# Making medical image reconstruction adversarially robust

Adva Wolf
Mentor: Yair Carmon

**Abstract**

We adapt a provable defense technique named "randomized smoothing" from classification to regression. We apply it to the problem of reconstructing MRI knee images from subsampled data: we attack the MRI-VN model [7] and successfully defend it using randomized smoothing. Our code is available in [16], and heavily based on [14, 13]. The code uses a variant of Tensorflow which is available in [15].

## 1 Introduction

Deep learning offers new approaches to image reconstruction, an inverse problem in which we are trying to estimate an object from a finite number of linear measurements. Stable and accurate solutions to this problems are crucial for MRI and CT scans, as well as other applications in the life sciences. Handcrafted signal processing methods to reconstruct tomographic images require many measurements to provide an accurate reconstruction. Current deep learning techniques which try to solve the same problem have the goal of achieving the same accuracy, but with a much smaller number of samples. Reaching this goal could potentially make MRI and CT scans shorter, which is very desirable.

However, many of the deep learning methods that have been developed so far for image reconstruction are not robust: a small adversarially-selected change in the input can create a large change in the output. For example, a small perturbation can create the appearance of non-existing tumors, or make the whole output image uniformly distorted [1]. This phenomenon also occurs in other deep learning tasks, and the study of attacks and defenses has developed to the field of adversarial machine learning.

In this project, we develop a defense technique for medical image reconstruction. Our starting point is "randomized smoothing" [3, 5, 6] a technique for making classification models provably robust. On the theoretical side, we extend the analysis of [2] from classification to regression. On the practical side, we improve upon the MRI-VN model [7]. The MRI-VN model and the improved variants we built take as input complex-valued undersampled Fourier coefficients together with additional sampling parameters and use supervised training to learn a deep neural network to reconstruct MRI images of knees. In addition, we implement a variant of the attack described in [1] by using projected gradient descent (PGD) in order to find strong adversarial attacks for our improved MRI-VN models and test our defense.

## 2 Related work

**Randomized smoothing.** Randomized smoothing was first introduced in [3] for improving and analyzing convergence rates of algorithms for non smooth convex optimization problems. [6, 5, 2] use similar technique as a defense against adversarial attacks. These works emphasize the importance of certified robustness: a defense that comes with theoretical, provable guarantees, in contrast to heuristic defense techniques which are vulnerable to new and improved attacks. As far as we know, randomized smoothing was used as a defense only in classification problems, and in this work we apply it to a regression problem.

**CS-MRI and variational networks.** Deep learning techniques for compressed sensing magnetic resonance imaging (CS-MRI) gained a lot of interest recently and may significantly change the field. Recent works vary in terms of the network's architecture, its flexibility, the size of the training dataset and more [4, 8]. In this work, we improve upon the MRI-VN model [7] which predicts MRI images of knees from subsampled MRI data and uses a variational network. Its design is inspired by the "field of expert" model [12], which generalized total variation regularization [9] and allows it to performs well even though its training set is relatively small.

Figure 1: From left to right, we see the full k-space of a single coil, the sensitivity map of a single coil and the ground truth, all cropped. Here we take the absolute value of the complex matrices element-wise and scale them between 0 and 1.

**Adversarial attacks on CS-MRI.** In [1] the authors find adversarial attacks on the CS-MRI models mentioned above, and in particular on the MRI-VN model. While they use Lagrange multipliers in the optimization problem of finding the adversarial examples, we use projected gradient ascent to get better attacks.

# 3 The Dataset

The dataset we use is available online [13] and was obtained by the authors of [7], for a type of parallel MR imaging protocol for knees named *Coronal Spin Density weighted with Fat Suppression*. In this mode, the MRI machine uses 15 *receiver coils*, each producing a Fourier transform of a part of the MRI image that we would like to reconstruct. Since each coil is sensitive to MR signal arising from a different region in the knee, each coil comes with position-dependent sensitivity map. The connection between the output $k_i$ of coil $i$ and the true image $y$ is given by $k_i = \mathcal{F}(S_i \circ y) + \text{noise}$, when $k_i$ represents the *full k-space* of coil $i$ (using the medical imaging community jargon). Here $S_i$ is the sensitivity map, $\circ$ is element-wise multiplication and $\mathcal{F}$ is the Fourier transform. From the fully sampled k-spaces and the sensitivity maps, the authors of [7] reconstructed the ground truth reference image by $y = \sum_{i=1}^{15} \mathcal{F}^{-1}(k_i) \circ S_i$. Now, we can create subsampled k-space by multiplying the fully sampled $k_i$ 's by an appropriate mask. We will denote by $k_{sub}$ and $S$ the subsampled k-space and the sensitivity maps of the 15 coils, and by $y$ the ground truth. Following the work done in [1], we use a constant mask that gives a 15% subsampling rate.

The training data consists of 200 triples $(k_{sub}^{(i)}, S^{(i)}, y^{(i)})$ obtained from 10 patients (the MRI scan of each patient produces around 35 *slices*, when each slice represents an image of a different slice of the knee. For each patient we take the central 20 slices). $k_{sub}^{(i)}$ and $S^{(i)}$ are complex matrices of size $(15, 640, 368)$ and $y^{(i)}$ is a complex matrix of size $(640, 368)$. No data augmentation was used. The validation set consists of 10 triples $(k_{sub}^{(i)}, S^{(i)}, y^{(i)})$ obtained from 10 new patients, when we took a single (central) slice from each patient's MRI scan. An example from the data set is given in Fig 1.

# 4 Methods

**Randomized smoothing for regression problems** Given a bounded base map $f$ (representing a neural network) , we build a "smoothed" map $g$ which is provably robust: a small perturbation in the input will result in a proportional perturbation of the output (in worst case). Here we give the full details on how to define $g$ and what is the theoretical guarantee, based on [3]

**Theorem 1.** *Let $f : \mathbb{R}^d \to \mathbb{R}^m$ be any bounded function. Let $\epsilon \sim \mathcal{N}\left(0, \sigma^2 I_{d\times d}\right)$. We define $g : \mathbb{R}^d \to \mathbb{R}^m$ as:*

$$g(x) = \mathbb{E}\left(f(x+\epsilon)\right).$$

*Then, $g$ is an $\frac{M}{\sqrt{2\pi}\sigma}$-Lipschitz map, where $M = (\max ||f||_2 + \min ||f||_2)$. In particular, for any $x, \delta \in \mathbb{R}^k$:*

$$||g(x) - g(x+\delta)||_2 \leq M\, erf\left(\frac{||\delta||_2}{2\sqrt{2}\sigma}\right) \leq \frac{M}{\sqrt{2\pi}\sigma}||\delta||_2$$

*where erf is the Gauss error function and the above bound is tight.*

The proof for this theorem is given in the appendix. We see that the Lipschitz constant $\frac{M}{\sqrt{2\pi}\sigma}$ will get smaller as the variance of $\epsilon$ grows. However, as it grows our function $g$ will be less likely to predict the same values as $f$. This reflects a trade-off between the accuracy of $g$ and its robustness.

**The MRI-VN model:** We give a short overview of the architecture of the MRI-VN model. The full description can be found in [7]. For an input of the form $(k_{sub}, S)$ the model first reconstructs the zero-filled image:

$$u^0 = \sum_{i=1}^{15} \mathcal{F}^{-1}(k_{sub,i} \circ M_{\text{mask}}) \circ S_i = A_S^*(k_{sub})$$

where $M_{\text{mask}}$ is the mask we used to subsample the full k-space. We denote by $A_S = (A_S^*)^*$ the adjoint operator to $A_S^*$. Then, there are 10 hidden layers of the form:

$$u^{t+1} = u^t - \left(K^t\right)^T \Psi^t \left(K^t u^t\right) + \lambda^t A_S^* \left(A_S u^t - k_{sub}\right), \quad 0 \le t < 10$$

when $\lambda^t$ is a scalar, $K^t$ is a convolutional operators with weights $\mathbf{k}^t$ and $\Psi^t$ is non-linear activation functions (a weighted combination of RBFs with weights $w_{ij}^t$). $\lambda^t, K^t, \Psi^t$ are different for each layer. The parameters of the model are $\theta^0, \ldots, \theta^{10}$ for $\theta^t = \left\{w_{ij}^t, \mathbf{k}^t, \lambda^t\right\}$, and during training the model finds the optimal parameters by minimizing:

$$\mathcal{L}(\theta) = \frac{1}{2N} \sum_{s=1}^{N} \left\| |u_s^T(\theta)|_\epsilon - |y_s|_\epsilon \right\|_2^2, \text{ where } |x|_\epsilon = \sqrt{x_{\text{re}}^2 + x_{\text{im}}^2 + \epsilon}.$$

To do that, the authors of [7] use a variant of projected mini-batch gradient descent named Inertial Proximal Alternating Linearized Minimization (IPALM, [10]) which controls the weights norm. At prediction, the output of the model is $\left|u^{10}\right|$.

**Robustness in practice: defending the MRI-VN model** We apply randomized smoothing similarly to the algorithm in the classification problem case [2, 5, 6]. We follow the next steps to obtain the new robust network $f_{rs}$:

a. **Noisy Training:** train the MRI-VN network by augmenting the data with noise, $\{(k_{sub}^{(i)} + \epsilon^{(i)}, S^{(i)}, y^{(i)}\}$ with i.i.d Gaussians. $\epsilon^{(i)} \sim \mathcal{N}\left(0, \sigma_{train}^2 I_{d\times d}\right)$

b. **Smoothed Prediction:** The prediction of $f_{rs}$ is:

$$f_{rs}(k_{sub}) = \frac{1}{K} \sum_{k=1}^{K} f_{\theta_{noise}}(k_{sub} + \eta^{(k)})$$

with the i.i.d Gaussians $\eta^{(k)} \sim \mathcal{N}\left(0, \sigma_{pred}^2 I_{d\times d}\right)$.

In practice, we use $\sigma_{train} = \sigma_{pred} = \sigma$ since the authors of [2] indicated that using $\sigma_{train} \neq \sigma_{pred}$ was ineffective, and choose the largest possible $K$ giving our memory limitations. For high enough $K$, step (b) gives a good approximation of the expectation $\mathbb{E}\left(f_{\theta_{noise}}(k_{sub} + \eta)\right)$. We note that the theorem above guarantees that $f_{rs}$ will be Lipschitz (for high enough $K$) even by skipping step (a) and using $f$ instead of $f_{\theta_{noise}}$. However, step (a) is necessary for keeping the accuracy of $f_{rs}$ reasonable.

**Testing the defense: the adversarial attack** Let $f$ be either the MRI-VN model or its smoothed version. We attack $f$ by using the framework of the authors of [1] and maximizing

$$E(r) = \frac{1}{2}\|f(k_{sub} + A_S r) - f(k_{sub})\|_2^2$$

using projected gradient ascent with momentum, with the constraint $\|r\|_2 < R$. Originally the authors of [1] used Lagrange Multipliers with hyper-parameter $\lambda$ instead of PGD, but by using PGD we get better attacks.

# 5 Results and Discussion

When training the MRI-VN and its smoothed versions all the network's parameters where set as in [1]. We test the accuracy and the robustness of the models on the validation set. We use the Root Mean Squared

| | No attack | | $\|r\| = 1.5$ | | $\|r\| = 2.5$ | | $\|r\| = 3.5$ | |
|---|---|---|---|---|---|---|---|---|
| Model | RMSE | SSIM | RMSE | SSIM | RMSE | SSIM | RMSE | SSIM |
| Original | **14.76** | **0.906** | 22.70 | 0.827 | 28.63 | 0.765 | 33.44 | 0.718 |
| $\sigma = 0.01$, $K = 9$ | 14.84 | 0.902 | 17.58 | 0.887 | 20.50 | 0.872 | 24.15 | 0.855 |
| $\sigma = 0.02$, $K = 9$ | 15.13 | 0.898 | **17.02** | **0.889** | 19.25 | **0.878** | 22.30 | 0.856 |
| $\sigma = 0.05$, $K = 9$ | 16.46 | 0.875 | 17.21 | 0.874 | **18.18** | 0.872 | 19.79 | **0.869** |
| $\sigma = 0.1$, $K = 9$ | 17.79 | 0.851 | 18.02 | 0.851 | 18.52 | 0.850 | **19.15** | 0.848 |

Table 1: The performance of the models on the validation set and the adversarial examples, for various perturbation sizes.

Error (RSME) and the Structural Similarity (SSIM) [11] metrics to evaluate it quantitatively. These metrics are given by:

$$\mathrm{RMSE}(f, \{k_{sub}^{(i)}, S^{(i)}, y^{(i)}\}_{i=1}^{N}) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(f(k_{sub}^{(i)}) - y^{(i)})^2} \text{ and}$$

$$\mathrm{SSIM}(f, \{k_{sub}^{(i)}, S^{(i)}, y^{(i)}\}_{i=1}^{N}) = \frac{1}{N}\sum_{i=1}^{N}\mathrm{ssim}(f(k_{sub}^{(i)}), y^{(i)}) \quad \mathrm{ssim}(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}.$$

We look for adversarial examples using the code of [1], while changing the algorithm as we described in section 4. This change produces better attacks. We tuned the PGD attack and set the learning rate to 0.5, the momentum to 0.9 and the maximal number of iteration to 60. Using the PGD attack, we create adversarial examples for a range of perturbation sizes for each example in the validation set. We test our defense while setting $K = 9$, the maximal $K$ we could use due to memory limitations and compute the metrics mentioned above, listing the results in Table 1.
As predicted, without any attack the accuracy decreases with $\sigma$, but as the effect of the perturbation grows the robustness grows with $\sigma$. In order to understand better the robustness of our model for different perturbations, in Figure 2 we plot the perturbation effect and the prediction error as a function of the size of the perturbation. Also here we see that increased $\sigma$ provides a more robust model. We also see that even for the smallest value of $\sigma = 0.01$ the defense is working. The right figure shows a trade-off between accuracy and robustness. For $\sigma = 0.1$ we get the worst accuracy on small values of $|r|$.

We also provide in Fig. 3 qualitative results, in which we see our attacks and defense on images of two different patients. We see that in all models the adversarial example (each chosen specifically for each model) is not visible to the human eye. We also see clearly that the prediction of the original model on the adversarial example is heavily distorted. Even for small $\sigma = 0.01$ which provides an accuracy similar to the original model, the distortion is reduced. The defense is working at its best for $\sigma = 0.1$, but we can see that the predictions are blurry and the accuracy decreases.

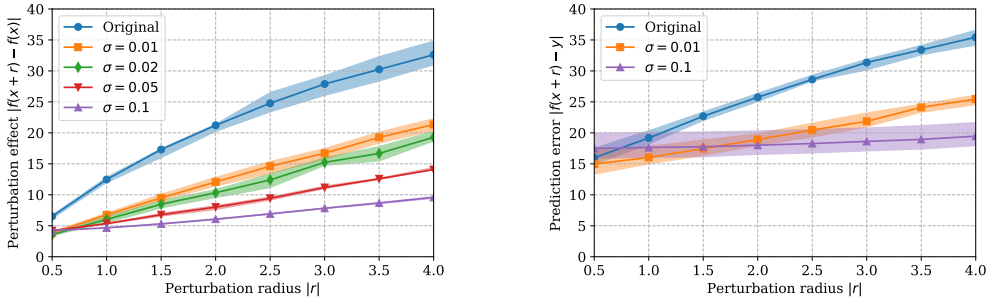


Figure 2: The bold lines represent the mean among the 10 adversarial examples for a single perturbation size, and the boundary of the shaded region is the 25th and 75th percentile. On left, we see how $|f(x+r) - f(x)|$ behaves as $|r|$ grows, and on the right we see prediction error $|f(x+r) - y|$ as a function of $|r|$.

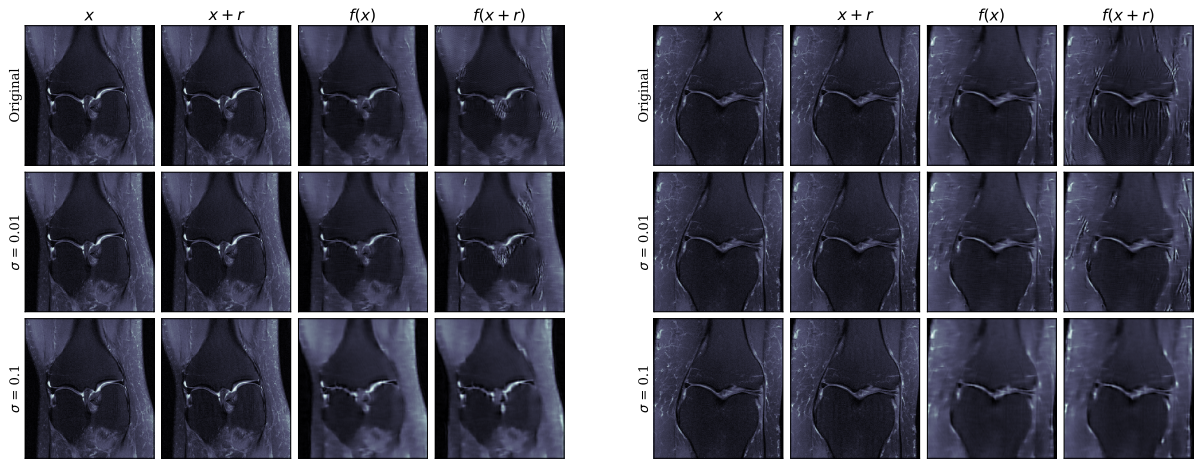


Figure 3: Reconstruct, attack and defense results for two patients. For each patient, from left to right: ground truth, ground truth with small perturbation (the adversarial example), prediction without the perturbation, prediction on the adversarial example. The rows represent different models.

# 6 Conclusion And Future Work

We conclude that randomized smoothing is effective even for small values of $\sigma$ and $K$. We note that for the values of $\sigma$ we used, the theoretical guarantee gave a large Lipschitz constant that was irrelevant to the scale of the problem. Still, we defend successfully the MRI-VN model and by varying $\sigma$, we saw a trade-off between the accuracy and the robustness, both quantitatively and qualitatively.

In future work, we would like to test randomized smoothing on MRI-reconstruction systems with different architecture and datasets. We would like to test the effect of the hyper-parameter $K$. Mathematically, we are interested in generalizing randomized smoothing to other norms (e.g. $\ell_1$, $\ell_\infty$) using other distibutions than the normal distribution.

# 7 Appendix

Here we provide the proof of the main theorem, using methods from [3].

*Proof:* We have that

$$\begin{aligned}||g(x) - g(x+\delta)||_2 &= || \int_{\mathbb{R}^d} f(w) \left[\mu(w-x) - \mu(w-x-\delta)\right] dw||_2 \\ &= || \int_{D^+} f(w) \left[\mu(w-x) - \mu(w-x-\delta)\right] dw - \int_{D^-} f(w) \left[\mu(w-x-\delta) - \mu(w-x)\right] dw||_2\end{aligned}$$

when $D^+ = \{w : \mu(w-x) > \mu(w-x-\delta)\} = \{w : ||w-x||_2 < ||w-x-\delta||_2\}$ and
$D^- = \{w : \mu(w-x) < \mu(w-x-\delta)\} = \{w : ||w-x||_2 > ||w-x-\delta||_2\}$. We notice that

$$\int_{D^+} \left[\mu(w-x) - \mu(w-x-\delta)\right] dw = \int_{D^-} \left[\mu(w-x-\delta) - \mu(w-x)\right] dw$$

Then:

$$||g(x) - g(x+\delta)||_2 \leq (\max ||f||_2 + \min ||f||_2) \int_{D^+} \left[\mu(w-x) - \mu(w-x-\delta)\right] dw$$

using Jensen's and Holder's inequalities. Now using exactly the same computation as in [3] we get:

$$I = \int_{D^+} \left[\mu(w-x) - \mu(w-x-\delta)\right] dw = \text{erf}\left(\frac{||\delta||_2}{2\sqrt{2}\sigma}\right) \leq \frac{1}{\sqrt{2\pi}\sigma} ||\delta||_2$$

We can see that the bound is tight for the step function $f(x) = 1\{x > 0\}$. At $x = 0$, we have $|g(\frac{\delta}{2}) - g(-\frac{\delta}{2})| = \text{erf}\left(\frac{||\delta||_2}{2\sqrt{2}\sigma}\right)$ and similar example works for higher dimensions.

## Acknowledgments

I am grateful for the guidance and assistance of Yair Carmon, who suggested the idea for this project and assisted me with debugging TensorFlow. I am very grateful for John Duchi, who allowed me to use his GPUs. I am also grateful for the authors of [1, 7] who made their code and dataset available online [14, 13], on which I heavily based my code.

## References

[1] Vegard Antun, Francesco Renna, Clarice Poon, Ben Adcock and Anders C. Hansen. On instabilities of deep learning in image reconstruction - Does AI come at a cost? arXiv preprint arXiv:1902.05300, 2019.

[2] Jeremy M Cohen, Elan Rosenfeld, and J Zico Kolter. Certified adversarial robustness via randomized smoothing. arXiv preprint arXiv:1902.02918, 2019.

[3] Duchi, J. C., Bartlett, P. L., and Wainwright, M. J. Randomized smoothing for stochastic optimization. SIAM Journal on Optimization, 2012.

[4] B. Zhu, J. Z. Liu, S. F. Cauley, B. R. Rosen and M. S. Rosen, 'Image reconstruction by domain-transform manifold learning', Nature, vol. 555, no. 7697, p. 487, Mar. 2018.

[5] Lecuyer, M., Atlidakis, V., Geambasu, R., Hsu, D., and Jana, S. Certified robustness to adversarial examples with differential privacy. In IEEE Symposium on Security and Privacy (SP), 2019.

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

[7] K. Hammernik, T. Klatzer, E. Kobler, M. P. Recht, D. K. Sodickson, T. Pock and F. Knoll, Learning a variational network for reconstruction of accelerated MRI data, Magnetic resonance in medicine, vol. 79, no. 6, pp. 3055-3071, 2018.

[8] G. Yang, S. Yu, H. Dong, G. Slabaugh, P. L. Dragotti, X. Ye, F. Liu, S. Arridge, J. Keegan, Y. Guo et al., DAGAN: Deep de-aliasing generative adversarial networks for fast compressed sensing MRI reconstruction, IEEE Transactions on Medical Imaging, 2017

[9] Rudin LI, Osher S, Fatemi E. Nonlinear total variation based noiseremoval algorithms. Physica D 1992;60:259-268.

[10] Pock T, Sabach S. Inertial proximal alternating linearized minimization (iPALM) for nonconvex and nonsmooth problems. SIAM J Imaging Sci 2016;9:1756-1787.

[11] Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 2004;13:600?612

[12] Roth S, Black MJ. Fields of experts. Int J Comput Vis 2009;82:205-229.

[13] github.com/VLOGroup/mri-variationalnetwork

[14] github.com/vegarant/Invfool

[15] github.com/VLOGroup/tensorflow-icg

[16] drive.google.com/drive/folders/1lg60Dyx4ftIM5NSGJBFz6046sKvopi0g?usp=sharing