Learned Convex Regularizers for Inverse Problems

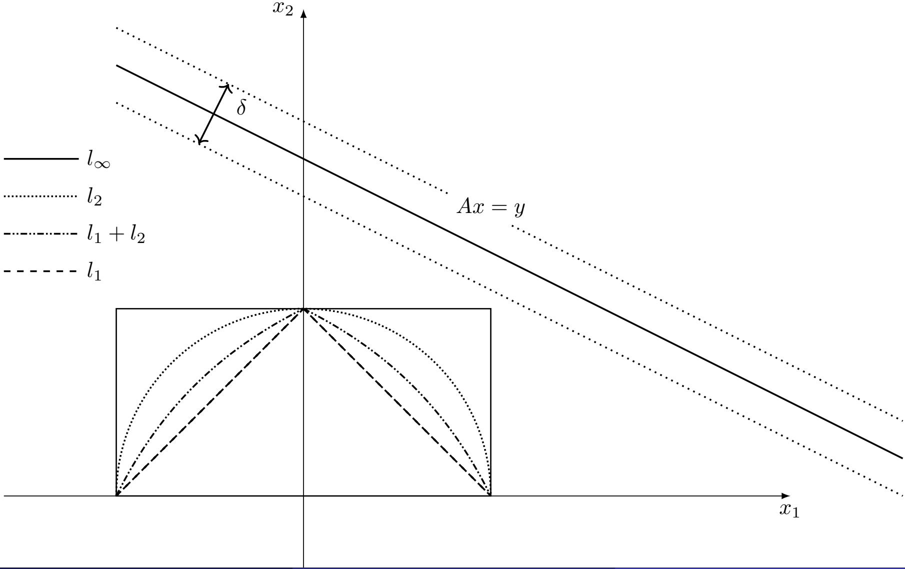
Marko Lalovic

Mathematical Methods for Medical Imaging Seminar

July 12, 2022

Motivational Examples

Example: Find solution of $y = Ax$, $A : \mathbb{R}^n \to \mathbb{R}^m$, $m < n$



Motivational Examples - Continued

Examples:

- Generate realistic images in HD [1]
- Extract main building blocks of images [2]

1024 × 1024 images generated using the CELEBA-HQ dataset

Corrupted Reconstructed True

[1] Tero Karras et al. "Progressive Growing of GANs for Improved Quality, Stability, and Variation", arxiv 2018

[2] Longfei Liu et al. "X-GANs: Image Reconstruction Made Easy for Extreme Cases", arxiv 2018

Introduction

Main ideas [3]:

- Use deep learning to solve inverse problems
- With adaptation: **enforce convexity on the learned regularizer**
- Be pragmatic: lack of large amount of paired data

More motivation:

- Can show some convergence guarantees
- Design provable reconstruction algorithms
- This is still less explored and poorly understood

[3] Subhadip Mukherjee, Sören Dittmer, Zakhar Shumaylov, Sebastian Lunz, Ozan Öktem, Carola-Bibiane Schönlieb "Learned Convex Regularizers for Inverse Problems", arxiv 2021

Inverse Problems in Computed Tomography

- Estimate model parameters $\boldsymbol{x}^* \in \mathbb{X}$ from data

$$\boldsymbol{y} = \mathcal{A}(\boldsymbol{x}^*) + \boldsymbol{e} \in \mathbb{Y}$$

- Forward operator $\mathcal{A} : \mathbb{X} \to Y$
- $\mathbb{X}, \mathbb{Y}$ Hilbert spaces (after discretization, $\mathbb{X} = \mathbb{R}^n$ and $\mathbb{Y} = \mathbb{R}^m$)

Variational Reconstruction

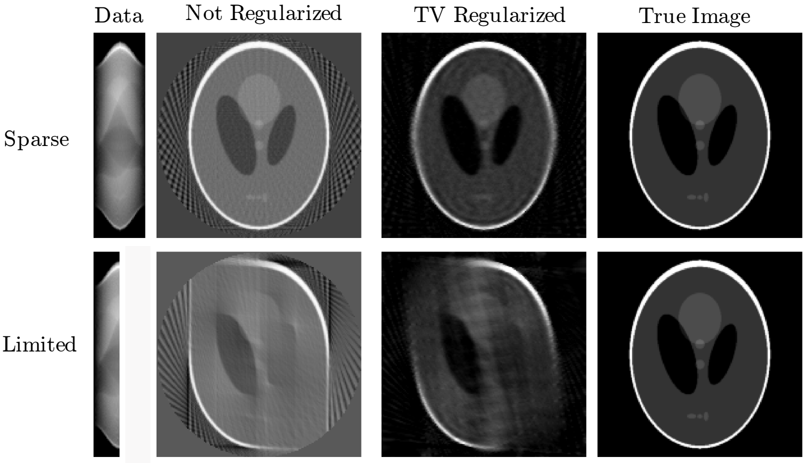
$$\min_{\boldsymbol{x} \in \mathbb{X}} \mathcal{L}_{\mathbb{Y}} \left(\mathcal{A}(\boldsymbol{x}), \boldsymbol{y} \right) + \lambda \mathcal{R}(\boldsymbol{x})$$

Where:

- $\mathcal{L}_{\mathbb{Y}} : \mathbb{Y} \times \mathbb{Y} \to \mathbb{R}$ measures data fidelity
- $\mathcal{R} : \mathbb{X} \to \mathbb{R}$ penalizes undesirable solutions

Classical Reconstruction Methods

- Image-size: 160 x 160, angles: 40 (20), degrees: 0 - 180 (90) [4]
- TV promotes sparsity in the image gradient: $\mathcal{R}(\boldsymbol{x}) = \|\nabla \boldsymbol{x}\|_1$



[4] https://github.com/markolalovic/learned-convex-regularizers

Statistical Bayesian Formulation

- $\boldsymbol{x}^*$ and $\boldsymbol{y}$ are modeled as realizations of X and Y, which are $\mathbb{X}$- and $\mathbb{Y}$-valued random variables, respectively and

$$Y = \mathcal{A}(X) + \boldsymbol{e}$$

- Data likelihood: $\pi_{Y|X}(Y = y | X = \boldsymbol{x}^*) = \pi_{\text{noise}}\left(\boldsymbol{y} - \mathcal{A}(\boldsymbol{x}^*)\right)$
- Prior: $\pi_X(\boldsymbol{x})$
- Posterior distribution:

$$\pi_{X|Y}(\boldsymbol{x}|\boldsymbol{y}) = \frac{\pi_{Y|X}(\boldsymbol{y}|\boldsymbol{x})\,\pi_X(\boldsymbol{x})}{Z(y)}$$

Supervised Learning

- Training data: i.i.d. samples $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^N$ from the joint distribution $\pi_{X,Y}$
- Parametric reconstruction operator: $G_\theta : \mathbb{Y} \to \mathbb{X}$, $\theta \in \Theta$
- Loss function: $\mathcal{L}_\mathbb{X} : \mathbb{X} \times \mathbb{X} \to \mathbb{R}_+$
- Risk minimization: $\min\limits_{\theta \in \Theta} \mathbb{E}_{\pi_{X,Y}} \left[\mathcal{L}_\mathbb{X}(X, \mathcal{G}_\theta(Y)) \right]$
- Empirical risk minimization: $\min\limits_{\theta \in \Theta} \sum\limits_{i=1}^N \mathcal{L}_\mathbb{X}(x_i, \mathcal{G}_\theta(y_i))$
- Example:
  - Using 0-1 loss and computing the mode, leads to so-called *maximum a-posterior probability* (MAP) estimate
  - Using Gibbs-type prior $\pi_X(\boldsymbol{x}) \propto \exp\left(-\lambda \mathcal{R}(\boldsymbol{x})\right)$ is equivalent to classical variational reconstruction framework.

Unsupervised Learning

Proposed Approach

Keep the variational framework and only try to learn a suitable regularizer from the training data

Where:

- Training data: i.i.d. samples $\{\boldsymbol{x}_i\}_{i=1}^{N_X}$ from π_X and $\{\boldsymbol{y}_i\}_{i=1}^{N_Y}$ from π_Y
- Empirical risk minimization approach cannot be applied
- Statistical characterization is an open problem

Adversarial Learning

We want to:

- Train regularization functional
- To suppress characteristic artifacts in the reconstruction
- Because of the ill-posedness of the forward operator

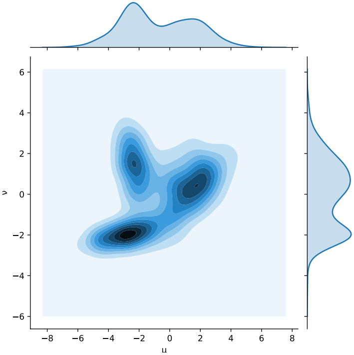
How to do this:

- Minimize distributional distance between:
  - True images, for example by using phantom images
  - Naive reconstructions, by using the pseudo-inverse on the data

Wasserstein Distance

- The Wasserstein-1 distance between two distributions $\mathbb{P}_1$ and $\mathbb{P}_2$

$$\text{Wass}(\mathbb{P}_1, \mathbb{P}_2) := \inf_{\gamma \in \Pi(\mathbb{P}_1, \mathbb{P}_2)} \|x_1 - x_2\| \, d\gamma(x_1, x_2)$$



Minimal path length to transport mass $\mathbb{P}_1$ to $\mathbb{P}_2$ [5]

[5] (By Lambdabadger licensed under CC BY-SA 4.0)

Adversarial Regularizer

- Variational reconstruction: $\min_{x \in \mathbb{X}} \|\mathcal{A}(\boldsymbol{x}) - \boldsymbol{y}\|_2^2 + \lambda \mathcal{R}_\theta(\boldsymbol{x})$
- Two-step sequential approach:
  - Learning:

$$\theta^* = \arg\min_\theta \mathbb{E}_{\pi_X} \left[\mathcal{R}_\theta(X)\right] - \mathbb{E}_{\mathcal{A}^\dagger_\# \pi_Y} \left[\mathcal{R}_\theta(X)\right]$$

subject to $\mathcal{R}_\theta \in 1$ - Lipschitz

  - Reconstruction: $\hat{x} = \arg\min_{\boldsymbol{x} \in \mathbb{X}} \|\mathcal{A}(\boldsymbol{x}) - \boldsymbol{y}\|_2^2 + \lambda \mathcal{R}_{\theta^*}(\boldsymbol{x})$
- The 1-Lipschitz constraint is enforced by adding a gradient-penalty term

$$\lambda_{gp} \, \mathbb{E}_{\pi_{X^{(\epsilon)}}} \left[\left(\left\| \nabla R_\theta \left(X^{(\epsilon)} \right) \right\|_2 - 1 \right)^2 \right]$$

- $X^{(\epsilon)}$ is uniformly sampled on the line-segment between X and $\mathcal{A}^\dagger Y$

Importance of 1-Lipschitz Constraint

- View $\mathcal{R}_\theta$ as a classifier that learns to discriminate π_X from $\pi_{\mathcal{A}^\dagger_\# \pi_Y}$
- Suppose the variational problem is solved via gradient-descent, starting with $\boldsymbol{x}_0$ such that $\nabla_{\boldsymbol{x}} \left(\|\mathcal{A}(\boldsymbol{x} - \boldsymbol{y}\|_2^2 \right)_{\boldsymbol{x}=\boldsymbol{x}_0} = 0$
- x_0 is a sample from $\pi_{\mathcal{A}^\dagger_\# \pi_Y}$ so $\mathcal{R}_\theta(\boldsymbol{x}_0)$ is large
- $\boldsymbol{x} = \boldsymbol{x}_0 - \eta \nabla \mathcal{R}_\theta(\boldsymbol{x}_0)$
- The output of $\mathcal{R}_\theta$ does not change much going from $\boldsymbol{x}_0$ to $\boldsymbol{x}_1$

$$|\mathcal{R}_\theta(\boldsymbol{x}_1) - \mathcal{R}_\theta(\boldsymbol{x}_0)| \leq \|\boldsymbol{x}_1 - \boldsymbol{x}_0\| = \eta \left\|\nabla \mathcal{R}_\theta(\boldsymbol{x}_0\right\|_2 \leq \eta$$

- Preventing learning sharp boundaries

Adversarial Convex Regularizer

- Let $\mathcal{R}_\theta(\boldsymbol{x}) = \mathcal{R}'_\theta(\boldsymbol{x}) + \rho_0 \|x\|_2^2$ where $\mathcal{R}'_\theta$ is convex and Lipschitz

Results:

- Existence and uniqueness: follow by strong-convexity
- Stability: $\hat{x}_\lambda(y)$ is continuous in y, in particular ($\mathcal{A}$ is assumed to be linear and bounded, β_1 is the operator norm)

$$\left\|\hat{x}_\lambda(y^{\delta_1}) - \hat{x}_\lambda(y)\right\|_2 \leq \frac{\beta_1 \delta_1}{\lambda \rho_0} \qquad \text{if} \qquad \left\|y^{\delta_1} - y\right\|_2 \leq \delta_1$$

- Convergence: $\hat{x}_\lambda(y) \to x^\dagger$ if $\lambda \to 0$ and $\frac{\delta}{\lambda} \to 0$ when $\delta = \|\boldsymbol{e}\|_2 \to 0$, where

$$x^\dagger = \arg\min_{x \in \mathbb{X}} \mathcal{R}_\theta \qquad \text{subject to} \quad \mathcal{A}(x) = y^0$$

- Implies existence of convergent sub-gradient algorithm

Adversarial Convex Regularizer - Architecture

From convex theory:

- Let $f_i : \mathbb{R} \to \mathbb{R}$ be convex, then so is $\sum_i \beta_i f_i$ for $\beta_i \geq 0$
- Let $f_1, f_2 : \mathbb{R} \to \mathbb{R}$ be convex, $f_1(x) \leq f_1(y)$ whenever $x \leq y \implies f_1 \circ f_2$ is convex:

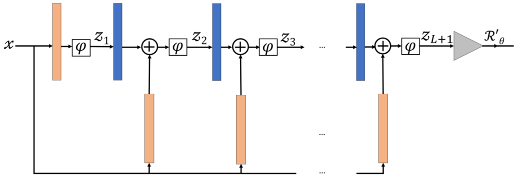
$$
\begin{aligned}
f_2(\lambda x_1 + (1-\lambda)x_2) &\leq \lambda f_2(x_1) + (1-\lambda) f(x_2) \\
\implies (f_1 \circ f_2)(\lambda x_1 + (1-\lambda)x_2) &\leq f_1(\lambda f_2(x_1) + (1-\lambda) f_2(x_2)) \\
&\leq \lambda\,(f_1 \circ f_2)(x_1) + (1-\lambda)(f_1 \circ f_2)(x_2)
\end{aligned}
$$

Input Convex Neural Network (ICNN):

- $z^{(1)}(\boldsymbol{x}) = \phi\left(W_x^{(1)} x + b^{(1)}\right)$
- ϕ acts component-wise such as Rectified Linear Unit (ReLU) $x \mapsto \max(0, x)$ is convex and monotonically non-decreasing

Adversarial Convex Regularizer - Architecture Contd.

- $z^{(2)}(\boldsymbol{x}) = \phi\left(W_z^{(1)} z^{(1)}(x) + W_x^{(2)} x + b^{(2)}\right)$, $W_z^{(1)} \geq 0$, is convex in x
- $z^{(i+1)}(\boldsymbol{x}) = \phi\left(W_z^{(i)} z^{(i)}(x) + W_x^{(i+1)} x + b^{(i+1)}\right)$, $i = 1, 2, \ldots, L = 10$
- $\mathcal{R}_\theta = \sum_j z_j^{(L+1)}(x) + \rho_0' \sum_{k=1}^{M} \left\|U^{(k)} x\right\|_1 + \rho_0 \left\|x\right\|_2^2$
- $\sum_j z_j^{(L+1)}(x)$ is convex and filter-bank term is convex, norms are penalized to impose 1-Lipschitz condition
- The squared ℓ_2 term makes the regularizer strongly-convex



Convergence of sub-gradient method

- We have objective functional of the form:

$$J(\boldsymbol{x}) = \underbrace{\|\mathcal{A}(\boldsymbol{x}) - \boldsymbol{y}\|_2^2 + \lambda\rho_0 \|\boldsymbol{x}\|_2^2}_{f(\boldsymbol{x}) \text{ smooth, strongly-convex}} \quad + \quad \underbrace{\lambda\mathcal{R}'_\theta(\boldsymbol{x})}_{\text{g}(\mathbf{x}) \text{ convex, Lipschitz}}$$

- The sub-gradient method

$$x_{k+1} = x_k - \eta_k \left(\nabla f(x_k) + u_k\right) \quad \text{where} \quad u_k \in \partial g(x_k)$$

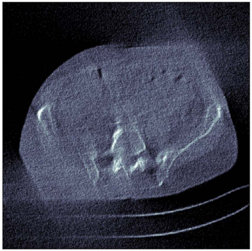
- Converges:
  - Let $e_k = \|x_k - \hat{x}\|_2^2$, derive the inequality $e_{k+1} \leq e_k - \text{Quant.}(\lambda, \rho_0, L_\nabla)$, if

$$\eta_k = \lambda\rho_0 \frac{\|x_k - \hat{x}\|_2^2}{\|\nabla f(x_k) + u_k\|_2^2}$$

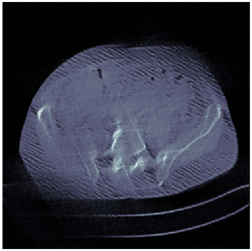
  - Take the limit on both sides, limit exists by monotonicity and boundedness from below

Limited-Angle CT Results

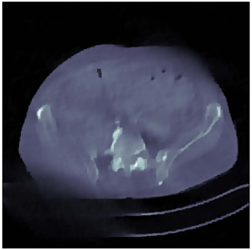
Not Regularized



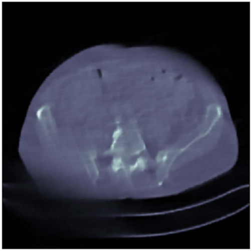
Adversarial Regularizer



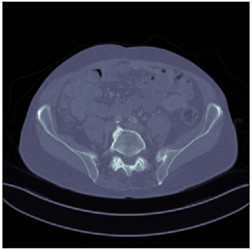
TV Regularized



Convex Adv. Regularizer



True Image



Open Questions

- Convex regularizers often underestimate the high-amplitude components of the true image
- The convexity does not seem to be a significant restriction
- There are 23 citations according to Google:

About 23 results (**0,02** sec)

Learned convex regularizers for inverse problems

☐ Search within citing articles

Learning to optimize: A primer and a benchmark — [PDF] arxiv.org

T Chen, X Chen, W Chen, H Heaton, J Liu... - arXiv preprint arXiv ..., 2021 - arxiv.org

Learning to optimize (L2O) is an emerging approach that leverages machine learning to develop optimization methods, aiming at reducing the laborious iterations of hand ...

☆ Save ⁹⁹ Cite Cited by 32 Related articles All 7 versions »

Bayesian imaging using Plug & Play priors: when Langevin meets Tweedie — [PDF] siam.org

R Laumont, VD Bortoli, A Almansa, J Delon... - SIAM Journal on Imaging ..., 2022 - SIAM

Since the seminal work of Venkatakrishnan, Bouman, and Wohlberg [Proceedings of the Global Conference on Signal and Information Processing, IEEE, 2013, pp. 945--948] in ...

☆ Save ⁹⁹ Cite Cited by 13 Related articles All 8 versions

[HTML] Feasibility-based fixed point networks — [HTML] springeropen.com

H Heaton, S Wu Fung, A Gibali... - Fixed Point ..., 2021 - fixedpointtheoryandapplications ...

Inverse problems consist of recovering a signal from a collection of noisy measurements. These problems can often be cast as feasibility problems; however, additional regularization ...

☆ Save ⁹⁹ Cite Cited by 11 Related articles All 8 versions »

A computationally efficient reconstruction algorithm for circular cone-beam computed tomography using shallow neural networks — [HTML] mdpi.com

MJ Lagerwerf, DM Pelt, WJ Palenstijn, KJ Batenburg - Journal of Imaging, 2020 - mdpi.com

Circular cone-beam (CCB) Computed Tomography (CT) has become an integral part of industrial quality control, materials science and medical imaging. The need to acquire and ...

☆ Save ⁹⁹ Cite Cited by 5 Related articles All 13 versions »

Wasserstein Distance - Justification

- The Kantorovich duality allows to equivalently characterize via

$$\mathsf{Wass}(\mathbb{P}_n, \mathbb{P}_r) := \sup_{f \in 1-\mathsf{Lip}} \varepsilon_{U \sim \mathbb{P}_n} f(U) - \varepsilon_{U \sim \mathbb{P}_r} f(U)$$

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

Assumptions

- Data Manifold Assumption (DMA): The measure $\mathbb{P}_r$ is supported on a weakly compact set $\mathcal{M}$
- Denote by $P_{\mathcal{M}} : D \to \mathcal{M}$, $u \mapsto \arg\min_{v \in \mathcal{M}} \|u - v\|$ 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

Theorem

Assume DMA and low-noise assumption. Then the distance function to the data manifold

$$u \mapsto \min_{v \in \mathcal{M}} \|u - v\|_2$$

is a maximizer to the Wasserstein Loss

$$\sup_{f \in 1-\mathsf{Lip}} \mathbb{E}_{U \sim \mathbb{P}_n} f(U) - \mathbb{E}_{U \sim \mathbb{P}_r} f(U)$$

Approximating f^*

Idea from Wasserstein Generative Adversarial Networks (WGANs)

- Use a neural network (critic) to approximate f^*
- Train the network with the loss

$$\mathbb{E}_{U\sim\mathcal{P}_r}[\Psi_{Theta}(U)]-\mathbb{E}_{U\sim\mathcal{P}_n}[\Psi_{Theta}(U)]+\mu\cdot\mathbb{E}\left[(\|\nabla_u\Psi_\Theta(U)\|_*-1)_+^2\right]$$

- 1-Lipschitz constraint into penalty term (WGAN-GP)