Finnish Centre of Excellence in Inverse Modelling and Imaging 2018-2025

UNIVERSITY OF OULU

UCL

LEARNED MODEL-BASED RECONSTRUCTIONS FOR INVERSE PROBLEMS:
ROBUSTNESS AND CONVERGENCE GUARANTEES

Andreas Hauptmann

University of Oulu

Research Unit of Mathematical Sciences

&

University College London

Department of Computer Science

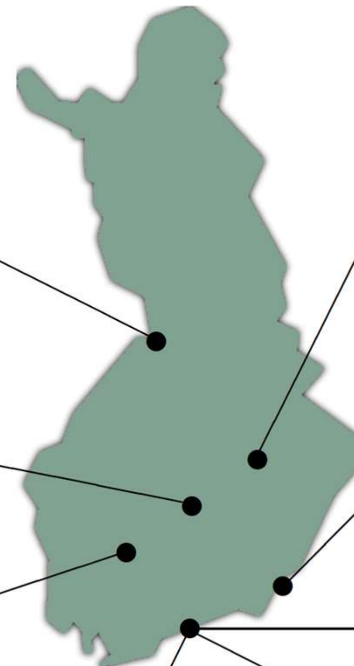
Mathematics of Data Science seminar

DTU, Lyngby, Denmark

23 November 2023

Finnish Centre of Excellence in
Inverse Modelling and Imaging
2018-2025

CENTRES OF EXCELLENCE
IN RESEARCH

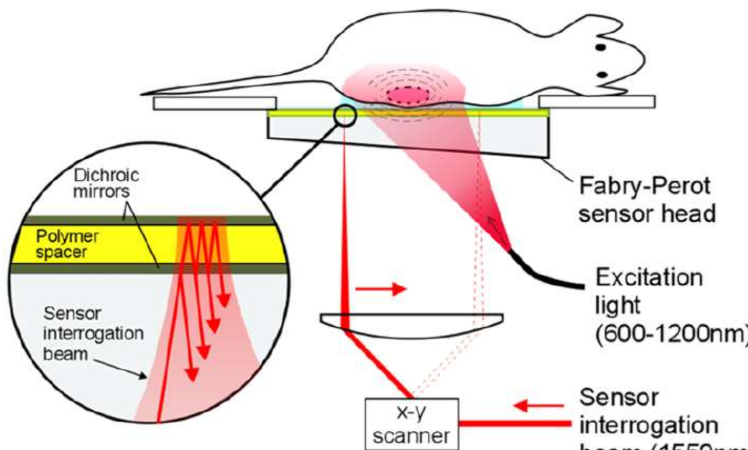


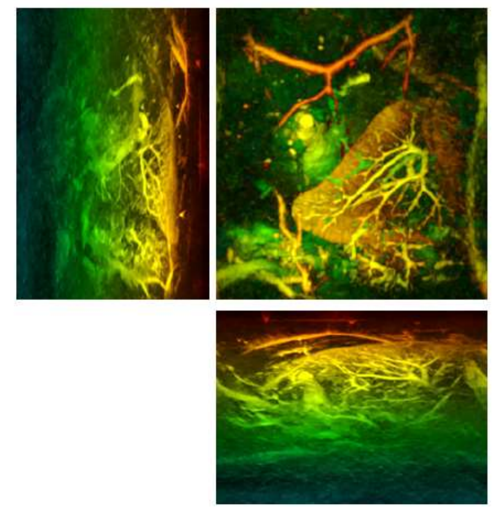
Flagship of Advanced Mathematics
for Sensing, Imaging and Modelling (FAME)
2024-2031

LIMITED-VIEW PHOTOACOUSTIC TOMOGRAPHY

- Linear inverse problem $Ax = y$:
  Recover initial pressure x from measured acoustic signal y

- Planar ultrasound sensor:
  - Limited-view setting
  - (Potential) Sparse-sampling for speed-up





[Jathoul et al., *Nature Photonics*, 2015]

- 3D imaging is expensive:
  - Image (volume) size
  - Data size: high temporal sampling (5x)
  - Forward operator: Wave equation ~12 sec.

THE VARIATIONAL APPROACH

Classic variational approach: find x from measurement y as a minimiser of

$$x \in \arg\min_{x'} \left\{ J(x') \right\} = \arg\min_{x'} \left\{ \mathcal{D}(x'; y) + \lambda \mathcal{R}(x') \right\}.$$

$$\mathcal{D}(x; y) = \frac{1}{2} \|\mathcal{A}x - y\|_2^2$$

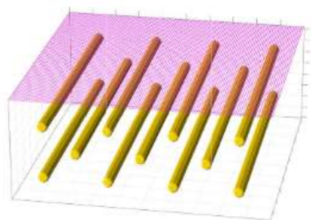
and

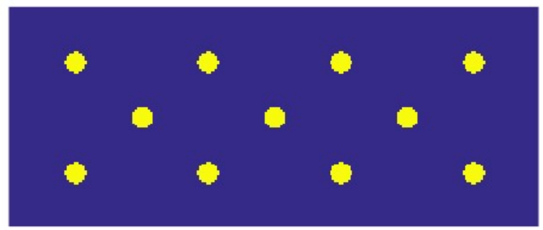
$$\nabla \mathcal{D}(x; y) := \mathcal{A}^*(\mathcal{A}x - y)$$

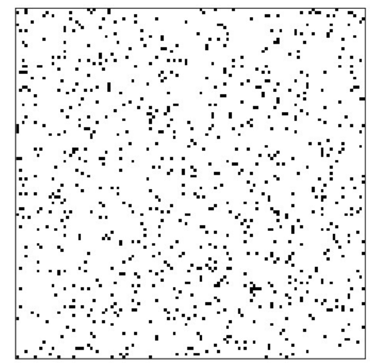
A classic gradient descent scheme would be given by

$$x_{i+1} = x_i - \gamma_{k+1} \left(\mathcal{A}^*(\mathcal{A}x_i - y) + \lambda \nabla \mathcal{R}(x_i) \right)$$

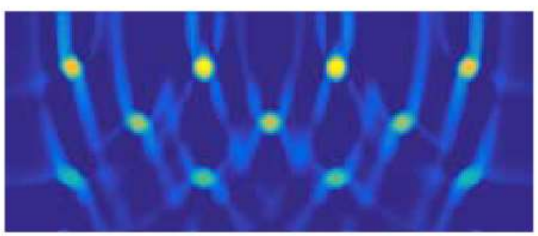
LIMITED-VIEW: ITERATIVE RECONSTRUCTIONS, REGULARISATION AND SUB-SAMPLED DATA



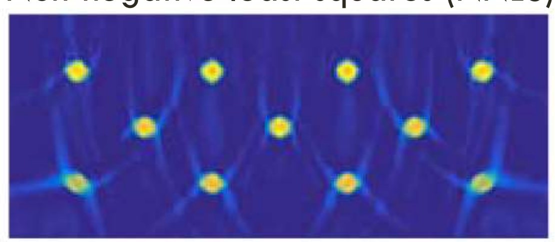


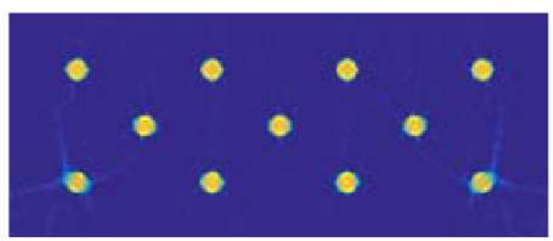


Time reversal

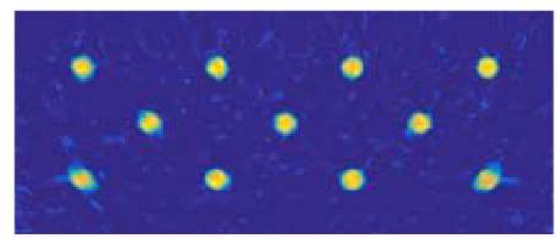


Iterative reconstruction:
Non-negative least squares (NNLS)

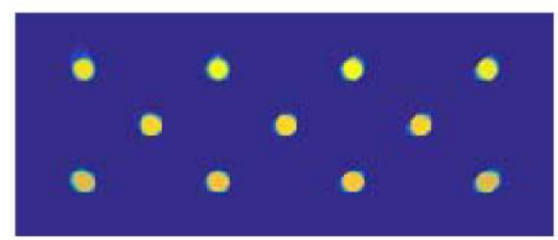




NNLS



TV



THE FORMAL NOTION OF A "CONVERGENT REGULARISATION"

Given the general variational formulation

$$x^* = \arg\min_x D(x, y) + \alpha R(x) \ ,$$

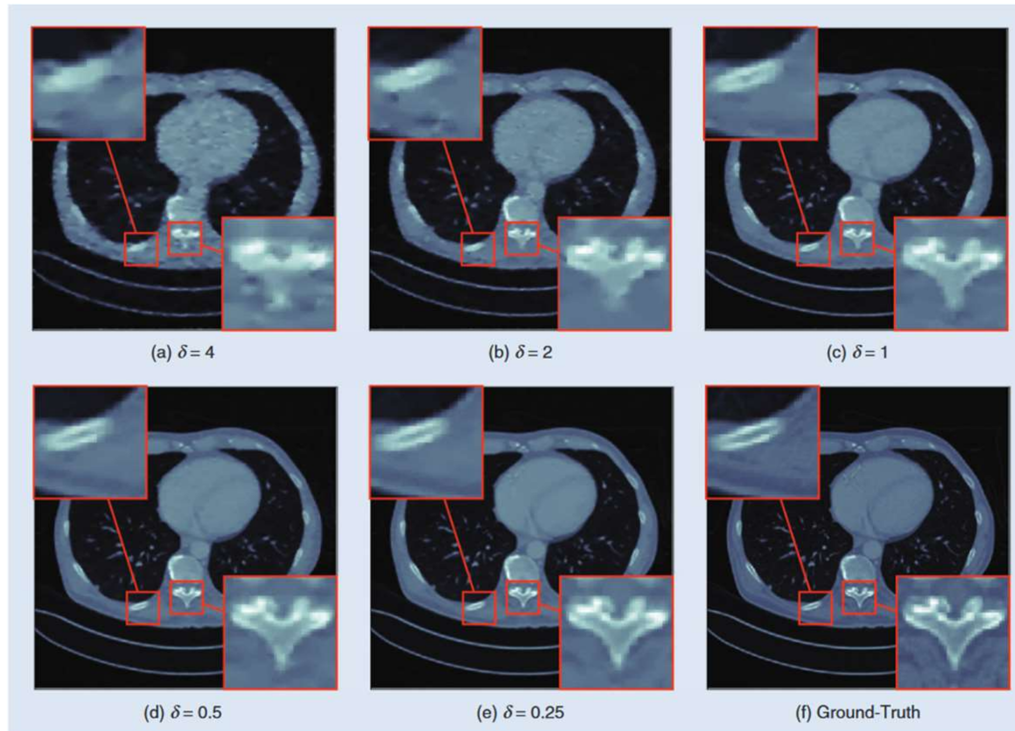
We say that a regularisation is convergent when:

- The solutions x^* are well-defined and depend continuously on the regularisation parameter α and noise level δ.

- When noise vanishes, i.e., $\delta \to 0$ anc then solutions converge to the so-call R-minimising solution:

  $\hat{x} \in \arg\min_x R(x)$ subject to $y^0 =$

  with y^0 the noise free data



[Scherzer, Grasmair, Grossauer, Haltmeier, Variational Methods in Imaging, 2009]

BENEFITS AND LIMITATIONS

Positive:

- We can quantify and analyse solutions
- The reconstruction operator is well-defined as the solution of a variational problem
  → No (training) data dependency
- We know that obtained solutions are "data-consistent" and converge (continuously) to solutions of the measurement equation $Ax = y^0$, if noise vanishes.

Negative:

- Slow convergence: can take 100 - 1000 of iterations.
- Limited expressivity: Reconstruction quality depends on prior information encoded in the regulariser → Balance representation of data and desirable analytical conditions.
- Unfortunately, computing good solutions is not as straight-forward as it may seem:
  → Choice of regulariser
  → Choice of regularisation parameter

THE DATA-DRIVEN APPROACH

- Previous limitations can be overcome by data-driven approaches:

  → Simply speaking, instead of hand-crafting a regularisation and prior, we can learn the prior information from the data itself

  → More efficient reconstruction operators or optimisation schemes can be learned to compute solutions

BUT: We may lose some (or even all) of the theoretical conditions we required before. (Depending on the approach taken as we see shortly)

LEARNED ITERATIVE RECONSTRUCTIONS

Classic variational approach: find x from measurement y as a minimiser of

$$x \in \arg\min_{x'} \left\{ J(x') \right\} = \arg\min_{x'} \left\{ \mathcal{D}(x';y) + \lambda \mathcal{R}(x') \right\}.$$

$$\mathcal{D}(x;y) = \frac{1}{2}\|\mathcal{A}x - y\|_2^2$$

and

$$\nabla \mathcal{D}(x;y) := \mathcal{A}^*(\mathcal{A}x - y)$$

A classic gradient descent scheme would be given by

$$x_{i+1} = x_i - \gamma_{k+1}\left(\mathcal{A}^*(\mathcal{A}x_i - y) + \lambda \nabla \mathcal{R}(x_i)\right)$$

Pro:

- Interpretable
- Convergence & reconstruction guarantees

Contra:

- Slow to converge
- Difficult to choose regulariser and parameter

LEARNED ITERATIVE RECONSTRUCTIONS

Classic variational approach: find x from measurement y as a minimiser of

$$x \in \arg\min_{x'} \left\{ J(x') \right\} = \arg\min_{x'} \left\{ \mathcal{D}(x'; y) + \lambda \mathcal{R}(x') \right\}.$$

$$\mathcal{D}(x; y) = \frac{1}{2} \|\mathcal{A}x - y\|_2^2$$

and

$$\nabla \mathcal{D}(x; y) := \mathcal{A}^*(\mathcal{A}x - y)$$

A simple learned gradient-like scheme would be given by

$$x_{i+1} = \mathcal{G}_{\theta_i}\big(x_i, \mathcal{A}^*(\mathcal{A}\, x_i - y)\big), \; i = 0, \dots, N-1.$$

This defines a reconstruction operator when stopped after N iterates:

$$\mathcal{A}_\theta^\dagger(y) := x_N \quad \text{where } \theta = (\theta_0, \dots, \theta_{N-1})$$

and initialisation $x_0 = \mathcal{A}^\dagger{}_\theta(y)$.

[Adler & Öktem, 2018], [Putzky & Welling, 2017]

TRAINING PROCEDURE

Given supervised training data $(x^{(j)}, y^{(j)}) \in X \times Y$.

Then an optimal parameter is found by

$$\min_{\theta} \frac{1}{m} \sum_{j=1}^{m} \mathrm{L}_\theta(x^{(j)}, y^{(j)})$$

where the loss function is given as

$$\mathrm{L}_\theta(x, y) := \left\| \mathcal{A}^\dagger_\theta(y) - x \right\|_X^2 \quad \text{for } (x, y) \in X \times Y.$$

Greedy training: Require iterate-wise optimality.

Given only a loss function for the i:th unrolled iterate:

$$\mathrm{L}_{\theta_i}(x_i, y) = \left\| \mathcal{G}_{\theta_i}\big(x_i, \mathcal{A}^*(\mathcal{A}(x_i) - y)\big) - x \right\|_X^2$$

where $x_i := \mathcal{G}_{\theta_{i-1}}\big(x_{i-1}, \mathcal{A}^*(\mathcal{A}(x_{i-1}) - y)\big)$.

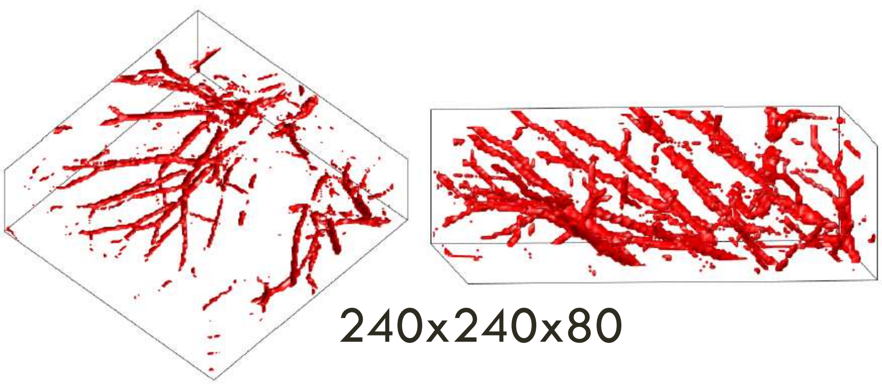
This constitutes an upper bound to end-to-end networks.

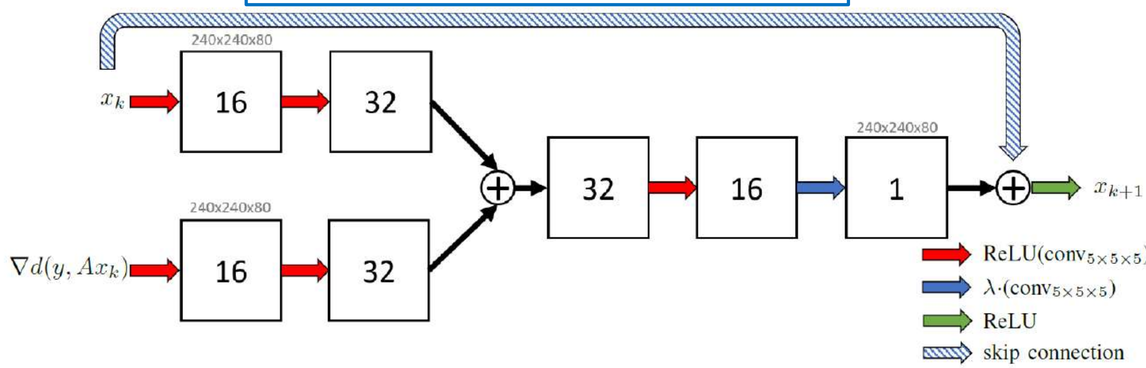
- End-to-end training is not (readily) scalable depending on:
  - Memory limitations
  - Operator evaluation: Repeated application of forward/adjoint operator
    - 3D PAT → 1 (unrolled) iteration takes ~25sec. (forward + adjoint)

NETWORK AND TRAINING

- With the computation of the gradient, total training time for 5 iterations takes 7 days
- Compare: End-to-end training would take about ∼140 days

$$x_{i+1} = \mathcal{G}_{\theta_i}\left(x_i, \mathcal{A}^*(\mathcal{A}x_i - y)\right)$$

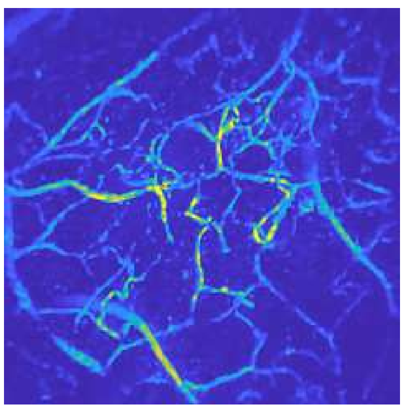




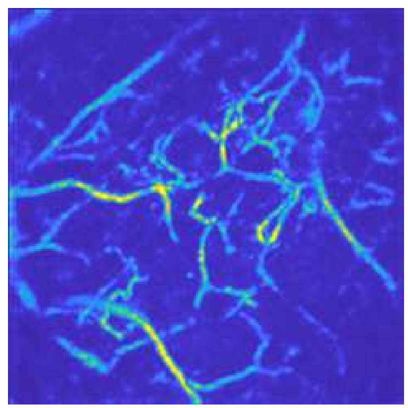
APPLICATION TO HUMAN IN-VIVO MEASUREMENTS

- Reduces reconstruction time by a factor 4 (by reduction of iterations)
- Considerably improves reconstruction quality

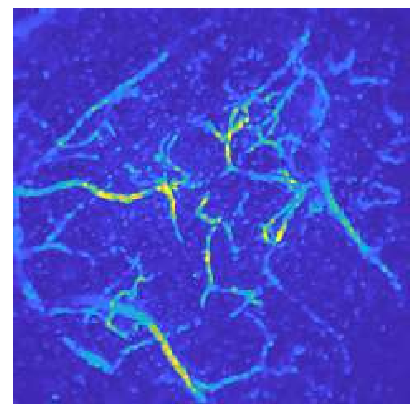
Reference
Fully-sampled data



Learned Reconstruction
4x sub-sampled, 5 Iterations,
Time: 2.5 min., PSNR: 41.40



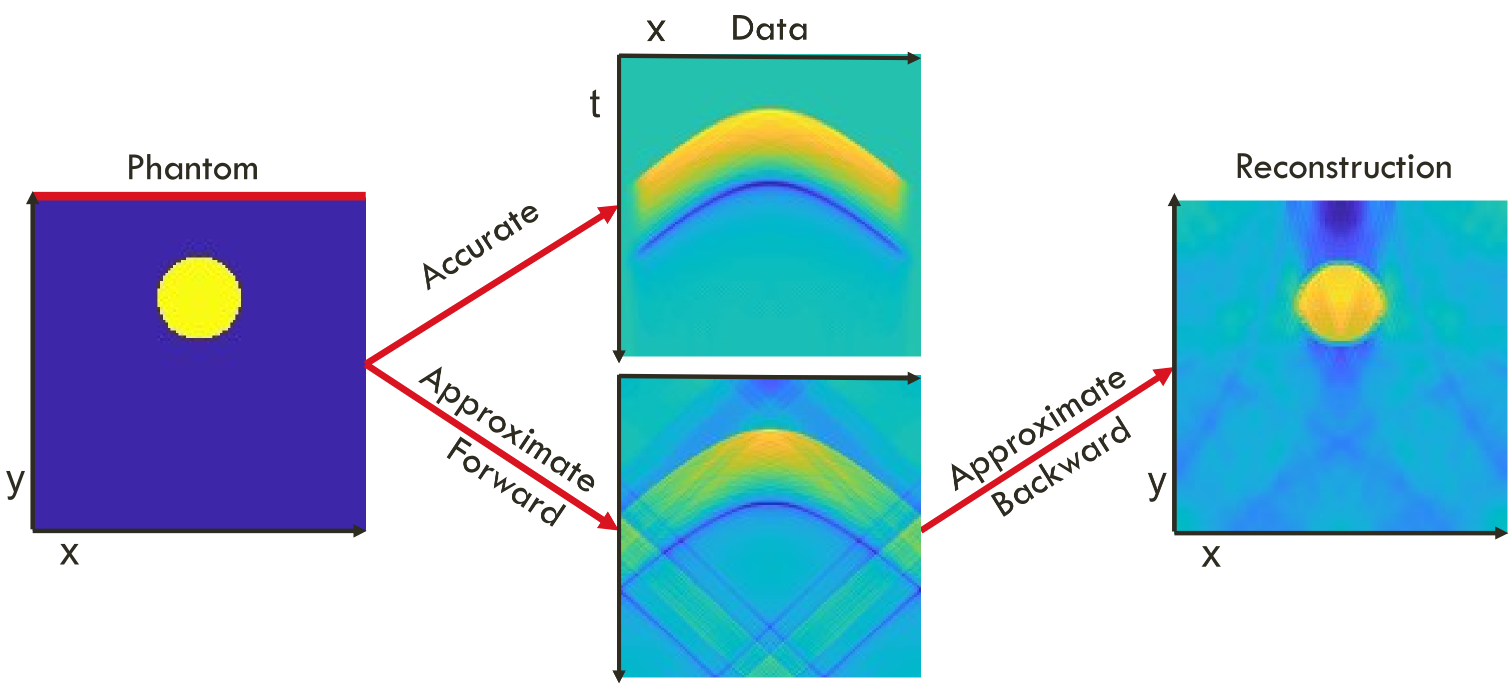
Total Variation Reconstruction
4x sub-sampled, 20 Iterations,
Time: 10 min., PSNR: 38.05



[Hauptmann et al., *IEEE Transactions on Medical Imaging*, 2018]

UTILISING A REDUCED MODEL

- Bottleneck of iterative reconstruction time is the application of the forward model
  - Use a fast approximate model in the iterative reconstruction instead (8x faster)
  - But approximate model introduces additional artefacts



UTILISING A REDUCED MODEL: IMPLICIT CORRECTION

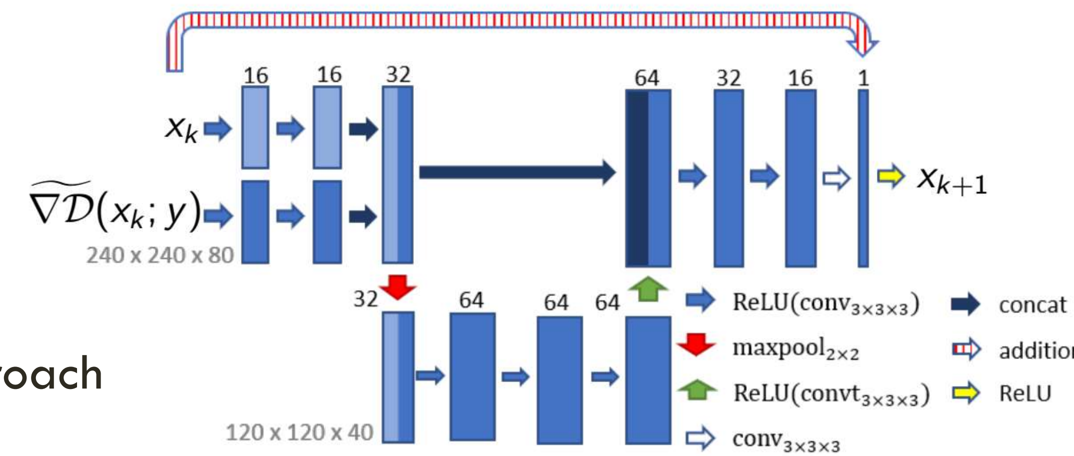
We formulate the updates now using an approximate gradient

$$x_{k+1} = \mathcal{G}_{\theta_k}(\widetilde{\nabla\mathcal{D}}(x_k; y), x_k)$$

with

$$\widetilde{\nabla\mathcal{D}}(x_k; y) := \widetilde{A}^*(\widetilde{A}x_k - y).$$

- Trained supervised on reference reconstruction from fully sampled data
- *5 iterates are trained in a greedy approach*



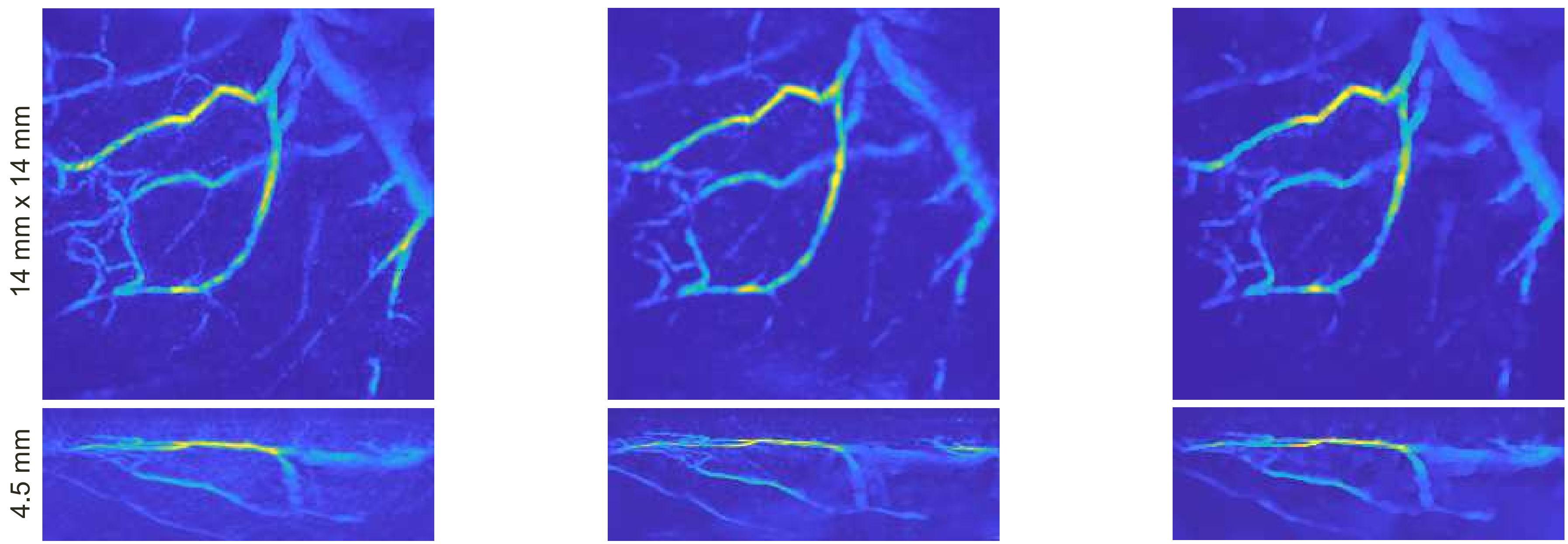
ACCELERATION BY USING AN APPROXIMATE MODEL

- Reduces reconstruction time by another factor of ~8 (→ 32x compared to TV)

Reference
Fully-sampled data

Learned Reconstruction
4x sub-sampled, 5 Iterations,
Time: 20 sec., PSNR: 42.18

Total Variation Reconstruction
4x sub-sampled, 20 Iterations,
Time: 10 min., PSNR: 41.16



[Hauptmann et al., *Machine Learning for Medical Image Reconstruction*, 2018]

UNIVERSITY OF OULU UCL

RECAP, WHY WE NEED LEARNING:

- Image quality depends on multiple factors, such as:
  - Acquisition time
  - Signal strength (radiation exposure)
  - Patient movements
  - Cost-point
- Advanced mathematical techniques used to compensate, but:
  - Can be slow → Not applicable for real-time
  - Analytic prior → Do not describe targets well
  - Accurate models → Computationally expensive

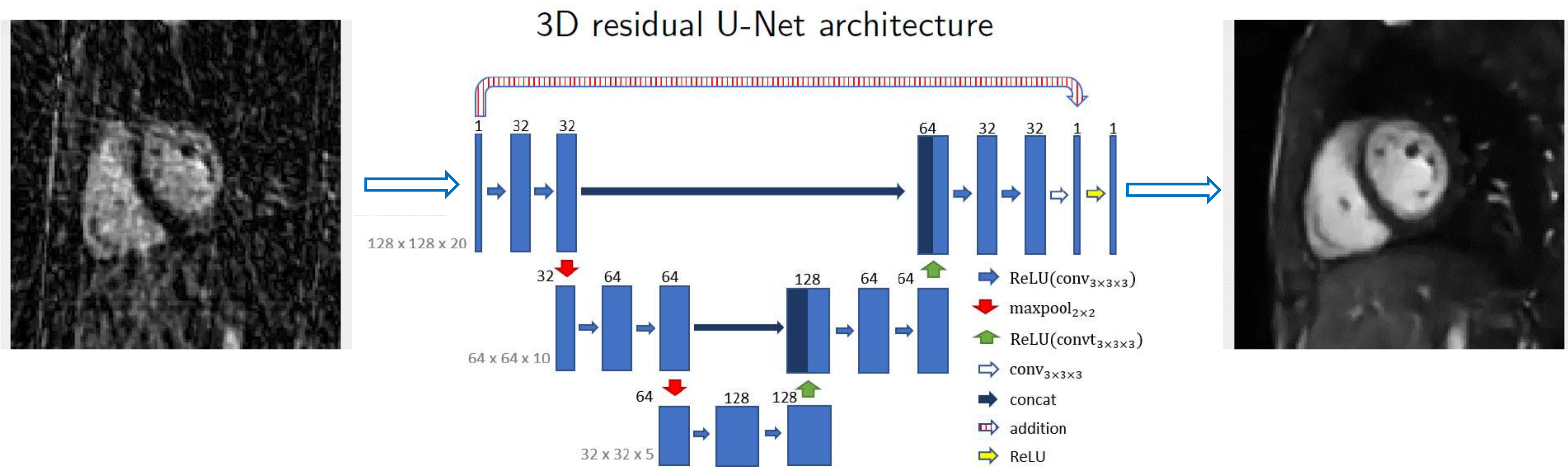
THE DATA-DRIVEN APPROACH: TWO-STEP

We now want to learn a parameterised reconstruction operator, such that

$$R_\theta(y) \approx x.$$

Two-step approach: 1.) Compute a reconstruction (undersampled, zero-filled k-space data)

2.) Train a network Λ_θ as post-processing to remove artefacts and noise



[Hauptmann, et al., Magnetic Resonance in Medicine, 2019]

THE DATA-DRIVEN APPROACH: ITERATIVE

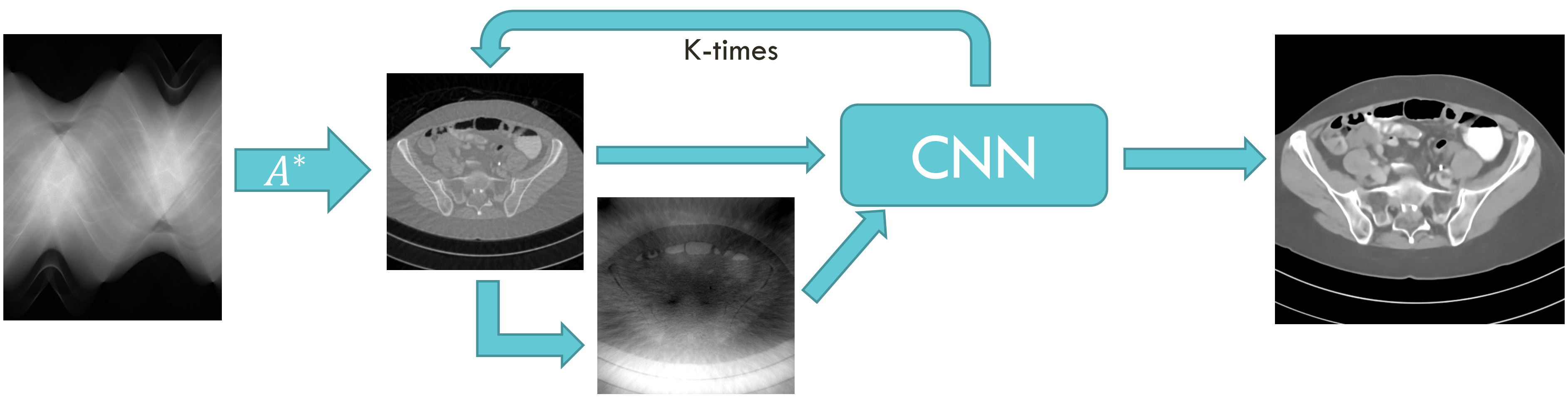
More powerful and successful methods rather compute reconstructions iteratively, where an updating operator Λ_{θ_k} is learned. For instance, in general form as:

$$x^{k+1} = \Lambda_{\theta_k}\left(x^{k+1}, \nabla D(Ax, y)\right).$$

Note, for
$D(Ax, y) = ||Ax - y||_2^2$
We get
$\nabla||Ax - y||_2^2 = 2A^*(Ax - y).$

These include many popular approaches such as:

- Variational Networks [Hammernik et al., *Magnetic resonance in medicine*, 2018]
- Learned Gradient Schemes [Adler & Öktem, *Inverse Problems*, 2017]
- Plug-and-Play type approaches [Venkatakrishnan, Bouman, Wohlberg, *GlobalSIP*, 2013]



EMPIRICAL SUCCESS WITHOUT THEORETICAL GUARANTEES?

- Most successful methods come without theoretical guarantees

- Convergence proofs can be established by restricting the networks:
  - Contractiveness/non-expansive
  - Convexity
  - Invertibility

- **Disclaimer:** Limiting expressivity

  → Worse quantitative performance

WHAT CAN WE SAY THEORETICALLY?

- How stable are learned reconstruction methods?
- Do learned unrolled/iterative approaches converge?
- Do we minimise the variational cost function, or a related one?
- Is the learned reconstruction a (formal) regularisation, i.e., can we say something about the case of vanishing noise?

PHYSICS-DRIVEN MACHINE LEARNING FOR COMPUTATIONAL IMAGING

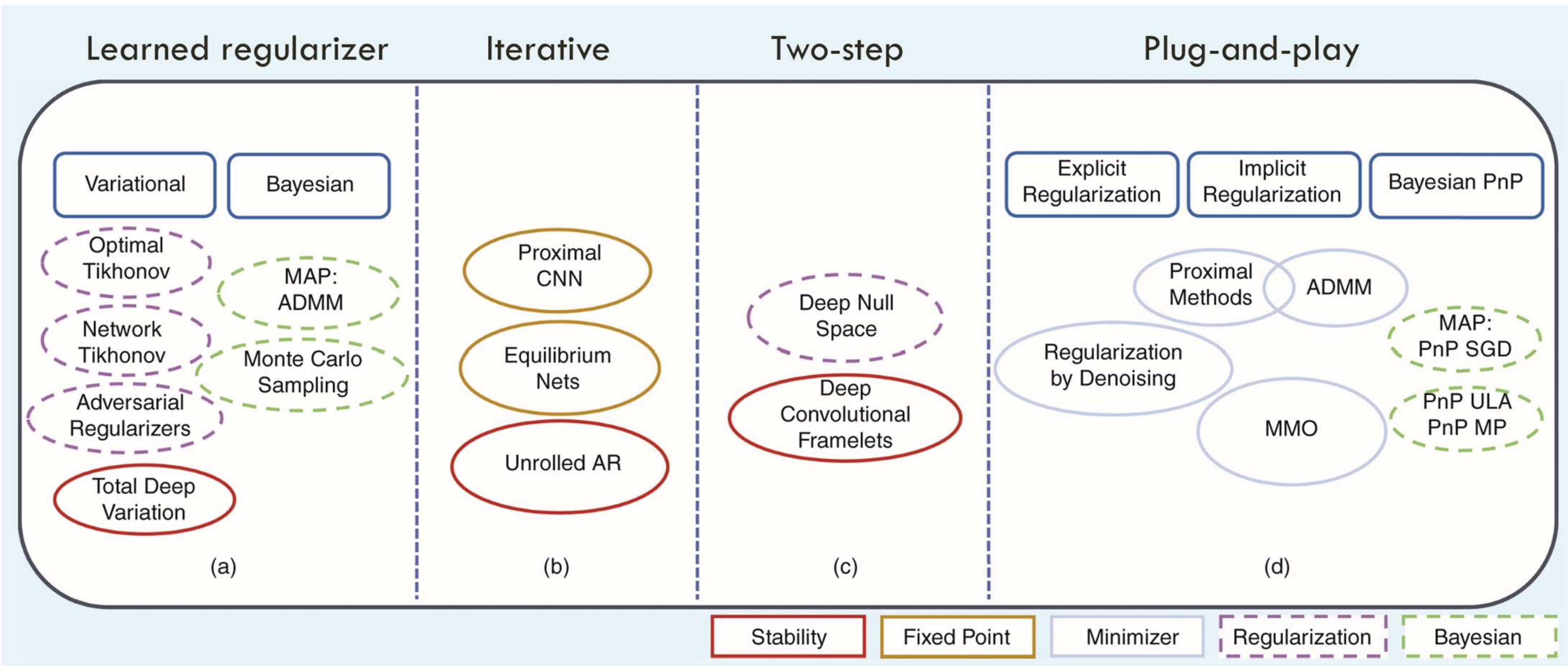
Learned Reconstruction Methods With Convergence Guarantees

A survey of concepts and applications

Subhadip Mukherjee, Andreas Hauptmann, Ozan Öktem, Marcelo Pereyra, and Carola-Bibiane Schönlieb

IEEE SIGNAL PROCESSING MAGAZINE | January 2023 |

OVERVIEW OF EXISTING APPROACHES



Stability Versus Accuracy

Consider a trained reconstruction operator $\mathcal{R}_\theta$ with fixed network parameters (learned from training data). The reconstruction produced by $\mathcal{R}_\theta$ is said to be stable if $\mathcal{R}_\theta : \mathbb{Y} \to \mathbb{X}$ is a continuous function of the observed data. Formally, stability demands that

$$\| \mathcal{R}_\theta(y+w) - \mathcal{R}_\theta(y) \|_{\mathbb{X}} \to 0 \text{ as } \| w \|_{\mathbb{Y}} \to 0.$$

One possibility for a stability analysis is to consider the Lipschitz constant L of the mapping $\mathcal{R}_\theta$, which is given by the smallest $L > 0$, such that

$$\| \mathcal{R}_\theta(y_1) - \mathcal{R}_\theta(y_2) \| \leq L \| y_1 - y_2 \|, \text{for all } y_1, y_2 \in \mathbb{Y}. \quad \text{(S1)}$$

1. Note, since deep neural networks are compositions of affine functions and smoothly varying nonlinear activation functions, a reconstruction operator R_θ is continuous and a constant L exists.
   → That makes the mapping formally stable, but L might be large

Stability Versus Accuracy

Consider a trained reconstruction operator $\mathcal{R}_\theta$ with fixed network parameters (learned from training data). The reconstruction produced by $\mathcal{R}_\theta$ is said to be stable if $\mathcal{R}_\theta : \mathbb{Y} \to \mathbb{X}$ is a continuous function of the observed data. Formally, stability demands that

$$\| \mathcal{R}_\theta(y+w) - \mathcal{R}_\theta(y) \|_{\mathbb{X}} \to 0 \text{ as } \| w \|_{\mathbb{Y}} \to 0.$$

One possibility for a stability analysis is to consider the Lipschitz constant L of the mapping $\mathcal{R}_\theta$, which is given by the smallest $L > 0$, such that

$$\| \mathcal{R}_\theta(y_1) - \mathcal{R}_\theta(y_2) \| \leq L \| y_1 - y_2 \|, \text{ for all } y_1, y_2 \in \mathbb{Y}. \quad \text{(S1)}$$

Additionally, a consequence of (S1) is that the reconstruction of a slightly perturbed image must satisfy

$$\| \mathcal{R}_\theta(\mathcal{A}(x+\eta)) - \mathcal{R}_\theta(\mathcal{A}x) \| \leq L \| \mathcal{A}\eta \|, \text{ for any perturbation } \eta.$$

2. The perturbation $||A\eta||$ could be arbitrarily small for small η.
   →If L is small the reconstruction operator is insensitive to these perturbations
   →An accurate R_θ must have a large Lipschitz constant L

Adversarial Robustness

The adversarial robustness of a trained reconstruction operator $\mathcal{R}_\theta$ is measured by the largest deviation caused in the reconstruction by a small perturbation in the data. For a given $y_0 = \mathcal{A}x_0 \in \mathbb{Y}$, where x_0 is the underlying image, and a given noise level ϵ_0, this is defined formally as [S1]

$$\delta_{\text{adv}} = \sup_{w: \|w\| \leq \epsilon_0} \| \mathcal{R}_\theta(y_0 + w) - \mathcal{R}_\theta(y_0) \|_2. \quad \text{(S2)}$$

If δ_{adv} is small for small ϵ_0, the reconstruction method $\mathcal{R}_\theta$ is said to be adversarially robust.

References

[S1] M. Genzel, J. Macdonald, and M. Marz, "Solving inverse problems with deep neural networks - robustness included," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Feb. 4, 2022, doi: 10.1109/TPAMI.2022.3148324.

[S2] V. Antun et al., "On instabilities of deep learning in image reconstruction and the potential costs of AI," *Proc. Nat. Acad. Sci.*, vol. 117, no. 48, pp. 30,088–30,095, 2020, doi: 10.1073/pnas.1907377117.

[S3] R. Alaifari, G. S. Alberti, and T. Gauksson, "Localized adversarial artifacts for compressed sensing MRI," 2022, *arXiv:2206.05289v1*.

- Concern about the adversarial stability (or lack thereof) of deep learning-based approaches has been raised [S2].

- Subsequent work [S1] performed a systematic comparison of data-driven methods with the classical (TV)-regularized solution.
  → Learned methods were found to be as robust as TV to adversarial noise.
  → For the FastMRI dataset learned methods were more resilient to large perturbations.

- Finally, [S3] showed that learned methods are more robust with respect to ℓ_∞-perturbations. (Capturing localised artifacts)

FIXED POINT AND OBJECTIVE CONVERGENCE

Fixed point convergence can be (comparably) easily achieved when considering a proximal gradient type update:

$$x^{k+1} = R(x^k) = \Lambda_\theta\left(x^k - \lambda_k \nabla D(Ax^k, y)\right).$$

When Λ_θ is trained to be 1-Lipschitz, i.e., with constant $\mathrm{L} < 1$ and $\lambda_k < ||A||_{op}^2$, then the above iterations are contractive and will converge to a fixed point

$$x^\infty = R(x^\infty).$$

This tells us that the iterations are stable,
BUT: This does not say anything about the "goodness" of x^∞.
→ Objective convergence is more desirable, but also more restrictive, in short:

We need to parameterise the network Λ_θ in such a way that it corresponds to the gradient of a (possibly convex) function (representing the regulariser).

[Gilton, Ongie, Willett, *IEEE Transactions on Computational Imaging*, 2021]

Objective Convergence of Plug-and-Play With Gradient Step Denoisers

The convergence of plug-and-play (PnP) denoisers used with half-quadratic splitting was established in [26]. The denoiser is constructed as a gradient step denoiser, as explained in the "PnP Denoising" section; i.e., $D_\sigma = \mathrm{Id} - \nabla g_\sigma$, where g_σ is proper, lower semicontinuous, and differentiable with an L-Lipschitz gradient. The PnP algorithm proposed in [26] takes the form $x_{k+1} = \mathrm{prox}_{\tau f}(x_k - \tau\lambda\nabla g_\sigma(x_k))$, where $f: \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ measures the data fidelity and is assumed to be convex and lower semicontinuous. Under these assumptions on f and g_σ, the following guarantees hold for $\tau < 1/\lambda L$:

1) The sequence $F(x_k)$, where $F = f + \lambda g_\sigma$, is nonincreasing and convergent.
2) Here, $\|x_{k+1} - x_k\|_2 \to 0$, which indicates that iterations are stable in the sense that they do not diverge if one iterates indefinitely.
3) All limit points of $\{x_k\}$ are stationary points of $F(x)$.

[Hurault, Leclaire, Papadakis, *arXiv:2110.03220*, 2021]

CONVERGENT REGULARISATION

In fact, we can even obtain a convergent regularisation strategy with a learned regulariser:

Learn just the regulariser such that

$$D(Ax, y) + \alpha R_\theta(x),$$

we can then enforce conditions to ensure well-posedness of the solution operator.

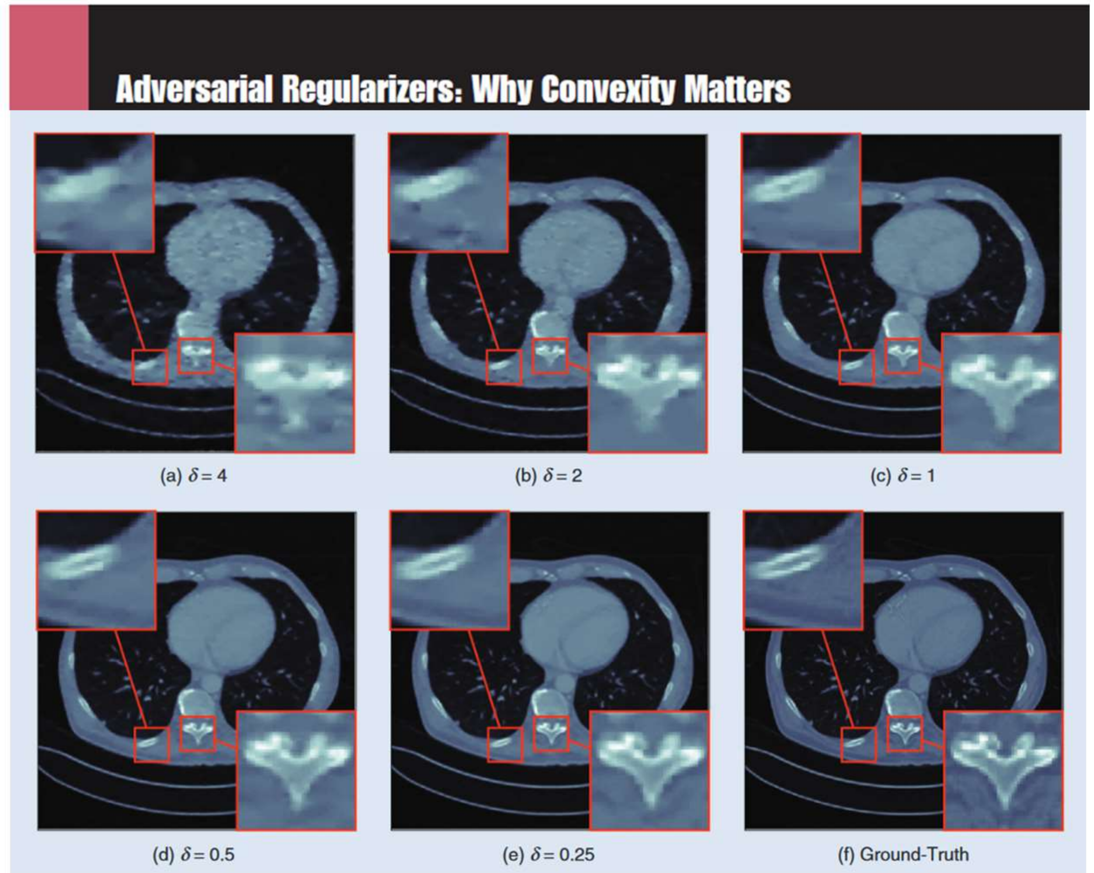
→ Simply put: when R_θ is convex we obtain a classical convergent regularisation

→ Composition with a regularisation functional g:

$$D(Ax, y) + \alpha g(R_\theta(x))$$

→ Plug-and-play with linear denoiser: Quadratic R

BUT: We need to solve again the variational problem, which is slow.



[Lunz, Öktem, Schönlieb, *NeurIPS*, 2018] [Li, Schwab, Antholzer, Haltmeier, Inverse Problems, 2020] [Hauptmann, Mukherjee, Schönlieb, Sherry, arXiv]

CONCLUSIONS

- Inverse problems and regularisation theory helps to understand the problem
- Provide convergence, stability, and data-consistent reconstructions
- Classical methods are reliable but have shortcomings:
  computation times, expressivity, hand-tuning
- Data-driven approaches can solve shortcomings, but guarantees may be lost
  → We can reintroduce varying levels of guarantees
  → The more theoretical guarantees we get, the more conditions are enforced

More restrictive conditions → Worse (quantitative) performance

WHAT'S TO COME?

- Currently: trade-off between performance and theoretical guarantees.

- But how much guarantee is needed, if performance is better?
  → Importance of challenges like FastMRI!
  → Do clinicians/engineers care?

- Untouched here: Generalisation and the role of training data
  → Here reconstruction guarantees can be certainly useful!
  → Need for more semi- or unsupervised methods?

Learned approaches are here to stay!