Adversarial Regularizers in Inverse Problems

SIAM CSE 2019 - Spokane, Washington

Sebastian Lunz

February 25, 2019

University of Cambridge

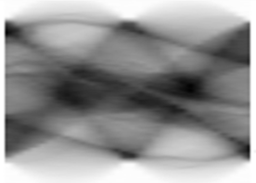
Inverse Problems

- Given measurement of the form

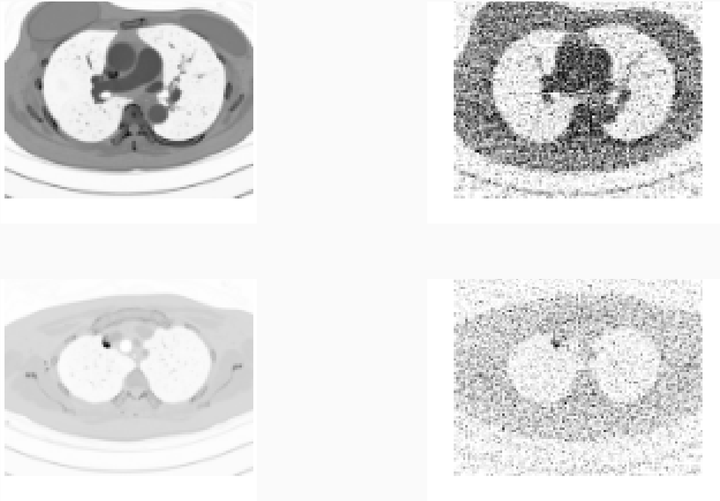
$$y = Ax + \epsilon$$

- Noise distribution $\epsilon \sim \mathbb{P}_e$ known
- Recover x from data y
- Naive reconstruction for *ill-posed* problem is unstable

Computed Tomography

$A \cdot$  $+\epsilon \;=$ 

Ill-posedness



Image

Reconstruction

Variational Regularization

- Reconstruct by solving variational problem

$$\text{argmin}_x \|Ax - y\|^2 + \lambda R(x)$$

- Motivated by MAP interpretation

$$\text{argmax}_x\, p(x|y) = \text{argmin}_x \frac{1}{2\sigma^2}\|Ax - y\|^2 - \log p(x)$$

- $R(x)$ encodes *prior* knowledge about the reconstruction.
- Corresponds to Gibbs prior $p(x) \sim \exp(-R(x))$
- Total variation regularization

$$TV(x) = \int \|\nabla x(t)\|_1 \, \mathrm{dt}$$

Deep Learning in Inverse Problems

Methods based on minimizing $\|\Psi_\Theta(y) - x\|$.

- Fully learned inversion
- Postprocessing
- Recurrent Inference Machines
- Learned gradient descent
- Learned PDHG

Learning a regularization functional

- Regularization by Denoising (RED):
- Deep Image Priors: Regularization by Architecture.
- GAN Image Priors: Projection onto Image of Generative Model.
- NETT

Distributional losses

- Heuristic: Regularization functional suppresses characteristic noise in reconstruction
- *Distribution* of Reconstructions should be close to *distribution* of ground truth images.
- Denote by $\mathbb{P}_r$ distribution of ground truth images, by $\mathbb{P}_Y$ measurement distribution.
- In practice, have access to the empirical associated to $\mathbb{P}_r$ and $\mathbb{P}_Y$.
- Pull back distribution $\mathbb{P}_Y$ via pseudo-inverse $A^\dagger$:

$$\mathbb{P}_n := A^\dagger_\sharp \mathbb{P}_Y$$

- Aim: Construct regularization functional that aligns reconstruction distribution with $\mathbb{P}_r$.

Wasserstein Loss

- The Wasserstein-1 distance between two distributions $\mathbb{P}_n$ and $\mathbb{P}_r$ is defined as

$$\mathrm{Wass}(\mathbb{P}_n, \mathbb{P}_r) := \inf_{\gamma \in \Pi(\mathbb{P}_n, \mathbb{P}_r)} \|x_1 - x_2\| \, \mathrm{d}\gamma(x_1, x_2)$$

- Minimal path length to 'transport' mass $\mathbb{P}_n$ to $\mathbb{P}_r$.
- The *Kantorovich duality* allows to equivalently characterize via

$$\mathrm{Wass}(\mathbb{P}_n, \mathbb{P}_r) = \sup_{f \ 1-Lip} \mathbb{E}_{X \sim \mathbb{P}_n} f(X) - \mathbb{E}_{X \sim \mathbb{P}_r} f(X)$$

- Denote now by f^* an optimizer to the dual formulation of the Wasserstein distance.

Decreasing the Wasserstein Distance

What happens if we do gradient descent over f^*?

- Definitions

$$g_\eta(x) := x - \eta \cdot \nabla_x f^*(x).$$

$$\mathbb{P}_\eta := (g_\eta)_\# \mathbb{P}_n$$

- Assume that $\eta \mapsto \mathrm{Wass}(\mathbb{P}_r, \mathbb{P}_\eta)$ admits a left and a right derivative at $\eta = 0$, and that they are equal. Then,

$$\frac{\mathrm{d}}{\mathrm{d}\eta} \mathrm{Wass}(\mathbb{P}_r, \mathbb{P}_\eta)|_{\eta=0} = -\mathbb{E}_{X \sim \mathbb{P}_n}\left[\|\nabla_x \Psi_\Theta(X)\|^2\right] = -1.$$

- This is the fastest decrease in Wasserstein distance for any regularization functional with normed gradients

Approximating f^*

- Idea in Wasserstein GANs: Use a neural network (critic), to approximate f^*.
- We employ a convolutional architecture for the network $\Psi_\Theta : \mathbb{R}^{n\times m} \to \mathbb{R}$.
- Train the Network with the loss

$$\mathbb{E}_{X\sim\mathbb{P}_r}[\Psi_\Theta(X)] - \mathbb{E}_{X\sim\mathbb{P}_n}[\Psi_\Theta(X)] + \lambda\cdot\mathbb{E}\left[(\|\nabla_x\Psi_\Theta(X)\|_* - 1)_+^2\right].$$

- Relaxation of Lipschitz constraint into penalty term (WGAN-GP)

Algorithm Summary

- Train a convolutional neural network as regularization functional by minimizing

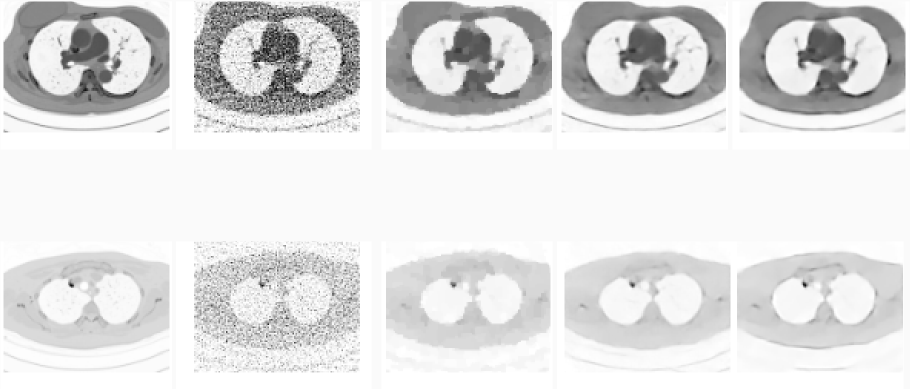
$$\mathbb{E}_{X\sim\mathbb{P}_r}\left[\Psi_\Theta(X)\right] - \mathbb{E}_{X\sim\mathbb{P}_n}\left[\Psi_\Theta(X)\right] + \lambda \cdot \mathbb{E}\left[(\|\nabla_x \Psi_\Theta(X)\|_* - 1)_+^2\right].$$

- Deploy the network on the inverse problem by solving

$$\operatorname{argmin}_x \|Ax - y\|^2 + \lambda \Psi_\Theta(x)$$

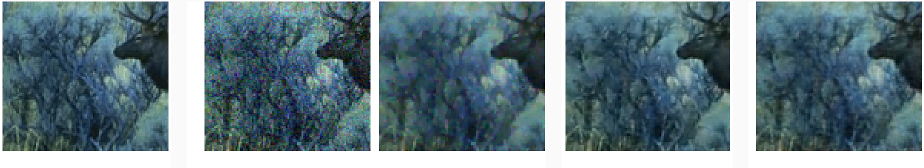
- Gradient of $\Psi_\Theta(x)$ available via backpropagation, prox not available
- Many optimization schemes possible. We employed gradient descent.

Computational Results



Ground truth | Filtered Backprojection | Total Variation | Post-Processing | Adversarial Regularizer

Denoising on BSDS



Ground Truth | Noisy Image | TV | Denoising AE | Adversarial Reg.

Ellipse data

Comparing to TV on synthetic ellipse data. The images are piecewise constant.

Parameter Heuristics

- Normed gradients of regularization functionals allow to estimate parameters easily from noise level
- Heuristic: Ground truth is a critical point of regularization functional
- Leads to formula

$$\lambda = 2\ \mathbb{E}_{e \sim p_n} \|A^* e\|_2,$$

- Note: Unlike direct unrolling schemes, can underregularize by choosing λ small.

How does this functional look like?

- Data Manifold Assumption: The measure $\mathbb{P}_r$ is supported on the weakly compact set $\mathcal{M}$, i.e. $\mathbb{P}_r(\mathcal{M}^c) = 0$
- Denote by $P_{\mathcal{M}} : D \to \mathcal{M}$, $x \to \operatorname{argmin}_{y \in \mathcal{M}} \|x - y\|$ the projection onto the data manifold
- Projection Assumption: $(P_{\mathcal{M}})_{\#}(\mathbb{P}_n) = \mathbb{P}_r$
- Corresponds to a low-noise assumption - noise level low in comparison to manifold curvature

Manifold Lemma

The distance function to the data manifold

$$d_{\mathcal{M}}(x) := \min_{y \in \mathcal{M}} \|x - y\|$$

is a maximizer to the functional

$$\text{argmax}_{f \in 1-Lip} \mathbb{E}_{X \sim \mathbb{P}_n} f(X) - \mathbb{E}_{X \sim \mathbb{P}_r} f(X)$$

Stability

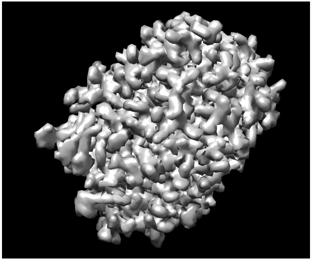
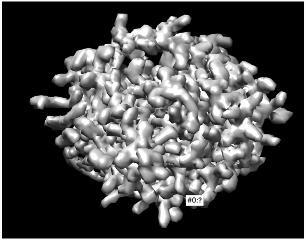
- Do not assume Ψ_Θ to be bounded from below
- Instead have Lipschitz assumption on Ψ_Θ
- Make usual sufficient assumptions for coercivity
- Let y_n be a sequence in Y with $y_n \to y$ in the norm topology and denote by x_n a sequence of minimizers of the functional

$$\operatorname{argmin}_{x \in X} \|Ax - y_n\|^2 + \lambda \Psi_\Theta(x)$$

  Then x_n has a weakly convergent subsequence and the limit x is a minimizer of $\|Ax - y\|^2 + \lambda \Psi_\Theta(x)$.

Unpaired Training Data

- Note that paired data of the form (x_i, y_i) from the joint distribution was never used.
- Instead, have access to both marginals x_i samples from $\mathbb{P}_r$, y_i samples from $\mathbb{P}_Y$.
- Use case: Real measurement data y, containing unmodeled effects
- Single Particle Analysis reconstruction: No need to for registration

Advantages & Disadvantages

Advantages

- Unpaired Training Data
- Greater Flexibility (i.e. EM algorithm)
- Interpretability. Non-implicit prior.
- Very quick training, small data sets
- Stability results

Disadvantages

- Slower evaluation than some direct methods (Post-Processing)
- Worse performance in term of PSNR