first column shows randomly selected original images from the test set, while the second and fourth column show the adversarial images. Reconstructions using BP are labeled BP-Rec. We show reconstructions in columns three, and five.

Dataset	**Orig.**	**C&W ℓ_0**				**JSMA**			
	Acc.	t_{avg}	Acc.	BP Acc.	PGD Acc.	t_{avg}	Acc.	BP Acc.	PGD Acc.
CIFAR-10	84.9%	18	8.7%	67.0%	-	34	2.7%	67.3%	-
MNIST	98.8%	15	0.9%	55.9%	14.1%	17	56.5 %	67.4%	93.5%
Fashion-MNIST	91.8%	16	5.27%	71.4%	75.1%	17	62.6 %	72.0%	76.6%

TABLE III

THE t_{AVG} COLUMN LISTS THE AVERAGE ADVERSARIAL NOISE BUDGET FOR EACH ATTACK. THE ORIG. ACC COLUMN LISTS THE ACCURACY OF THE NETWORK ON ORIGINAL TEST INPUTS, THE ACC. COLUMNS UNDER C&W ℓ_0 AND JSMA LIST NETWORK ACCURACY ON ADVERSARIAL INPUTS. THE BP ACC. COLUMN LISTS THE ACCURACY OF THE NETWORK ON INPUTS THAT HAVE BEEN CORRECTED USING BP.PGD ACC. SHOWS ACCURACY OF THE DEFENSE IN [19].

Lemma 7 (Proposition 2.3 in [29])**.** *For any $p > q > 0$ and $x \in \mathbb{C}^n$,*

$$\inf_{z \in M_k} \|x - z\|_p \leq \frac{1}{(k)^{\frac{1}{q}-\frac{1}{p}}} \|x\|_q$$

Proof of Theorem 1. Let $0 < \rho < 1$ be arbitrary. Since F is a unitary matrix, for any $S \subseteq [n]$ and $v \in \mathbb{C}^n$, we have

$$\|v_S\|_2 \leq \frac{\rho}{\sqrt{k}} \|v_{\overline{S}}\|_1 + \tau \|v\|_2 = \frac{\rho}{\sqrt{k}} \|v_{\overline{S}}\|_1 + \tau \|Fv\|_2 \quad (7)$$

where $\tau = 1$. Now let $S \subseteq [n]$ such that $\text{card}(S) \leq k$. Then, F satisfies the ℓ_2 robust null space property for S. Next, using Lemma 5 we get $\|v_S\|_1 \leq \rho \|v_{\overline{S}}\|_1 + \tau\sqrt{k}\|Fv\|_2$ for all $v \in \mathbb{C}^n$. Now let $x^\# = \text{BP}(y, F, \eta)$. Then we know $\|x^\#\|_1 \leq \|\hat{x}\|_1$. So, by fixing $S \subseteq [n]$ to be the support of $\hat{x}_{h(k)}$ and using Lemma 6 and the fact that $\|F(x^\# - \hat{x})\|_2 \leq 2\|e\|_2 \leq 2\eta$, we get:

$$\begin{aligned}
\|x^\# - \hat{x}\|_1 &\leq \frac{1+\rho}{1-\rho}(\|x^\#\|_1 - \|\hat{x}\|_1 + 2\|\hat{x}_{t(k)}\|_1) \\
&\quad + \frac{2\tau\sqrt{k}}{1-\rho}\|F(x^\# - \hat{x})\|_2 \\
&\leq \frac{1+\rho}{1-\rho}\left(2\|\hat{x}_{t(k)}\|_1\right) + \frac{4\tau\sqrt{k}}{1-\rho}\eta
\end{aligned}$$

Letting $\rho \to 0$ and recalling that $\tau = 1$ gives (2). Now let S be the support of $(x^\# - \hat{x})_{h(k)}$. Note $\|(x^\# - \hat{x})_{\overline{S}}\|_2 = \inf_{z \in M_k} \|(x^\# - \hat{x}) - z\|_2$. Then, using Lemma 7 and (7), we see that

$$\begin{aligned}
\|x^\# - \hat{x}\|_2 &\leq \|(x^\# - \hat{x})_{\overline{S}}\|_2 + \|(x^\# - \hat{x})_S\|_2 \\
&\leq \frac{1}{\sqrt{k}}\|(x^\# - \hat{x})\|_1 + \frac{\rho}{\sqrt{k}}\|(x^\# - \hat{x})_{\overline{S}}\|_1 \\
&\quad + \tau\|F(x^\# - x)\|_2 \\
&\leq \frac{1+\rho}{\sqrt{k}}\|(x^\# - \hat{x})\|_1 + 2\tau\eta \\
&\leq \frac{(1+\rho)^2}{\sqrt{k}(1-\rho)}\left(2\|\hat{x}_{t(k)}\|_1\right) + \frac{4\tau(1+\rho)}{(1-\rho)}\eta + 2\tau\eta
\end{aligned}$$

Recalling $\tau = 1$ and letting $\rho \to 0$ gives the desired result. □

Proof of Theorem 2. The proof follows the same structure as the proof of Theorem 1. Therefore we provide a sketch and leave out the complete derivation. Let $0 < \rho < 1$ be arbitrary. Since F is a unitary matrix, for any $S \subseteq [n]$ and $v \in \mathbb{C}^n$, we have

$$\|v_S\|_2 \leq \frac{\rho}{\sqrt{k}}\|v_{\overline{S}}\|_1 + \|v_S\|_2 \leq \frac{\rho}{\sqrt{k}}\|v_{\overline{S}}\|_1 + \sqrt{k}\|v\|_\infty$$

The rest of the argument is the same as in the proof of Theorem 1. □

We first establish the restricted isometry property for certain structured matrices. First, we give some definitions.

Definition 8. *Let A be a matrix in $\mathbb{C}^{m \times N}$, let $M \subseteq \mathbb{C}^N$, and let $\delta \geq 0$. We say that A satisfies the M-restricted isometry property (or M-RIP) with constant δ if*

$$(1-\delta)\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1+\delta)\|x\|_2^2$$

Layer	Type	Properties
1	Convolution	32 channels, 3×3 Kernel, No padding
2	Convolution	64 channels, 3×3 Kernel, No padding, Dropout with $p = 0.5$
3	Max-pooling	2×2, Dropout with $p = 0.5$
4	Fully Connected	128 neurons, Dropout with $p = 0.5$
5	Fully Connected	10 neurons

TABLE IV
NETWORK ARCHITECTURE USED FOR MNIST AND FASHION-MNIST DATASETS IN SECTION IV-A AND SECTION IV-B. THE FIRST FOUR LAYERS USE RELU ACTIVATIONS WHILE THE LAST LAYER USES A SOFTMAX ACTIVATION.

for all $x \in M$.

Definition 9. *We define* M_k *to be the set of all* k*-sparse vectors in* $\mathbb{C}^N$ *and similarly define* $M_{k,t}$ *to be the set of* (k,t)*-sparse vectors in* $\mathbb{C}^{2n}$*. In other words,*

$$M_{k,t} = \{x = [x_1 \;\; x_2]^T \in \mathbb{C}^{2n} : \\ x_1 \in \mathbb{C}^n, x_2 \in \mathbb{C}^n, \|x_1\|_0 \leq k, \|x_2\|_0 \leq t\}.$$

We define $S_{k,t}$ *to be the following collection of subsets of* $\{1, \ldots, 2n\}$*:*

$$S_{k,t} = \{S_1 \cup S_2 : S_1 \subseteq \{1, \ldots, n\}, S_2 \subseteq \{n+1, \ldots, 2n\}, \\ card(S_1) \leq k, card(S_2) \leq t\}$$

Note that $S_{k,t}$ *is the collection of supports of vectors in* $M_{k,t}$*.*

Theorem 10. *Let* $A = [F \;\; I] \in \mathbb{C}^{n \times 2n}$*, where* $F \in \mathbb{C}^{n \times n}$ *is a unitary matrix with* $|F_{ij}|^2 \leq \frac{c}{n}$ *and* $I \in \mathbb{C}^{n \times n}$ *is the identity matrix. Then*

$$\left(1 - \sqrt{\frac{ckt}{n}}\right) \|x\|_2^2 \leq \|Ax\|_2^2 \leq \left(1 + \sqrt{\frac{ckt}{n}}\right) \|x\|_2^2 \quad (8)$$

for all $x \in M_{k,t}$*. In other words,* A *satisfies the* $M_{k,t}$*-RIP property with constant* $\sqrt{\frac{ckt}{n}}$*.*

Proof. In this proof, if B denotes an matrix in $\mathbb{C}^{n \times n}$, then $\lambda_1(B), \ldots, \lambda_n(B)$ denote the eigenvalues of B ordered so that $|\lambda_1(B)| \leq \cdots \leq |\lambda_n(B)|$. It suffices to fix an $S = S_1 \cup S_2 \in S_{k,t}$ and prove (8) for all non-zero $x \in \mathbb{C}^S$.

Since $A_S^* A_S$ is normal, there is an orthonormal basis of eigenvectors $u_1, \ldots, u_{k+t}$ for $A_S^* A_S$, where u_i corresponds to the eigenvalue $\lambda_i(A_S^* A_S)$. For any non-zero $x \in \mathbb{C}^S$, we have $x = \sum_{i=1}^{k+t} c_i u_i$ for some $c_i \in \mathbb{C}$, so

$$\frac{\|Ax\|_2^2}{\|x\|_2^2} = \frac{\langle A_S^* A_S x, x\rangle}{\langle x, x\rangle} = \frac{\sum_{i=1}^{k+t} \lambda_i(A_S^* A_S) c_i^2}{\sum_{i=1}^{k+t} c_i^2}. \quad (9)$$

Thus it will suffice to prove that $|\lambda_i(A_S^* A_S) - 1| \leq \sqrt{\frac{ckt}{n}}$ for all i. Moreover,

$$|\lambda_i(A_S^* A_S) - 1| = |\lambda_i(A_S^* A_S - I)| \\ = \sqrt{\lambda_i\left((A_S^* A_S - I)^*(A_S^* A_S - I)\right)} \quad (10)$$

where the last equality holds because $A_S^* A_S - I$ is normal. By combining (9) and (10), we see that (8) will hold upon showing that the eigenvalues of $(A_S^* A_S - I)^*(A_S^* A_S - I)$ are bounded by ckt/n.

So far we have not used the structure of A, but now we must. Observe that $(A_S^* A_S - I)^*(A_S^* A_S - I)$ is a block diagonal matrix with two diagonal blocks of the form $X^* X$ and $X X^*$. Therefore the three matrices $(A_S^* A_S - I)^*(A_S^* A_S - I)$, $X^* X$, and $X X^*$ have the same non-zero eigenvalues. Moreover, X is simply the matrix F_{S_1} with those rows not indexed by S_2 deleted. The hypotheses on F imply that the entries of $X^* X$ satisfy $|(X^* X)_{ij}| \leq \frac{ct}{n}$. So the Gershgorin disc theorem implies that each eigenvalue λ of $X^* X$ and (hence) of $(A_S^* A_S - I)^*(A_S^* A_S - I)$ satisfies $|\lambda| \leq \frac{ckt}{n}$.

□

Lemma 11. *Let* $A \in \mathbb{C}^{n \times 2n}$*, if* $\|Ax\|_2^2 \leq (1+\delta)\|x\|_2^2$ *for all* $x \in M_{k,t}$*, then,* $\|A_S^* A_S - I\|_{2 \to 2} \leq \delta$*, for any* $S \in S_{k,t}$*.*

Proof. Let $S \in S_{k,t}$ be given. Then for any $x \in \mathbb{C}^S$, we have

$$\|A_S x\|_2^2 - \|x\|_2^2 \leq \delta \|x\|_2^2$$

We can re-write this as : $\|A_S x\|_2^2 - \|x\|_2^2 = \langle A_S x, A_S x\rangle - \langle x, x\rangle = \langle (A_S^* A_S - I)x, x\rangle$. Noting that $A_S^* A_S - I$ is Hermitian, we have:

$$\|A_S^* A_S - I\|_{2 \to 2} = \max_{x \in \mathbb{C}^S \setminus \{0\}} \frac{\langle (A_S^* A_S - I)x, x\rangle}{\|x\|_2^2} \leq \delta$$

□

Proof of Theorem 3. We will derive (6) by showing that the matrix A satisfies all the hypotheses in Theorem 4.33 in [29] for every vector in $M_{k,t}$.

First note that by Theorem 10, A satisfies the $M_{k,t}$*-RIP* property with constant $\delta_{k,t} := \sqrt{\frac{ckt}{n}}$. Therefore, by Lemma 11, for any $S \in S_{k,t}$, we have $\|A_S^* A_S - I\|_{2 \to 2} \leq \delta_{k,t}$. Since $A_S^* A_S$ is a positive semi-definite matrix, it has only non-negative eigenvalues that lie in the range $[1 - \delta_{k,t}, 1 + \delta_{k,t}]$. Since $\delta_{k,t} < 1$ by assumption, $A_S^* A_S$ is injective. Thus, we can set: $h = A_S (A_S^* A_S)^{-1} \text{sgn}(x_S)$ and get:

$$\|h\|_2 = \|A_S (A_S^* A_S)^{-1} \text{sgn}(x_S)\|_2 \\ \leq \|A_S\|_{2 \to 2} \|(A_S^* A_S)^{-1}\|_{2 \to 2} \|\text{sgn}(x_S)\|_2 \leq \tau \sqrt{k+t}$$

where $\tau = \frac{\sqrt{1+\delta_{k,t}}}{1-\delta_{k,t}}$ and we have used the following facts: since $\|A_S^* A_S - I\|_{2 \to 2} \leq \delta_{k,t} < 1$, we get that $\|(A_S^* A_S)^{-1}\|_{2 \to 2} \leq \frac{1}{1-\delta_{k,t}}$ and that the largest singular value of A_S is less than $\sqrt{1+\delta_{k,t}}$. Now let $u = A^* h$, then $\|u_S - \text{sgn}(x_S)\|_2 = 0$. Now we need to bound the value $\|u_{\overline{S}}\|_\infty$. Denoting row j of $A_{\overline{S}}^* A_S$ by the vector v_j, we see that it has at most $\max\{k,t\}$ non-zero entries and that $|(v_j)_l|^2 \leq \frac{c}{n}$ for $l = 1, \ldots, (k+t)$. Therefore, for any element $(u_{\overline{S}})_j$, we have:

$$|(u_{\overline{S}})_j| = |\langle (A_S^* A_S)^{-1} \text{sgn}(x_S), (v_j)^*\rangle| \\ \leq \|(A_S^* A_S)^{-1}\|_{2 \to 2} \|\text{sgn}(x_S)\|_2 \|v_j\|_2 \\ \leq \frac{\sqrt{k+t}}{1-\delta_{k,t}} \sqrt{\frac{\max\{k,t\}c}{n}}$$

Defining $\beta := \sqrt{\frac{\max\{k,t\}c}{n}}$ and $\theta := \frac{\sqrt{k+t}}{1-\delta_{k,t}}\beta$, we get $\|u_{\overline{S}}\|_\infty \leq \theta < 1$ and also observe that $\max_{l \in \overline{S}} \|A_S^* a_l\|_2 \leq \beta$.

Therefore, all the hypotheses of Theorem 4.33 in [29] have been satisfied. Note that $y = F\hat{x} + e = A[\hat{x}_{h(k)} \;\; e]^T + F\hat{x}_{t(k)}$, Therefore, setting $x^{\#} = \text{BP}(y, A, \|\hat{x}_{t(k)}\|_2)$, we use the fact $\|F\hat{x}_{t(k)}\|_2 = \|\hat{x}_{t(k)}\|_2$ combined with the bound in Theorem 4.33 in [29] to get (6):

$$\|\hat{x}^{\#} - \hat{x}_{h(k)}\|_2 \leq \left(\frac{2\tau\sqrt{k+t}}{1-\theta} \left(1 + \frac{\beta}{1-\delta_{k,t}} \right) + 2\tau \right) \|\hat{x}_{t(k)}\|_2$$

where we write $x^{\#} = [\hat{x}^{\#}, e^{\#}]^T$ with $\hat{x}^{\#}, e^{\#} \in \mathbb{C}^n$. □

REFERENCES

[1] M. Bafna, J. Murtagh, and N. Vyas, "Thwarting adversarial examples: An *l*_0-robust sparse fourier transform," in *Advances in Neural Information Processing Systems*, pp. 10075–10085, 2018.

[2] E. J. Candes, J. K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, vol. 59, no. 8, pp. 1207–1223, 2006.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[4] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, pp. 3104–3112, 2014.

[5] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[6] C. Szegedy, W. Zaremba, and I. Sutskever, "Intriguing properties of neural networks," *International Conference on Learning Representations*, 2014.

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

[8] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*, pp. 372–387, IEEE, 2016.

[9] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 39–57, IEEE, 2017.

[10] M. Hein and M. Andriushchenko, "Formal guarantees on the robustness of a classifier against adversarial manipulation," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pp. 2263–2273, 2017.

[11] M. Cisse, P. Bojanowski, E. Grave, Y. Dauphin, and N. Usunier, "Parseval networks: Improving robustness to adversarial examples," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 854–863, JMLR. org, 2017.

[12] A. Sinha, H. Namkoong, and J. Duchi, "Certifying some distributional robustness with principled adversarial training," *arXiv preprint arXiv:1710.10571*, 2017.

[13] E. Wong and J. Z. Kolter, "Provable defenses against adversarial examples via the convex outer adversarial polytope," *arXiv preprint arXiv:1711.00851*, 2017.

[14] M. Lecuyer, V. Atlidakis, R. Geambasu, D. Hsu, and S. Jana, "Certified robustness to adversarial examples with differential privacy," *arXiv preprint arXiv:1802.03471*, 2018.

[15] C. Dwork, A. Roth, *et al.*, "The algorithmic foundations of differential privacy," *Foundations and Trends® in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.

[16] B. Li, C. Chen, W. Wang, and L. Carin, "Second-order adversarial attack and certifiable robustness," *arXiv preprint arXiv:1809.03113*, 2018.

[17] F. Tramèr, A. Kurakin, N. Papernot, D. Boneh, and P. McDaniel, "Ensemble Adversarial Training: Attacks and Defenses," *ArXiv e-prints*, May 2017.

[18] W. Xu, D. Evans, and Y. Qi, "Feature Squeezing Mitigates and Detects Carlini/Wagner Adversarial Examples," *ArXiv e-prints*, May 2017.

[19] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[20] P. Samangouei, M. Kabkab, and R. Chellappa, "Defense-gan: Protecting classifiers against adversarial attacks using generative models," *arXiv preprint arXiv:1805.06605*, 2018.

[21] A. Prakash, N. Moran, S. Garber, A. DiLillo, and J. Storer, "Deflecting adversarial attacks with pixel deflection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8571–8580, 2018.

[22] C. Xie, J. Wang, Z. Zhang, Z. Ren, and A. Yuille, "Mitigating adversarial effects through randomization," *arXiv preprint arXiv:1711.01991*, 2017.

[23] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on information theory*, vol. 52, no. 2, pp. 489–509, 2006.

[24] D. L. Donoho *et al.*, "Compressed sensing," *IEEE Transactions on information theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[25] E. Candes and T. Tao, "Decoding by linear programming," *arXiv preprint math/0502327*, 2005.

[26] T. Blumensath and M. E. Davies, "Iterative hard thresholding for compressed sensing," *Applied and computational harmonic analysis*, vol. 27, no. 3, pp. 265–274, 2009.

[27] R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hedge, "Model-based compressive sensing," *IEEE Transactions on Information Theory*, vol. 56, no. 4, pp. 1982–2001, 2010.

[28] E. Candes, T. Tao, *et al.*, "The dantzig selector: Statistical estimation when p is much larger than n," *The annals of Statistics*, vol. 35, no. 6, pp. 2313–2351, 2007.

[29] S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing*. 2017.

[30] D. L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via 1 minimization," *Proceedings of the National Academy of Sciences*, vol. 100, no. 5, pp. 2197–2202, 2003.

[31] D. L. Donoho and M. Elad, "On the stability of the basis pursuit in the presence of noise," *Signal Processing*, vol. 86, no. 3, pp. 511–532, 2006.

[32] D. L. Donoho and X. Huo, "Uncertainty principles and ideal atomic decomposition," *IEEE transactions on information theory*, vol. 47, no. 7, pp. 2845–2862, 2001.

[33] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pp. 506–519, 2017.

[34] A. Athalye, N. Carlini, and D. Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," *arXiv preprint arXiv:1802.00420*, 2018.

[35] A. Krizhevsky, "Learning multiple layers of features from tiny images," tech. rep., Citeseer, 2009.

[36] Y. LeCun, "The mnist database of handwritten digits," *http://yann. lecun. com/exdb/mnist/*.

[37] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," *arXiv preprint arXiv:1708.07747*, 2017.

[38] S. Diamond and S. Boyd, "CVXPY: A Python-embedded modeling language for convex optimization," *Journal of Machine Learning Research*, vol. 17, no. 83, pp. 1–5, 2016.

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[40] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: a simple and accurate method to fool deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2574–2582, 2016.

[41] A. Shafahi, M. Najibi, M. A. Ghiasi, Z. Xu, J. Dickerson, C. Studer, L. S. Davis, G. Taylor, and T. Goldstein, "Adversarial training for free!," in *Advances in Neural Information Processing Systems*, pp. 3353–3364, 2019.

[42] E. Wong, L. Rice, and J. Z. Kolter, "Fast is better than free: Revisiting adversarial training," *arXiv preprint arXiv:2001.03994*, 2020.

[43] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," *arXiv preprint arXiv:1607.02533*, 2016.