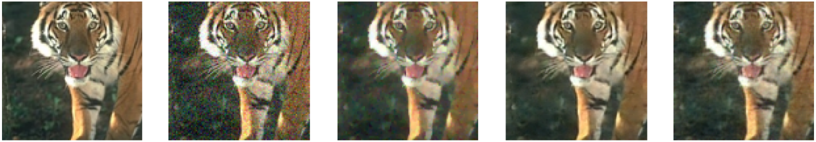
Adversarial Regularizers in Inverse Problems[1]



Ground truth | Noisy Image | TV reconstruction | Denoising NN | Adversarial Reg.

Seminar: Deep Learning in Inverse Problems
Institut für Mathematik, WiSe2021/2022
Viktor Stein, 07.01.2022

[1]Lunz et al.: "Adversarial Regularizers in Inverse Problems", *NeurIPS*, 2018

Table of Contents

Inverse problem: recover image $x \in X$ from measurements

$$y = Ax + \varepsilon \in Y, \qquad \text{where } A\colon X \to Y \textit{ linear.}$$

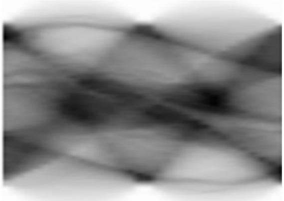
$\boldsymbol{A} \cdot$  $+\boldsymbol{\epsilon} \; =$ 

Fig. 1: In computer tomography, the noisy sinogram y is obtained by applying the forward operator A to the true image x and adding noise ε.

Inverse problem: recover image $x \in X$ from measurements

$$y = Ax + \varepsilon \in Y, \qquad \text{where } A \colon X \to Y \textit{ linear.}$$

Regularisation approach: for $\lambda > 0$ solve

$$\operatorname*{argmin}_{x \in X} \underbrace{\|Ax - y\|^2}_{\text{data term}} + \lambda \underbrace{f(x)}_{\substack{\text{a-priori}\\ \text{info}}}$$

Inverse problem: recover image $x \in X$ from measurements

$$y = Ax + \varepsilon \in Y, \qquad \text{where } A\colon X \to Y \textit{ linear.}$$

Regularisation approach: for $\lambda > 0$ solve

$$\operatorname*{argmin}_{x \in X} \underbrace{\|Ax - y\|^2}_{\text{data term}} + \lambda \underbrace{\Psi_\Theta(x)}_{\substack{\text{a-priori}\\ \text{info}}}$$

- **Key idea**: Replace regularizer f (for example $\mathrm{TV}(x) = \|\nabla x\|_1$) by neural network (NN) Ψ_Θ.
- Goal: Find best parameters Θ for NN architecture Ψ.

Why is this a good idea?

- Combination of mathematical methods (model-based) with deep learning (data driven).

Why is this a good idea?

- Only training NN to learn $y \mapsto x$ works badly if A complicated or few data available (e.g. MRI, CT).
- NN is independent of A ("plug-and-play"-structure).
- Is unsupervised.
- scales to high-dimensional parameter space.
- hand-crafted regularizers do not reflect true prior.

Table of Contents

NN as a critic

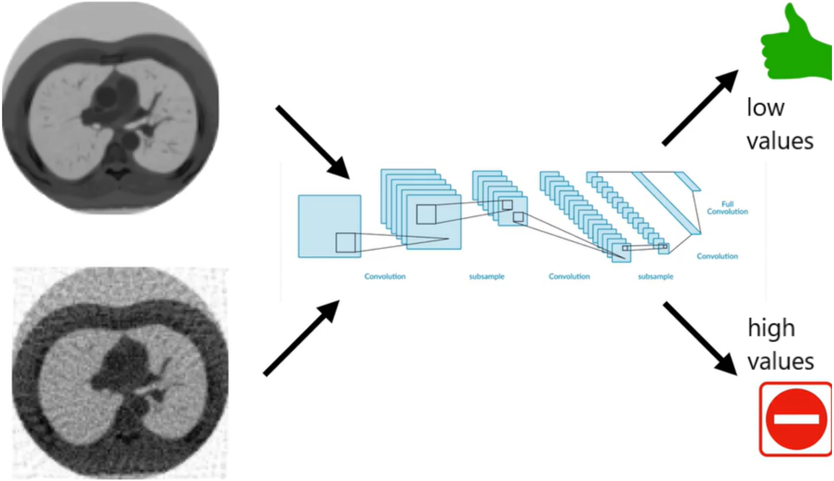


Fig. 1: NN learns to *discriminate* between distribution of true data and of unregularized reconstructions. (Source: https://youtu.be/ZfYm6Om4hec, modified)

Regularizers as critics

Let $\mathbb{P}_r$ and $\mathbb{P}_Y$ be the distributions of true images / measurements.

– $\mathbb{P}_Y$ can be mapped from Y to X via $A_\delta^\dagger := (AA^T + \delta I)^{-1} A^T$, a (regularized) pseudo-inverse of A.

Space	Y		X
Point	$y = Ax + \varepsilon$	$\longmapsto$	$A_\delta^\dagger y \approx x + A_\delta^\dagger \varepsilon$
Distribution	$\mathbb{P}_Y$		$\mathbb{P}_n := (A_\delta^\dagger)_\# \mathbb{P}_Y$

+ $A_\delta^\dagger$ is known for many A, e.g. Fourier, Radon.[2]

– Applying $A_\delta^\dagger$ to y amplifies noise.

[2] Jin et al.: "Deep Convolutional Neural Network for Inverse Problems in Imaging", *IEEE Transactions on Image Processing*, 2017

Regularizers as critics

Goal: Ψ_Θ takes high values on samples from $\mathbb{P}_n$ and low values on samples of $\mathbb{P}_r$. $\rightsquigarrow$ We want to find

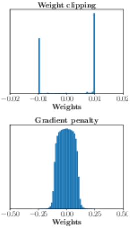
$$\operatorname*{argmin}_{\Theta} \mathbb{E}_{P_X \sim \mathbb{P}_r}[\Psi_\Theta(P_X)] - \mathbb{E}_{P_X \sim \mathbb{P}_n}[\Psi_\Theta(P_X)].$$

For $\mu > 0$, define the loss functional

$$\mathbb{E}_{P_X \sim \mathbb{P}_r}[\Psi_\Theta(P_X)] - \mathbb{E}_{P_X \sim \mathbb{P}_n}[\Psi_\Theta(P_X)] + \mu \underbrace{\mathbb{E}_{P_X \sim \tau}\left[\left((\|\nabla_x \Psi_\Theta(P_X)\| - 1)_+\right)^2\right]}_{\text{penalizes } \mathrm{Lip}(\Psi_\Theta) > 1 \;\; ^3},$$

whose minimizer Ψ_{Θ^*} approximates the maximizer f of

$$W_1(\mathbb{P}_r, \mathbb{P}_n) = \sup_{\substack{f \in 1\text{-Lip} \\ f\colon X \to \mathbb{R}}} \mathbb{E}_{P_X \sim \mathbb{P}_n}[f(P_X)] - \mathbb{E}_{P_X \sim \mathbb{P}_r}[f(P_X)].$$



[3] Gulrajani et al.: *Improved Training of Wasserstein GANs*, *NeurIPS*, 2017

Learning a regularizer: Algorithm

Data: Gradient penalty μ, batch size m, pseudo-inverse $A_\delta^\dagger$.

while Θ *has not converged* **do**

  for $i \in \{1, \dots, m\}$ **do**

    Sample ground truth image $x_i^{(r)} \sim \mathbb{P}_r$, measurement $y_i \sim \mathbb{P}_Y$, random number $\theta \sim U[0, 1]$;

    $x_i^{(n)} \leftarrow A_\delta^\dagger y_i$; `// Noisy reconstruction`

    $x_i \leftarrow \theta x_i^{(r)} + (1-\theta) x_i^{(n)}$; `// Random convex combination`

    $L_i \leftarrow \Psi_\Theta(x_i^{(r)}) - \Psi_\Theta(x_i^{(n)}) + \mu \left[(\|\nabla_{x_i} \Psi_\Theta(x_i)\|_2 - 1)_+ \right]^2$;

    `// Loss functional`

  $\Theta \leftarrow \mathsf{Adam}\left(\nabla_\Theta \sum_{i=1}^m L_i\right)$; `// Improved SGD`

Once NN trained, solve $\operatorname{argmin}_{x \in X} \|Ax - y\|_2^2 + \lambda \Psi_{\Theta^*}(x)$ with GD, choosing $\lambda = 2\,\mathbb{E}_{\varepsilon \sim p_n} \|A^* \varepsilon\|_2$, where A^* is the adjoint of A and p_n is the noise distribution.

Table of Contents

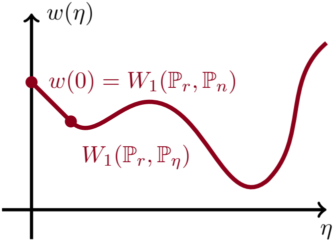
Regularizer's impact on $\mathbb{P}_n$

Assume X is a Hilbert space, Ψ_{Θ^*} is 1-Lipschitz, $\mathbb{P}_n$-a.e. differentiable and attains maximum in $W_1(\mathbb{P}_r, \mathbb{P}_n)$.

- Let x be noisy reconstruction sampled from $\mathbb{P}_n$.
- Image obtained by performing η-sized GD step over Ψ_{Θ^*} is $g_\eta(x) \coloneqq x - \eta \cdot \nabla_x \Psi_{\Theta^*}(x)$. Define $\mathbb{P}_\eta \coloneqq (g_\eta)_\# \mathbb{P}_n$.

Goal: $W_1(\mathbb{P}_r, \mathbb{P}_\eta) < W_1(\mathbb{P}_r, \mathbb{P}_n)$ for small $\eta > 0$, that is, $w'(0) < 0$, where $w(\eta) \coloneqq W_1(\mathbb{P}_r, \mathbb{P}_\eta)$.

$\rightsquigarrow$ Then gradient step is meaningful.



Regularizer's impact on $\mathbb{P}_n$

One can prove, provided the derivative exists, that

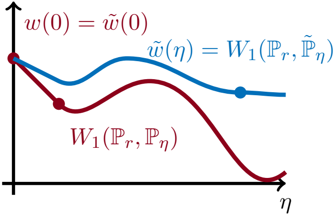
$$w'(0) = -\mathbb{E}_{P_X \sim \mathbb{P}_n}\left[\|\nabla_x \Psi_{\Theta^*}(P_X)\|^2\right] = -1.$$

Mild assumptions $\implies$ $\|\nabla_x \Psi_{\Theta^*}\|^2 \equiv 1$ $\mathbb{P}_n$-a.e.

$\leadsto$ Our loss has optimal W_1-decay rates:

for any other regularizer $f \colon X \to \mathbb{R}$ with $\|\nabla_x f(x)\| \leq 1$ we define $\tilde{g}_\eta(x) \coloneqq x - \eta \cdot \nabla f(x)$ and $\tilde{\mathbb{P}}_\eta \coloneqq (\tilde{g}_\eta)_\#(\mathbb{P}_n)$. Then

$$\tilde{w}'(0) \geq -1 = w'(0).$$



Weak data manifold assumption

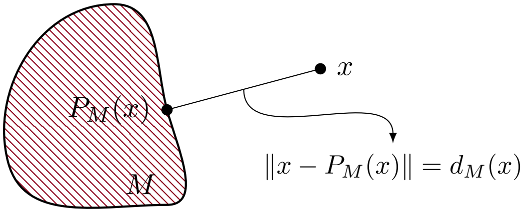
Assume $\mathbb{P}_r$ is supported on the weakly compact set $M \subset X$.

Regularizer should encode prior knowledge about $\mathbb{P}_r \rightsquigarrow$ penalize points far away from M. One choice:

$$d_M \colon X \to [0, \infty), \qquad x \mapsto \min_{y \in M} \|x - y\|.$$

The data manifold projection is

$$P_M \colon X \supset D \to M, \qquad x \mapsto \operatorname*{argmin}_{y \in M} \|x - y\|.$$



Distance function maximizes loss functional

Theorem (Distance function maximizes loss functional)

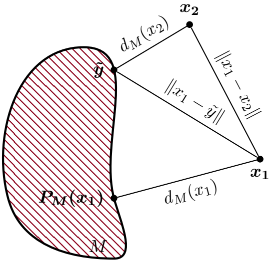
If additionally $(P_M)_\#(\mathbb{P}_n) = \mathbb{P}_r$ (low noise), then

$$d_M \in \underset{f \in 1\text{-Lip}}{\operatorname{argmax}} \; \mathbb{E}_{P_X \sim \mathbb{P}_n}[f(P_X)] - \mathbb{E}_{P_X \sim \mathbb{P}_r}[f(P_X)].$$

– Minimizer is *not unique*, as we can alter f outside of the convex hull of $\operatorname{supp}(\mathbb{P}_r) \cup \operatorname{supp}(\mathbb{P}_n)$ (provided it remains 1-Lipschitz).

1. d_M is 1-LIPSCHITZ: Let $x_1, x_2 \in X$ and $\tilde{y} \coloneqq P_M(x_2)$. Then

$$\begin{aligned} d_M(x_1) - d_M(x_2) &= \min_{y \in M} \|x_1 - y\| - \min_{y \in M} \|x_2 - y\| \\ &= \min_{y \in M} \|x_1 - y\| - \|x_2 - \tilde{y}\| \\ &\leq \|x_1 - \tilde{y}\| - \|x_2 - \tilde{y}\| \\ &\overset{\triangle\neq^{-1}}{\leq} \|x_1 - \tilde{y} - (x_2 - \tilde{y})\| \\ &= \|x_1 - x_2\| \end{aligned}$$



by the inverse triangle inequality.

Now exchange x_1 and x_2.

2. d_M **attains maximum:** Let $f\colon X \to \mathbb{R}$ be 1-LIPSCHITZ. As we assumed $(P_M)_\#(\mathbb{P}_n) = \mathbb{P}_r$ $(\star)$, we have

$$\begin{aligned}
\mathbb{E}_{P_X \sim \mathbb{P}_n}[f(P_X)] - \mathbb{E}_{P_X \sim \mathbb{P}_r}[f(P_X)] &\overset{(\star)}{=} \mathbb{E}_{P_X \sim \mathbb{P}_n}[f(P_X) - f(P_M(P_X))] \\
&\overset{\text{Lip}}{\leq} \mathbb{E}_{P_X \sim \mathbb{P}_n}[\|P_X - P_M(P_X)\|] \\
&\overset{(\ddagger)}{=} \mathbb{E}_{P_X \sim \mathbb{P}_n}[d_M(P_X)] \\
&= \mathbb{E}_{P_X \sim \mathbb{P}_n}[d_M(P_X) - d_M(P_M(P_X))] \\
&\overset{(\star)}{=} \mathbb{E}_{P_X \sim \mathbb{P}_n}[d_M(P_X)] - \mathbb{E}_{P_X \sim \mathbb{P}_r}[d_M(P_X)],
\end{aligned}$$

because the distance between x and $P_M(x)$ is $d_M(x)$ $(\ddagger)$.

Existence and Stability

Let f be weakly lower-semicontinuous and 1-Lipschitz with $\|f(x)\| \xrightarrow{\|x\|\to\infty} \infty$ and A be continuous.

Remark. The function $f := d_M$ fulfills the above assumptions.

Theorem (Existence of minimiser)

There exist a $x^* \in \operatorname{argmin}_{x\in X} \|Ax - y\|^2 + \lambda f(x)$ *for* $\lambda > 0$.

Theorem (Weak stability of the data term)

Let $(y_n)_{n\in\mathbb{N}} \subset Y$ *converge to* y *and* $(x_n)_{n\in\mathbb{N}} \subset X$ *be a sequence of minimizers of* $\|A\cdot - y_n\|^2 + \lambda f$. *Then* $(x_n)_{n\in\mathbb{N}}$ *has a weakly convergent subsequence, whose limit* x^* *minimizes* $\|A\cdot - y\|^2 + \lambda f$.

Proof. Consult [3, Appendix, p. 10-12].

Table of Contents

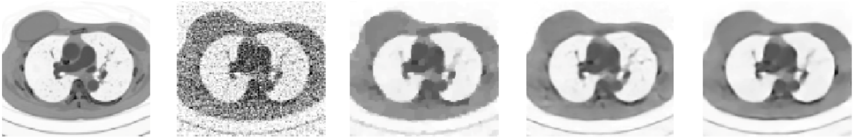
Results: Denoising

(a) Ground Truth (b) Noisy Image (c) TV (d) Denoising N.N. (e) Adversarial Reg

	Noisy Image	TV (supervised)	**Denoising NN (unsupervised)**	*Adversarial Regularizer*
PSNR (dB)	20.3	26.3	**28.8**	*28.2*
SSIM	.534	.836	**.908**	*.892*

Fig. 2: Performance on a denoising task ($A = \mathrm{id}$) on the BSDS dataset.

Results: Reconstruction



(a) Ground Truth (b) FBP (c) TV (d) Post-Processing (e) Adversarial Reg

		Model	- based	Supervised	*Unsupervised*
		(b)	(c)	(d)[4]	*(e)*
high noise	**PSNR (dB)**	14.9	27.7	31.2	*30.5*
	SSIM	.227	.890	.936	*.927*
low noise	**PSNR (dB)**	23.3	30.0	33.6	*32.5*
	SSIM	.604	.924	.955	*.946*

Fig. 3: Performance on a complicated reconstruction task.

[4] Jin et al.: "Deep Convolutional Neural Network for Inverse Problems in Imaging", *IEEE Transactions on Image Processing*, 2017

Outlook: Extensions

– **Local regularizers.** Samples many patches of pixels from image, then value of regularizer is average of values on patches. Architecture: choose convolution layers followed by global average pooling.
  $\rightsquigarrow$ less training data needed.
– **Recursive training.** When solving variational problem, regularizer "sees" partially reconstructed images (not ground truth, but not with learned noise either).

Outlook: Extensions

- **Local regularizers.** Samples many patches of pixels from image, then value of regularizer is average of values on patches. Architecture: choose convolution layers followed by global average pooling.
  $\rightsquigarrow$ less training data needed.
- **Recursive training.** When solving variational problem, regularizer "sees" partially reconstructed images (not ground truth, but not with learned noise either). Add those images to training data $\rightsquigarrow$ NN learns from own outputs. But: delicate choice, which images to add.

Conclusion

- Key idea: replace regularizer by NN.
- Training algorithm for NN inspired heavily by WGAN.
- Solve regularised problem with GD $\rightsquigarrow$ optimal W_1 decay-rates.
- Under weak assumption, optimal NN has favourable properties (similar to $d_{\mathrm{supp}(\mathbb{P}_r)}$).
- This approach outperforms TV-regularisation and is unsupervised.

Thank you for your attention!

References

[1] Ishaan Gulrajani et al. *Improved Training of Wasserstein GANs*. 2017. arXiv: 1704.00028 [cs.LG].

[2] Kyong Hwan Jin et al. "Deep Convolutional Neural Network for Inverse Problems in Imaging". In: *IEEE Transactions on Image Processing* 26.9 (2017), pp. 4509–4522.

[3] Sebastian Lunz et al. "Adversarial Regularizers in Inverse Problems". In: *NeurIPS* abs/1805.11572 (2018). arXiv: 1805.11572.

- Talk at SIAM CSE19: https://www.pathlms.com/siam/courses/10878/sections/14350/video_presentations/127337.
- Poster at NIPS2018: https://nips.cc/Conferences/2018/ScheduleMultitrack?event=11813.