Functional Analysis meets Deep Learning

Sam Power

Cambridge Centre for Analysis
Cantab Capital Institute for the Mathematics of Information

sp825@cam.ac.uk

June 23, 2018

Overview

- **Extremely** non-exhaustive!
- Recap some recent works of a colleague which bring together *Inverse Problems*, *Functional Analysis* and *Deep Learning*.
- Want to focus on *mathematical aspects* of how deep learning can be used

Reading List

1. '*Adversarial Regularizers in Inverse Problems*' (Lunz, Oktem, Schoenlieb)
   - Incorporating deep learning into the inverse problems pipeline.
2. '*Banach Wasserstein GAN*' (Adler, Lunz)
   - Using functional analysis to design new GAN objectives

Inverse Problems

- Given a *forward map* $\mathcal{A} : \mathcal{X} \to \mathcal{Y}$, posit a model

$$y = \mathcal{A}(x) + e \tag{1}$$

  where e is some sort of noise/corruption.
  - Examples: MRI, X-Rays, Image Denoising
  - We will focus on cases in which x is an *image*.
- Often $\mathcal{X}, \mathcal{Y}$ are *vector spaces* and $\mathcal{A}$ is a *linear map*.
- **Inverse Problem**: If we observe y, can we recover x?

Variational Regularisation

- Common approach: **variational regularisation** [VR]

$$\hat{x} \in \arg\min_x \left\{ \frac{1}{2}\|y - \mathcal{A}x\|^2 + \lambda \mathcal{R}(x) \right\} \tag{2}$$

- First term: '*Data Fidelity*' term
  - Encourages x to be consistent with the observation y
- Second term: '*Regularizer*' term
  - Encourages x to have *structure*
    - e.g. sparsity, smoothness, piecewise constant
  - Often makes minimisation problem *well-posed*.
  - λ allows us to tune how much structure to impose

Adversarial Regularizers in Inverse Problems 2018 - Lunz, Oktem, Schoenlieb

- Classically, variational methods are *model-based*: we know the forward operator, and design a regularizer.
- $\mathcal{R}$ can be given *heuristically*, or using *domain knowledge*
  - Sparsity $\rightarrow \mathcal{R}(x) = \|x\|_1$
  - Piecewise-Constant $\rightarrow \mathcal{R}(x) = \|x\|_{TV}$
- In *data-rich* settings, we might try to *learn* a regulariser *from the data*.
- Can then **combine** our knowledge of the forward model with a *data-driven* regularizer $\rightarrow$ Best of both worlds!

How *should* one learn a regularizer?

- Bi-level Regularization: Learn $\mathcal{R}$ so that
  - Solving [VR] recovers the correct solution with high accuracy.
  - $\rightarrow$ Supervised Learning
  - But $\cdots$ issues with scalability
- Adversarial Regularization: Learn $\mathcal{R}$ so that
  - For real input images x, $\mathcal{R}(x)$ is *small*.
  - For noisy reconstructions x, $\mathcal{R}(x)$ is **large**.
  - $\rightarrow$ Unsupervised Learning
  - and, *scalable training algorithm*!

Learning a regularizer

- Learn a regularizer $\mathcal{R}$ to minimise

$$\mathbf{E}_r \mathcal{R}(x) - \mathbf{E}_n \mathcal{R}(x) + \lambda \cdot \mathbf{E}\left[(\|\nabla \mathcal{R}(x)\| - 1)_+^2\right] \tag{3}$$

  where
  - $\mathbf{P}_r$ is the distribution of real images
  - $\mathbf{P}_n$ is the distribution of noisy reconstructions.
  - $\lambda > 0$ is a penalty parameter
- This encourages $\mathcal{R}$ to be
  - *large* on noisy reconstructions,
  - *small* on natural images, and
  - *approximately 1-Lipschitz*.
    - Compare Kantorovich formulation of Wasserstein distance.

Using the regularizer

- The regularizer is given by a neural network with parameters Θ

$$\mathcal{R}(x) = \Psi(x; \Theta) \tag{4}$$

- For a fully-trained $\mathcal{R}$, gradient flow with respect to $\mathcal{R}$ pushes noisy reconstructions $x \sim \mathbf{P}_n$ *closer* to $\mathbf{P}_r$, the distribution of real images.

$$\left.\frac{d}{dt}\right|_{t=0} \text{Wass}\left[(\phi^t_{\mathcal{R}})_{\#}\mathbf{P}_n, \mathbf{P}_r\right] < 0 \tag{5}$$

Solution Properties

- A minimiser of the functional

$$\mathcal{R} \mapsto \mathbf{E}_r \mathcal{R}(x) - \mathbf{E}_n \mathcal{R}(x) \tag{6}$$

  is given by $\mathcal{R}(x) = d(x, \mathcal{M})$, the distance to the 'data manifold' $\mathcal{M}$.
- If our trained $\mathcal{R}$ is truly 1-Lipschitz, then solutions to the Variational Regularization problem obey a *weak stability* property.
  - Informally: If $y_n \to y$, and x_n solves [VR](y_n), then x_n has a subsequence which converges to the solution of [VR](y).

Conclusions

- New framework:
  - Allows knowledge of forward model to be put to use
  - Inherits theory from variational formulation
  - Uses data to design effective regularizers
  - Combines aspects of *model-based* and *data-driven* approaches to build practical models.

Banach Wasserstein GAN
2018 - Adler, Lunz

- Observation: Most GAN objectives assume an underlying ℓ^2 geometry
- Observation: there are **many** norms we could use!
- $\rightarrow$ How would one train a GAN when the geometry is different?

Wasserstein GAN & Lipschitz Geometry

- In WGAN, one learns a discriminator which is Lipschitz
  - $\cdots$ with respect to the ℓ^2 norm
  - This can be enforced by penalising $(\|\nabla D(x)\|_2 - 1)_+$
- If we want to be Lipschitz with respect to a norm $\|\cdot\|$,
  - then our gradients should be bounded in the *dual norm*, $\|\cdot\|_*$.
  - This leads to the **BW-GAN** penalty

$$L = \frac{1}{\gamma}\left[\mathbf{E}_{gen} D(X) - \mathbf{E}_{real} D(X)\right] + \lambda \mathbf{E}\left[\left(\frac{1}{\gamma}\|\nabla D(X)\|_* - 1\right)_+^2\right] \quad (7)$$

Exploring other Banach Spaces

- There is a rich variety of *Banach spaces* which encode different features of functions
- The *Sobolev spaces* $W^{s,p}$ are one such family.
  - For $s = 0$, we recover the familiar L^p spaces.
  - As s increases, we see smoother functions.
    - (informally, s-times differentiable functions)
    - This places greater importance on *local features* of the image.
  - For *negative* s, we are in spaces of 'negative smoothness'.
    - These focus on low-frequency Fourier modes, and thus prioritize the *global structure* of the image.
- As s and p are varied, different aspects of images are emphasised.

Conclusions

- In WGAN and variants, ℓ^2 geometry is implicit
  - $\cdots$ but not necessarily appropriate
- It is possible (and practical) to train GANs using alternative norms.
- Other norms are capable of leading to higher-quality samples.
- Using families of Banach spaces with 'free parameters' allows for some fine-tuning of which image features are emphasised.