ICML | 2020

Improving Robustness of Deep-Learning-Based Image Reconstruction

Ankit Raj[1], *Yoram Bresler*[1], *Bo Li*[2]

[1] Department of ECE, [2] Department of CS

University of Illinois at Urbana-Champaign

Acknowledgment - This work was supported in part by US Army MURI Award W911NF-15-1-0479

Overview

- Deep-learning-based inverse problem solvers recently proven to be sensitive to perturbations.
- Instability stems from the combined system (deep network + underlying inverse problem).

Contributions:

- Proposed a min-max formulation to build a ***robust*** model.
- Introduced an ***auxiliary network*** to generate adversarial examples for which the image recon network tries to minimize the recon loss.
- Significant improvement of robustness using the proposed approach over other methods for deep networks.
- Theoretically analyzed a simple linear network - found that min-max formulation results in singular-value filter regularized solution mitigating the effect of adversarial examples due to ill-conditioning.

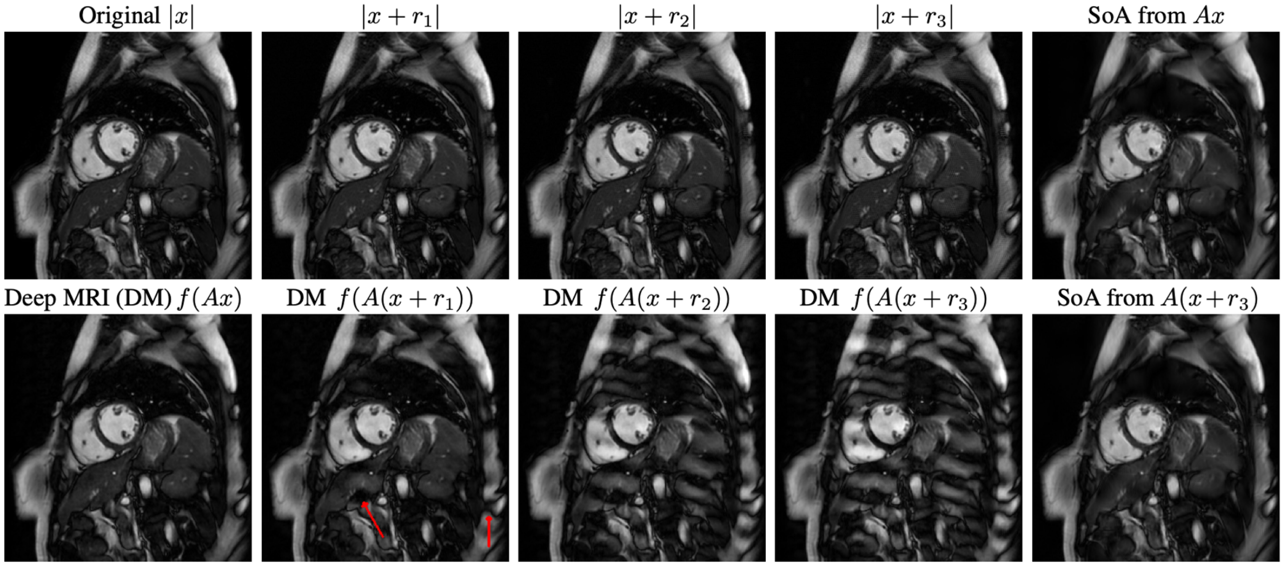
Attacks on DL-based Inverse problems solvers [1]

- Recent work shows deep learning typically yields unstable methods for image reconstruction.
- Evaluated 3 different types of instabilities:
  - **Tiny perturbation in the image domain results in severe artifacts.**
  - Small **structural** change which is not recovered.
  - Increasing number of samples does not improve recovery.

[1] Antun et al. On instabilities of deep learning in Image Recon and the potential costs of AI, PNAS ‘20

Instabilities to perturbation in *Image-Domain* [1]

$$y = Ax, \quad A \in \mathbb{R}^{m \times n} \qquad y' = A(x + r)$$



Attack is obtained by solving: $\max_r ||f(y + Ar) - x||^2 - \frac{\lambda}{2}||r||^2$

Modeling perturbations in *x or y-domain*?

Our argument - study of perturbation in x-domain is *sub-optimal* for inverse problems.

REASON - 1

- Perturbation in ***x*** may not be able to model all possible perturbations in ***y.***
- δ - perturbation in ***x*** leads to $A\delta$ perturbation in ***y.***
- ➡ Constrains the perturbation to be in Range(***A***).
- ➡ Not possible to model all possible perturbations when ***A*** does not have full-row rank.

Reason-2: Effect of Ill-Conditioning

$$A = \begin{bmatrix} 1 & 0 \\ 0 & a \end{bmatrix} \text{ and } f = \begin{bmatrix} 1 & 0 \\ 0 & 1/a \end{bmatrix} \qquad |a| \ll 1 \qquad \delta = \begin{bmatrix} 0 \\ \epsilon \end{bmatrix}$$

Perturbation in ***x***: $||f(A(x+\delta)) - x||_2 = ||\delta||_2 = \epsilon$

Perturbation in ***y***: $||f(Ax+\delta) - x||_2 = ||f\delta||_2 = \frac{\epsilon}{a}$

➡ For ill-conditioned measurement operator, an ideal inverse can be highly vulnerable to even a small perturbation in the ***measurement-space***, which is totally missed in the ***x-space*** formulation.

Reason-3: Measurement Operator Perturbations

- Suppose there is mismatch between ***A*** *used in training****,*** and the ***A*** actually generating the measurements.
- Let actual $A' = A + \tilde{A}$ ⇨ perturbation $\tilde{A}x$ in y-space.
- Typically $\tilde{A}x \notin Range(A)$, which the ***x-space*** formulation can't model.

Adversarial Training Framework for IR

$$\min_{\theta} \mathbb{E}_x \max_{\delta:\|\delta\|_2 \leq \epsilon} \|f(Ax;\theta) - x\|^2 + \lambda \|f(Ax+\delta;\theta) - x\|^2$$

- Ideal framework for adversarial training.
- Very expensive during training.
- Finding perturbation specific to each training sample.

↓ A sub-optimal approximation

$$\min_{\theta} \max_{\delta:\|\delta\|_2 \leq \epsilon} \mathbb{E}_x \|f(Ax;\theta) - x\|_2^2 + \lambda \|f(Ax+\delta;\theta) - x\|_2^2$$

- Tractable training.
- Finding perturbation common to many training samples.
- Not the ideal scheme. Why?

Desiderata for Adversarial Training

- Perturbation specific to the sample.
- Reasonably feasible to train in adversarial way.

$$\delta = \arg \max_{\delta:||\delta||_2 \leq \epsilon} ||f(y+\delta;\theta) - x||_2^2$$

Idea: model this perturbation using a deep network $G(y;\phi)$

Advantages:

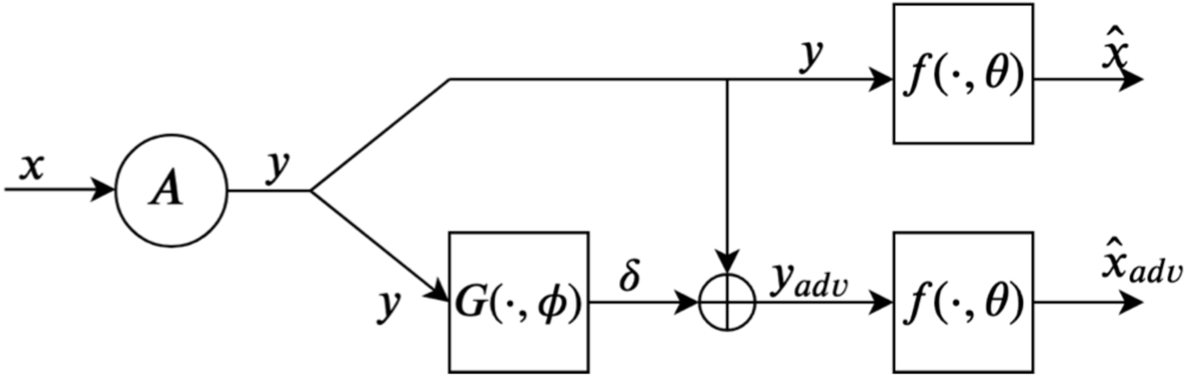
- This approach eliminates the need to solve the inner-max using hand-crafted method.
- Since ***G(.)*** is parameterized, and takes ***y*** as input, a well-trained ***G*** results in optimal perturbation, given ***y***.

Modified Objective

$$\min_{\theta} \max_{\phi : \|G(\cdot,\phi)\|_2 \leq \epsilon} \mathbb{E}_x \|f(Ax;\theta) - x\|^2 + \lambda \|f(Ax + G(Ax;\phi);\theta) - x\|^2$$

$$\min_{\theta} \max_{\phi} \mathbb{E}_x \underbrace{\|f(Ax;\theta) - x\|^2}_{\text{True Recon. term}} + \lambda_1 \underbrace{\|f(Ax + G(Ax;\phi);\theta) - x\|^2}_{\text{Adversarial term}} + \lambda_2 \underbrace{\max\{0, \|G(Ax;\phi)\|_2^2 - \epsilon^2\}}_{\text{Bounded perturbation term}}$$

Training Schematic



Robustness Metric

$$\Delta_{\max}(x_0, \epsilon) = \max_{\|\delta\|_2 \leq \epsilon} \|f(Ax_0 + \delta) - x_0\|^2$$

- Determines the reconstruction error due to the worst-case additive perturbation over the ϵ--ball around the measurement.
- Solved empirically using Projected Gradient Ascent.

$$\hat{\rho}(\epsilon) = \frac{1}{N} \sum_{i=1}^{N} \Delta_{\max}(x_i, \epsilon)$$

Smaller value implies more robust network

Experiments - Comparison Benchmarks

End-to-end Training (No Regularization):

$$\min_{\theta} \mathbb{E}_x \| f(Ax;\theta) - x\|^2$$

L2-norm Regularization ("weight decay"):

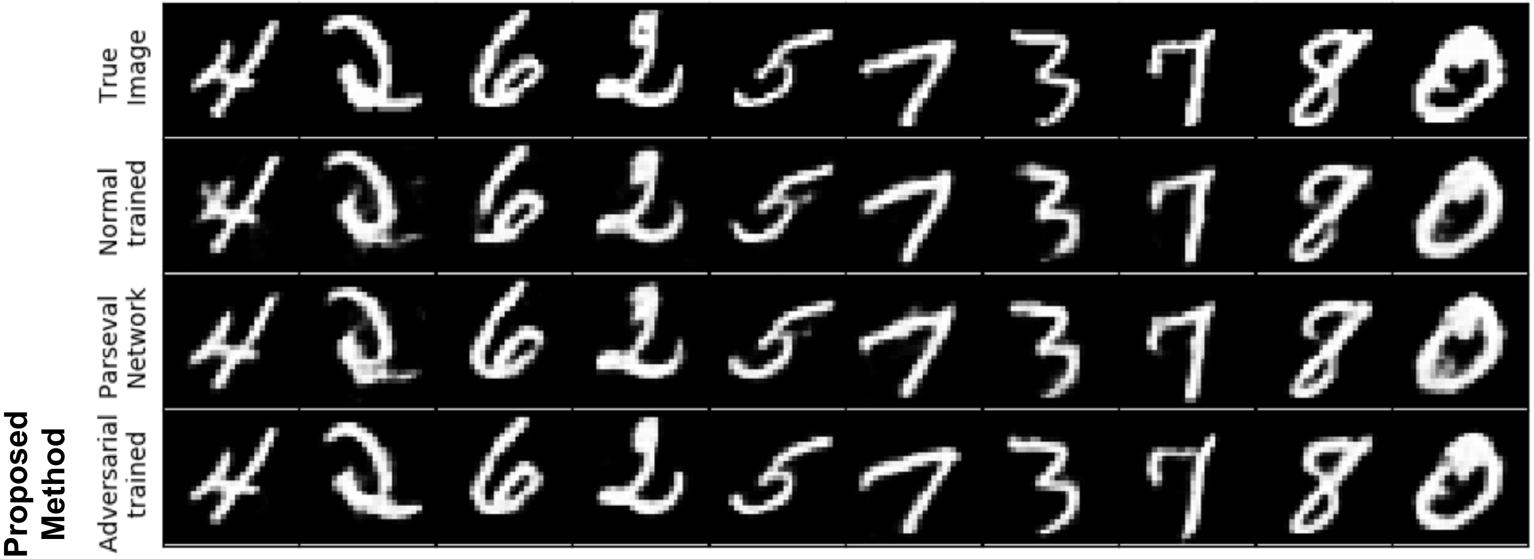
$$\min_{\theta} \mathbb{E}_x \| f(Ax;\theta) - x\|^2 + \mu\|\theta\|^2$$

Parseval Networks:

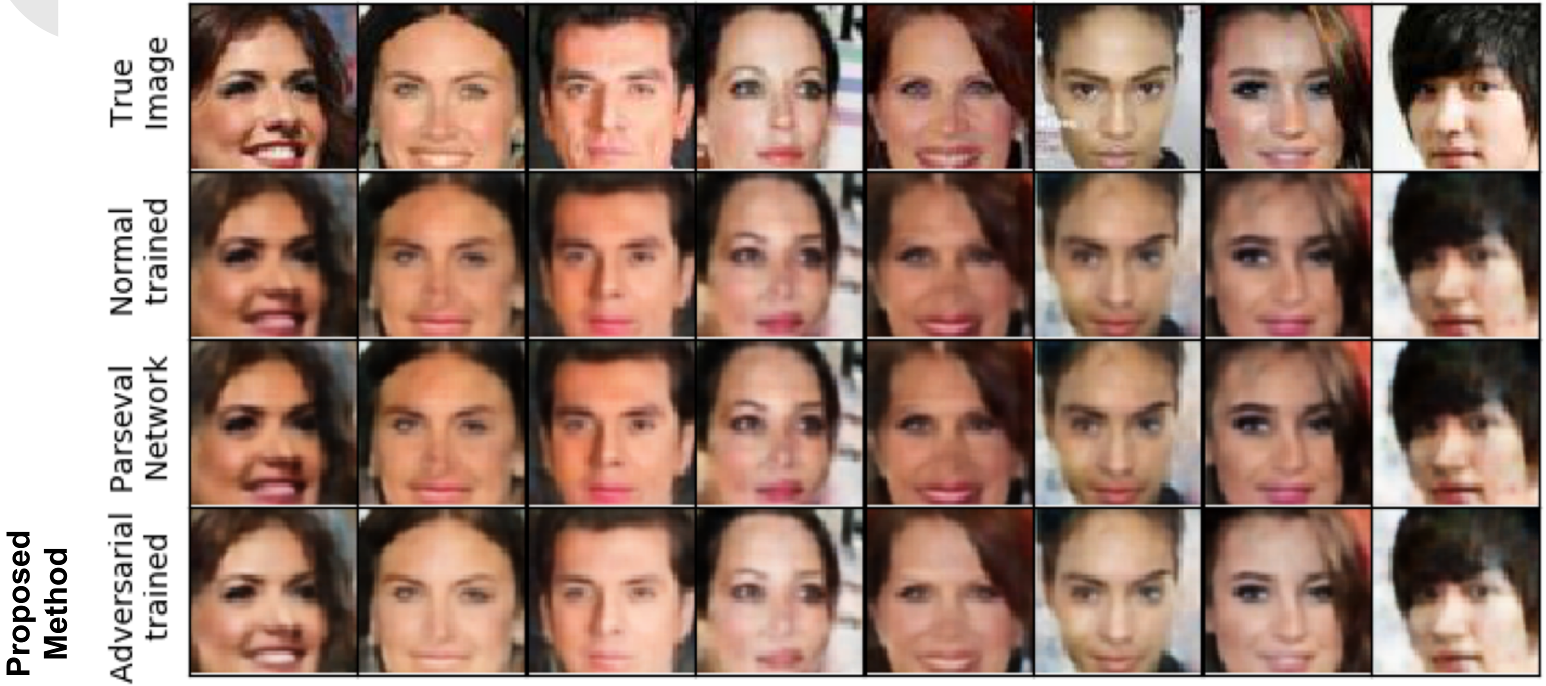
$$\min_{\theta} \mathbb{E}_x \| f(Ax;\theta) - x\|^2 + \frac{\beta}{2}(\sum_{i \in S_{fc}} \|W_i^T W_i - I_i\|_2^2 + \sum_{j \in S_c} \|\mathbf{W_j}^T \mathbf{W_j} - \frac{I_j}{k_j}\|_2^2)$$

Qualitative Results: MNIST

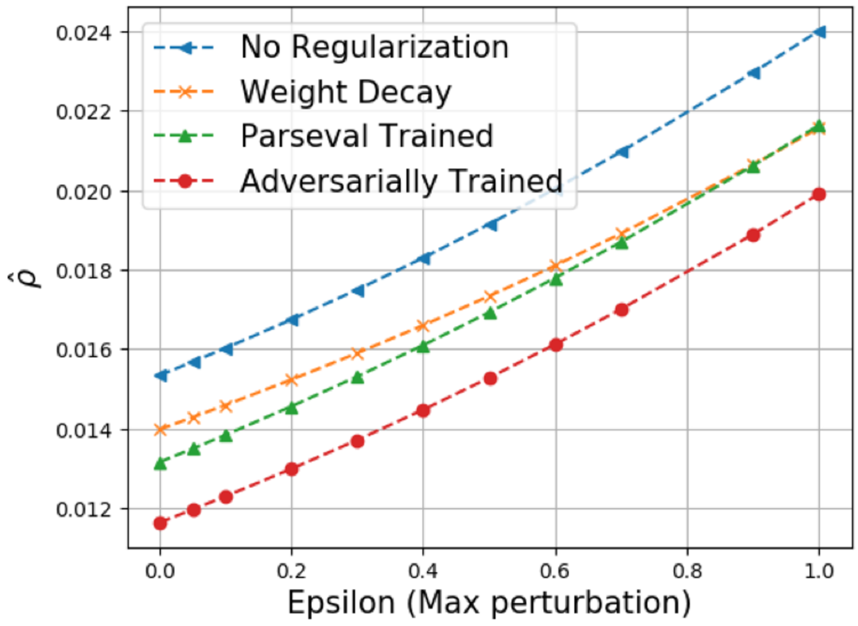
Compressed Sensing (with Gaussian Measurement Matrix): Recon using deep CNN



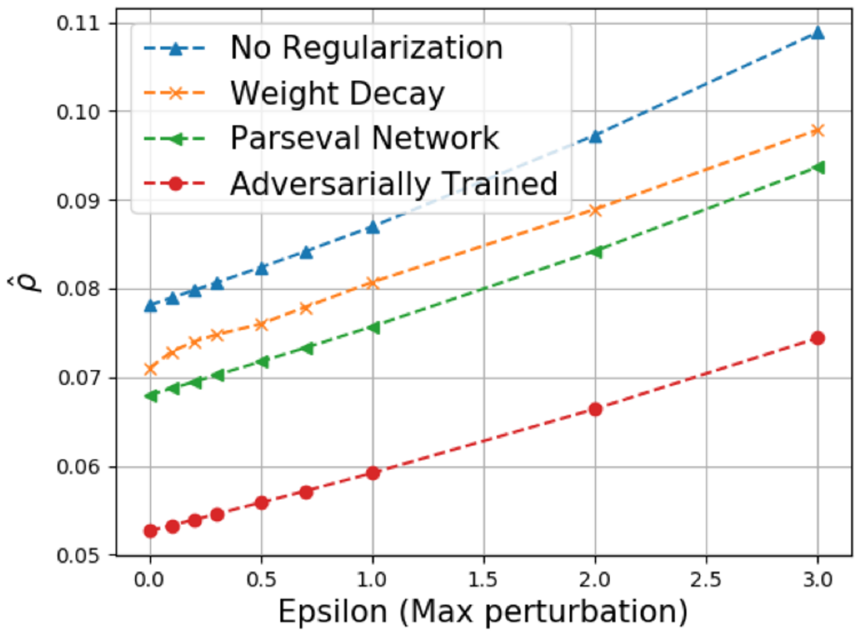
Qualitative Results: CelebA



Quantitative Results



MNIST



CelebA

Experiment on Real X-ray Images

- Implemented the proposed adversarial training algorithm on FBPConvNet [2] for low-dose CT reconstruction.
- For fast computation of forward projection (Radon transform) and filtered backprojection (FBP - numerical inverse Radon transform) on GPUs, we used the Astra toolbox [3].
- Dataset: Anonymized clinical CT images [4]: 884 slices for training, and 221 slices for evaluation.
- Measurements obtained by computing parallel-beam projections of the CT images at 143 view angles uniformly spaced on [0, 180].

[2] Jin et al. Deep CNN for Inverse Problems in Imaging, IEEE Trans. On Image Proc., 2017
[3] Van Aarle, W., et al. "Fast and flexible X-ray tomography using the ASTRA toolbox." Optics Express 2016
[4] Prof. Michael Vannier, Dept. Radiology, Univ. of Chicago, personal communication.

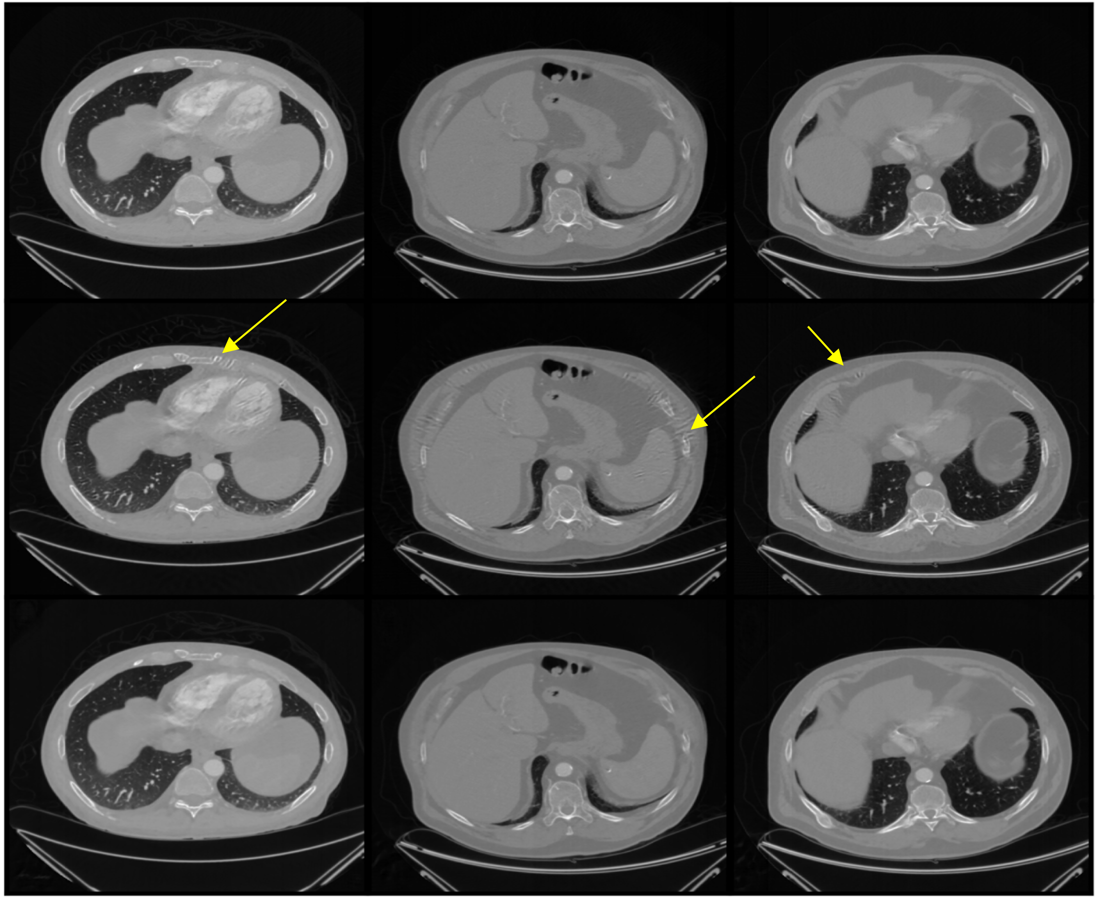
Qualitative Results for CT Recon

True Image

Normal trained

Proposed Method

Adversarial trained



Theoretical Analysis

$$\min_{\theta} \max_{\delta:\|\delta\|_2 \leq \epsilon} \mathbb{E}_x \|f(Ax;\theta) - x\|_2^2 + \lambda \|f(Ax+\delta;\theta) - x\|_2^2 \qquad (6)$$

Assumptions+Notation:

- f is a one-layer feed-forward network with no non-linearity i.e. $f = B$.
- Data is normalized i.e. $E(x) = 0, \mathrm{COV}(x) = I$
- Matrices ***A*** and ***B*** have SVDs: $A = USV^T \quad B = MQP^T$
- ***S*** is a diagonal matrix with singular values ordered by increasing magnitude

Theorem: If the above ***assumptions*** are satisfied, then the optimal ***B*** obtained by solving (6) is a modified pseudo-inverse of ***A***, with $M = V, P = U$ and *Q* a filtered inverse of ***S***:

$$Q = \mathrm{diag}\left(q_m, \ldots, q_m, 1/S_{m+1}, \ldots, 1/S_n\right),$$

$$q_m = \frac{\sum_{i=1}^m S_i}{\sum_{i=1}^m S_i^2 + \frac{\lambda}{1+\lambda}\epsilon^2}$$

with largest entry q_m of multiplicity *m* that depends on ϵ, λ and $\{s_i\}_{i=1}^n$

Revisit: simple ill-conditioned case

$$A = \begin{bmatrix} 1 & 0 \\ 0 & a \end{bmatrix} \text{ and } f = \begin{bmatrix} 1 & 0 \\ 0 & 1/a \end{bmatrix}$$

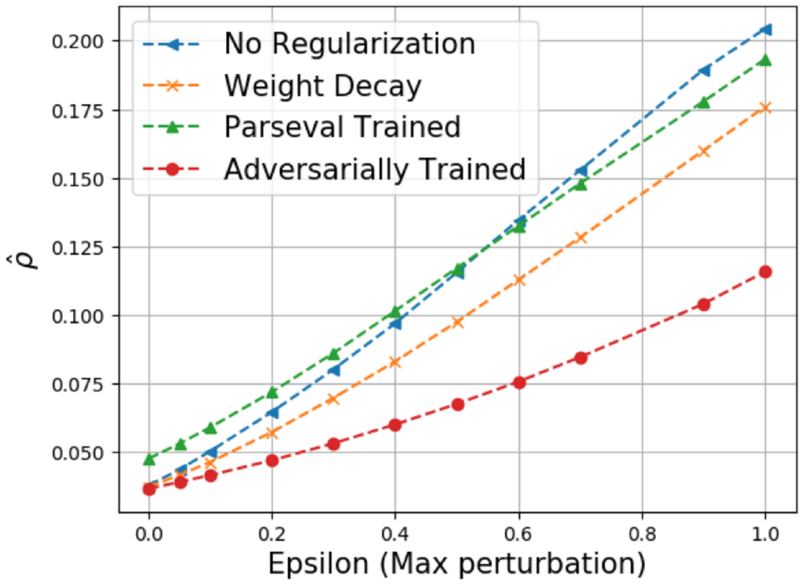
Modified pseudo-inverse after adv. training: $\hat{f} = \begin{bmatrix} 1 & 0 \\ 0 & \frac{a}{a^2+0.5\epsilon^2} \end{bmatrix}$

$$\delta = [0, \epsilon]^T \Longrightarrow \|\hat{f}\delta\| \ll \|f\delta\| \text{ for } a \to 0 \text{ and } \epsilon \nrightarrow 0$$

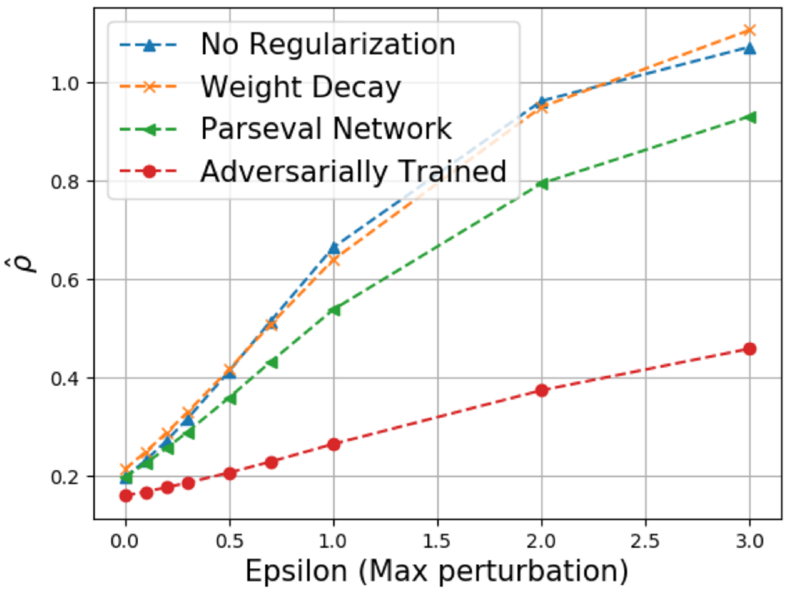
Important points:

- For unperturbed ***y*,** true inverse better than modified inverse.
- But for the true inverse, small perturbation results in severe degradation
- ***Trade-off behavior***

Results for relatively ill-conditioned DCT sub-matrix



MNIST



CelebA

Take-home

- Conventionally trained (and even regularized) deep-learning-based image reconstruction networks are ***vulnerable*** to adversarial perturbations in the measurement.
- Proposed a min-max formulation to build ***robust*** DL-based image reconstruction.
- To make this tractable, we introduced an ***auxiliary network*** to generate adversarial examples for which the image recon network tries to minimize the recon loss.
- Analyzed a simple linear network - found that min-max formulation results in singular-value filter regularized solution mitigating the effect of adversarial examples due to ill-conditioning of the measurement operator.
- Empirical results show that behavior depends on the ***conditioning*** of the measurement operator.