Lecture 2: Machine learning in the context of inverse problems - Learning priors and post-processing

Ozan Öktem Jonas Adler

Department of Mathematics
KTH - Royal Institute of Technology, Stockholm

April 18, 2018
Mini-course: Mathematics of Deep Learning
with an Emphasis on Inverse Problems

Georg-August-Universität Göttingen

Lecture overview

- Inverse problems and regularisation
  - Ill-posedness & regularisation
  - Analytic methods
  - Iterative methods with early stopping
  - Variational methods
- Learning parameters in regulariser
  - Bi-level optimisation
  - Examples of bi-level optimisation
- Sparse models
  - Sparse recovery
  - Joint reconstruction & dictionary learning
  - Dictionary learning
  - Convolutional Sparse Model
- Prior with black box denoiser

Inverse problems and regularisation

Inverse problem

Inverse problem: Recover (reconstruct) an estimate of signal $f^* \in X$ from data $g \in Y$ assuming

$$g = \mathcal{A}(f^*) + \delta g.$$

- Reconstruction space: X (normed) vector space, elements represent possible signals. We consider $X \subset$ real valued functions defined on $\Omega \subset \mathbb{R}^n$.
- Data space: Y (normed) vector space, elements represent possible data. We consider $Y \subset$ real valued functions defined on manifold $\mathbb{M}$.
  Acquisition geometry: Digitisation of data, i.e., sampling scheme in $\mathbb{M}$.
- Forward operator: $\mathcal{A} \colon X \to Y$, the deterministic part of a simulator (mathematical model) for how data is generated.
- Data noise component: δg generated by a Y-valued random variable.
- Data noise level: $\delta := \|\delta g\|_Y$.
- Reconstruction (operator): Mapping $\mathcal{A}^\dagger \colon Y \to X$ such that $\mathcal{A}^\dagger(g) \approx f^*$.

Ill-posedness

- Inverse problem: Recover $f^* \in X$ from $g = \mathcal{A}(f^*) + \delta g$.
- Ill-posedness
  - Existence

Ill-posedness

- Inverse problem: Recover $f^* \in X$ from $g = \mathcal{A}(f^*) + \delta g$.
- Ill-posedness
  - Existence

- Existence: Relax notion of a solution, e.g.,

$$\min_{f \in X} \mathcal{L}\big(\mathcal{A}(f), g\big)$$

given $\mathcal{L} \colon Y \times Y \to \mathbb{R}$ (data discrepancy).

Ill-posedness

- Inverse problem: Recover $f^* \in X$ from $g = \mathcal{A}(f^*) + \delta g$.
- Ill-posedness
    - Existence
    - Uniqueness
    - Stability

- Existence: Relax notion of a solution, e.g.,

$$\min_{f \in X} \mathcal{L}\big(\mathcal{A}(f), g\big)$$

given $\mathcal{L} \colon Y \times Y \to \mathbb{R}$ (data discrepancy).
- New difficulties:
    - Multiple (possibly infinitely many) solutions.
    - Generic solution not continuous w.r.t. data.

Ill-posedness

- Inverse problem: Recover $f^* \in X$ from $g = \mathcal{A}(f^*) + \delta g$.
- Ill-posedness
    - Existence
    - Uniqueness
    - Stability

- Regularisation: Replace original ill-posed inverse problem by a well-posed one that is convergent as noise level tends to zero.
- Key components:
    - Data model: Forward operator $\mathcal{A}$ and statistical properties of data.
    - Prior model: A priori info. about f^*.
    - Regularisation parameter: Compromise between fitting data and stability, usually based on an estimate of the noise level.

Ill-posedness

- Inverse problem: Recover $f^* \in X$ from $g = \mathcal{A}(f^*) + \delta g$.
- Ill-posedness
  - Existence
  - Uniqueness
  - Stability
- Well-defined regularisation: Existence and stability.

- Regularisation: Replace original ill-posed inverse problem by a well-posed one that is convergent as noise level tends to zero.
- Key components:
  - Data model: Forward operator $\mathcal{A}$ and statistical properties of data.
  - Prior model: A priori info. about f^*.
  - Regularisation parameter: Compromise between fitting data and stability, usually based on an estimate of the noise level.

Ill-posedness

- Inverse problem: Recover $f^* \in X$ from $g = \mathcal{A}(f^*) + \delta g$.
- Ill-posedness
  - Uniqueness
  - Stability
- Well-defined regularisation: Existence and stability.
- Regularisation: Replace original ill-posed inverse problem by a well-posed one that is convergent as noise level tends to zero.
- Key components:
  - Data model: Forward operator $\mathcal{A}$ and statistical properties of data.
  - Prior model: A priori info. about f^*.
  - Regularisation parameter: Compromise between fitting data and stability, usually based on an estimate of the noise level.

Statistical properties of data: Captured by data discrepancy $\mathcal{L}\colon Y \times Y \to \mathbb{R}$, a suitable affine transform of negative log-likelihood of data (Bertero et al., 2008).

- Additive Gaussian noise (zero mean, covariance Σ): $\mathcal{L}(g,h) := \|g-h\|_2^2$ or $\mathcal{L}(g,h) := \|g-h\|_{\Sigma^{-1}}^2$.

Ill-posedness

- Inverse problem: Recover $f^* \in X$ from $g = \mathcal{A}(f^*) + \delta g$.
- Ill-posedness
    - Uniqueness
    - Stability
- Well-defined regularisation: Existence and stability.

- Regularisation: Replace original ill-posed inverse problem by a well-posed one that is convergent as noise level tends to zero.
- Key components:
    - Data model: Forward operator $\mathcal{A}$ and statistical properties of data.
    - Prior model: A priori info. about f^*.
    - Regularisation parameter: Compromise between fitting data and stability, usually based on an estimate of the noise level.

Statistical properties of data: Captured by data discrepancy $\mathcal{L}\colon Y \times Y \to \mathbb{R}$, a suitable affine transform of negative log-likelihood of data (Bertero et al., 2008).

- Poisson data (mean = measurement): $\mathcal{L}(g, h) := \sum_{i=1}^{m} \big[h_i \log g_i - g_i \big]$.

Ill-posedness

- Inverse problem: Recover $f^* \in X$ from $g = \mathcal{A}(f^*) + \delta g$.
- Ill-posedness
  - Existence
  - Uniqueness
  - Stability
- Well-defined regularisation: Existence and stability.
- Regularisation: Replace original ill-posed inverse problem by a well-posed one that is convergent as noise level tends to zero.
- Key components:
  - Data model: Forward operator $\mathcal{A}$ and statistical properties of data.
  - Prior model: A priori info. about f^*.
  - Regularisation parameter: Compromise between fitting data and stability, usually based on an estimate of the noise level.

Statistical properties of data: Captured by data discrepancy $\mathcal{L}\colon Y \times Y \to \mathbb{R}$, a suitable affine transform of negative log-likelihood of data (Bertero et al., 2008).

- Impulse noise, e.g., salt and pepper noise: $\mathcal{L}(g, h) := \|g - h\|_0$ or $\mathcal{L}(g, h) := \|g - h\|_1$.

Ill-posedness

- Inverse problem: Recover $f^* \in X$ from $g = \mathcal{A}(f^*) + \delta g$.
- Ill-posedness
  - Existence
  - Uniqueness
  - Stability
- Well-defined regularisation: Existence and stability.

- Regularisation: Replace original ill-posed inverse problem by a well-posed one that is convergent as noise level tends to zero.
- Key components:
  - Data model: Forward operator $\mathcal{A}$ and statistical properties of data.
  - Prior model: A priori info. about f^*.
  - Regularisation parameter: Compromise between fitting data and stability, usually based on an estimate of the noise level.

Main challenges:

- Choosing prior model $\implies$ reconstruction is a well-defined regularisation.
- Choose regularisation parameter.
- Computationally feasible data model.

- Inverse problem: Recover $f^* \in X$ from $g = \mathcal{A}(f^*) + \delta g$.
- Reconstruction method: Parametrised family $\{\mathcal{A}_\theta^\dagger\}_\theta$ where $\mathcal{A}_\theta^\dagger \colon Y \to X$.
- Existence: For every $g \in Y$, there exist a solution $\mathcal{A}_\theta^\dagger(g) \in X$ given fixed θ.
  $\implies$ Makes it possible to define reconstruction method as mapping.
- Stability: $g \mapsto \mathcal{A}_\theta^\dagger(g)$ is continuous in relevant topology for fixed θ.
  $\implies$ Small variations in data does not result in large variations in reconstruction.

- Inverse problem: Recover $f^* \in X$ from $g = \mathcal{A}(f^*) + \delta g$.
- Reconstruction method: Parametrised family $\{\mathcal{A}_\theta^\dagger\}_\theta$ where $\mathcal{A}_\theta^\dagger \colon Y \to X$.
- Existence: For every $g \in Y$, there exist a solution $\mathcal{A}_\theta^\dagger(g) \in X$ given fixed θ.
  $\implies$ Makes it possible to define reconstruction method as mapping.
- Stability: $g \mapsto \mathcal{A}_\theta^\dagger(g)$ is continuous in relevant topology for fixed θ.
  $\implies$ Small variations in data does not result in large variations in reconstruction.
- Convergence: $f_{\theta,\delta} := \mathcal{A}_\theta^\dagger(g_0 + \delta g)$ where $g_0 := \mathcal{A}(f^*)$ and $\delta := \|\delta g\|$. Show that there exists decreasing $\delta \mapsto \theta(\delta)$ (parameter selection rule) such that

$$f_{\theta(\delta),\delta} \to f \quad \text{as } \delta \to 0 \text{ where } f \text{ solves } \mathcal{A}(f) = g_0.$$

- Inverse problem: Recover $f^* \in X$ from $g = \mathcal{A}(f^*) + \delta g$.
- Reconstruction method: Parametrised family $\{\mathcal{A}_\theta^\dagger\}_\theta$ where $\mathcal{A}_\theta^\dagger \colon Y \to X$.
- Existence: For every $g \in Y$, there exist a solution $\mathcal{A}_\theta^\dagger(g) \in X$ given fixed θ.
  $\implies$ Makes it possible to define reconstruction method as mapping.
- Stability: $g \mapsto \mathcal{A}_\theta^\dagger(g)$ is continuous in relevant topology for fixed θ.
  $\implies$ Small variations in data does not result in large variations in reconstruction.
- Convergence: $f_{\theta,\delta} := \mathcal{A}_\theta^\dagger(g_0 + \delta g)$ where $g_0 := \mathcal{A}(f^*)$ and $\delta := \|\delta g\|$. Show that there exists decreasing $\delta \mapsto \theta(\delta)$ (parameter selection rule) such that

$$f_{\theta(\delta),\delta} \to f \quad \text{as } \delta \to 0 \text{ where } f \text{ solves } \mathcal{A}(f) = g_0.$$

- Convergence rates: Estimate difference between $f_{\theta,\delta}$ and a minimal norm solution. Need regularity assumptions of f^* (source conditions).
- Stability estimates: Bounds to the difference between $f_{\theta,\delta}$ and $f_{\theta,0}$ depending on δ.

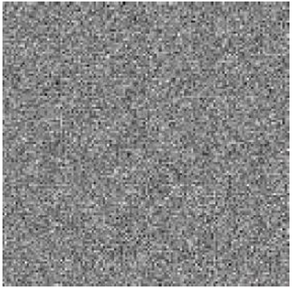
- Inverse problem: $\mathcal{A} = \mathrm{Id} \implies$ remove noise from a signal/image.

Signal: Original image

- Inverse problem: $\mathcal{A} = \mathrm{Id} \implies$ remove noise from a signal/image.

Signal: Original image

$+$

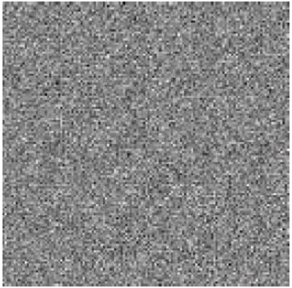


Noise

- Inverse problem: $\mathcal{A} = \mathrm{Id} \implies$ remove noise from a signal/image.

Signal: Original image

$+$



Noise

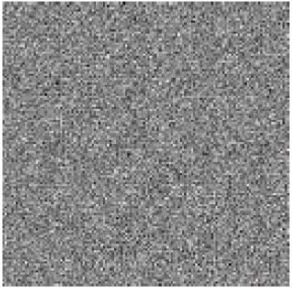
$=$

Data: Noisy image

- Inverse problem: $\mathcal{A} = \mathrm{Id} \implies$ remove noise from a signal/image.
- Removal of additive zero-mean white noise is essentially a solved problem in image processing

Signal: Original image

$+$



Noise

$=$

Data: Noisy image

- Deconvolution: $\mathcal{A}(f) = k * f$
  - k kernel.
  - k unknown (blind deconvolution).
- Tomography: $\mathcal{A}(f)(\ell) = \int_\ell f$.
- PDE parameter estimation: Forward operator $=$ solution to a PDE.
- ...

Type of regularisation methods

- Analytic methods.
- Iterative methods with early stopping.
- Variational methods.
- Learned iterative methods.

Analytic methods

- Data model: $g = \mathcal{A}(f^*)$, i.e., no specific adaptation to handle noise.
- Prior model: Features of f^* stably recoverable, e.g., if feature is a mollified version $\implies$ signal is bandlimited.
- Reconstruction method: Highly specific (depends on forward operator and acquisition geometry). $\mathcal{A}^\dagger_\theta(g) =$ stably recoverable features of f^*, main example is when features is a mollified version of the signal.
- Regularisation parameter: θ parametrises features, e.g., if feature is a mollified version then θ is typically the bandlimit, which is set using Shannon-Nyquist sampling theory.
- Examples:
  - Filtered backprojection: Recovering bandlimited function from its ray transform data (Natterer & Wübbeling, 2001).
  - Lambda-tomography: Recovering wavefront set (singularities) of function from its ray transform data (Quinto, 1993; Krishnan & Quinto, 2015).
  - Approximate inverse (Schuster, 2007; Louis, 1996).

Iterative regularisation with early stopping

- Data model: Both forward operator and data discrepancy $\mathcal{L}\colon Y \times Y \to \mathbb{R}$ are exchangeable.
- Prior model: Iterates are semi-convergent.
- Reconstruction method: $\mathcal{A}^{\dagger}_{\theta}(g)$ given by a fixed point iteration scheme for minimising $f \mapsto \mathcal{L}\bigl(\mathcal{A}(f), g\bigr)$.
- Regularisation parameter: θ number of iterates.
- Examples:
  - Conjugate gradient least squares with variants (Engl et al., 2000; Bakushinsky & Kokurin, 2004; Kaltenbacher et al., 2008; M. Burger et al., 2015).
  - Algebraic reconstruction technique with variants (Hansen, 1997; Byrne, 2008).
  - ML-EM (Natterer & Wübbeling, 2001; Byrne & Eggermont, 2015).

Variational regularisation

- Data model: Both forward operator and data discrepancy $\mathcal{L}\colon Y \times Y \to \mathbb{R}$ are exchangeable.
- Prior model: Accounted for by regulariser $\mathcal{R}_\theta\colon X \to \mathbb{R}$.
- Reconstruction method: $\mathcal{A}_\theta^\dagger(g)$ is solution to a variational problem (penalised log-likelihood):
$$\min_{f \in X} \Big[\mathcal{L}\big(\mathcal{A}(f), g\big) + \mathcal{R}_\theta(f)\Big] \quad \text{for a fixed } \theta.$$
- Regularisation parameter: θ parametrises regulariser.
- Examples:
  - Tikhonov regularisation (Engl et al., 2000).
  - Total variation regularisation (Scherzer et al., 2009; Caselles et al., 2015).
  - ...

Prior information	Regularisation functional $\mathcal{R}_\theta(f) := \theta\,\mathcal{R}(f)$
$f^* - \rho$ is sparse for some known $\rho \in X$.	$\mathcal{R}(f) = \lVert f - \rho \rVert_p$ with $0 \leq p \leq 1$ $\mathcal{R}(f) = \displaystyle\int_\Omega \Bigl(f(x) \ln \frac{f(x)}{\rho(x)} - f(x) + \rho(x) \Bigr)\,dx$
∇f^* is sparse.	$\mathcal{R}(f) = \lVert \nabla f \rVert_p = \Bigl(\displaystyle\int_\Omega \bigl\lvert \nabla f(x) \bigr\rvert^p\,dx \Bigr)^{1/p}$ with $0 \leq p \leq 1$. Case with $p = 1$ is TV-regularisation.
f^* is smooth.	$\mathcal{R}(f) = \lVert \nabla f \rVert_2 = \sqrt{\displaystyle\int_\Omega \bigl\lvert \nabla f(x) \bigr\rvert^2\,dx}.$
f^* is sparse w.r.t. $\{\phi_i\}_i$.	$\mathcal{R}(f) = \Bigl(\displaystyle\sum_i \lvert \langle f, \phi_i \rangle \rvert^p \Bigr)^{1/p}$ with $0 \leq p \leq 1$.

Three type of methods: A posteriori, a priori, and error-free parameter choice rules (Engl et al., 2000) (Bertero & Boccacci, 1998, Section 5.6).

A posteriori rules: Access to a reasonably tight estimate of the data discrepancy and/or value of regulariser at true solution, i.e., know $\epsilon > 0$ and/or $E > 0$ such that

$$\mathcal{L}\big(\mathcal{A}(f^*), g\big) \leq \epsilon \quad \text{for } g := \mathcal{A}(f^*) + \delta g \qquad \text{and/or} \quad \mathcal{R}(f^*) \leq E.$$

- Morozov principle: Choose θ so $\mathcal{L}\big(\mathcal{A}(f_\theta), g\big) \leq \epsilon$ (Morozov, 1966).
- Miller method: Choose θ so that $\mathcal{L}\big(\mathcal{A}(f_\theta), g^\delta\big) \leq \epsilon$ and $\mathcal{R}(f_\theta) \leq E$ (Miller, 1970).

Here, f^* is the true (unknown) solution and $f_\theta := \mathcal{A}_\theta^\dagger(g)$ is the regularised solution.

Three type of methods: A posteriori, a priori, and error-free parameter choice rules (Engl et al., 2000) (Bertero & Boccacci, 1998, Section 5.6).

A priori rules: Determine the regularisation parameter solely from knowledge of the noise level in data.

Variational regularisation methods

Common parameter choice rules

Three type of methods: A posteriori, a priori, and error-free parameter choice rules (Engl et al., 2000) (Bertero & Boccacci, 1998, Section 5.6).

Error-free parameter choice rules: Use data to guide choice of parameter, e.g., by balancing principles between the error in the fidelity and the regularisation terms.

- Generalised cross-validation: Let $f_{k,\theta} \in X$ denote the regularised solution when we have removed the k:th component g_k of the data g. Choose θ in order to predict missing data values (Golub et al., 1979), i.e.,

$$\mathcal{A}(f_{k,\theta})_k \approx g_k \quad \text{by minimising} \quad \sum_{i=1}^{m} \bigl|\mathcal{A}(f_{k,\theta}) - g_k\bigr|.$$

- L-curve: θ is chosen where log-log plot of $\theta \mapsto \bigl(\mathcal{L}\bigl(\mathcal{A}(f_\theta), g\bigr), \mathcal{R}(f_\theta)\bigr)$ has highest curvature (i.e., a corner) (Hansen, 1992).

Three type of methods: A posteriori, a priori, and error-free parameter choice rules (Engl et al., 2000) (Bertero & Boccacci, 1998, Section 5.6).

Current status:

- Most of the work on parameter choice techniques addresses the case of a single scalar parameter.
- Much of the theory is developed for additive Gaussian noise, i.e., when data discrepancy $\mathcal{L}$ is a 2-norm.
- For error-free parameter choice rules, convergence $f_{\theta(\delta)} \to f^*$ as $\delta \to 0$ cannot be guaranteed (Bakushinskii, 1984).
- Error-free parameter choice rules computationally very demanding (requires solutions for varying values of regularisation parameter).
- Although many rules have been proposed, very few of them are used in practice.

Learning parameters in regulariser

Bi-level optimisation

Variational model: Define $\mathcal{A}_\theta^\dagger \colon Y \to X$ as

$$\mathcal{A}_\theta^\dagger(g) \in \arg\min_f \Big[\mathcal{L}\big(\mathcal{A}(f), g\big) + \mathcal{R}_\theta(f)\Big] \quad \text{for } g \in Y.$$

- Supervised training data $(f_i, g_i) \in X \times Y$ such that $g_i \approx \mathcal{A}(f_i)$.
- Learn regularisation parameter θ from supervised training data by minimising empirical risk:

$$\theta^* \in \arg\min_\theta \Big[\frac{1}{m}\sum_{i=1}^m \ell_X\big(\mathcal{A}_\theta^\dagger(g_i), f_i\big)\Big]$$

where $\ell_X \colon X \times X \to \mathbb{R}$ is a loss function.

Bi-level optimisation

- Bi-level optimisation: Find reconstruction method $\mathcal{A}^{\dagger}_{\theta^*} \colon Y \to X$ from training data (f_i, g_i) where

$$\begin{cases} \theta^* \in \arg\min_{\theta} \Big[\dfrac{1}{m} \displaystyle\sum_{i=1}^{m} \ell_X\big(\mathcal{A}^{\dagger}_{\theta}(g_i), f_i\big) \Big] \\ \mathcal{A}^{\dagger}_{\theta}(g) \in \arg\min_{f} \Big[\mathcal{L}\big(\mathcal{A}(f), g\big) + \mathcal{R}_{\theta}(f) \Big] \end{cases}$$

- θ^* yields a minimiser of the variational model that minimises the empirical risk.
- Existence of solution to bi-level optimisation far from obvious, needs to be proved. Uniqueness does not hold in general.
- Computing derivative of $\theta \mapsto \mathcal{A}^{\dagger}_{\theta}(g)$ is non-trivial and computationally demanding.

Example of bi-level optimisation

Anisotropic weighted Dirichlet/total variation: (Haber & Tenorio, 2003)

$$\mathcal{R}_\theta(f) := \left\|\theta(\,\cdot\,)\nabla f(\,\cdot\,)\right\|_2^2 \quad \text{where } \theta\colon \Omega \to \mathbb{R}$$

$$\mathcal{R}_\theta(f) := \left\|\theta\big(|\nabla f(\,\cdot\,)|\big)\right\|_1 \quad \text{where } \theta\colon \mathbb{R} \to \mathbb{R}$$

Example of bi-level optimisation

Anisotropic weighted Dirichlet/total variation: (Haber & Tenorio, 2003)

$$\mathcal{R}_\theta(f) := \left\|\theta(\,\cdot\,)\nabla f(\,\cdot\,)\right\|_2^2 \quad \text{where } \theta\colon \Omega \to \mathbb{R}$$

$$\mathcal{R}_\theta(f) := \left\|\theta\big(|\nabla f(\,\cdot\,)|\big)\right\|_1 \quad \text{where } \theta\colon \mathbb{R} \to \mathbb{R}$$

Total generalised variation: 2nd order case (TGV^2) (Bredies et al., 2010):

$$\mathcal{R}_\theta(f) = \min_{\boldsymbol{\nu}}\Big[\theta_1\|\nabla f - \boldsymbol{\nu}\|_1 + \theta_2\|\boldsymbol{\nabla}\boldsymbol{\nu}\|_1\Big] \quad \text{where } \boldsymbol{\nu}\colon \mathbb{R}^n \to \mathbb{R}^n$$

with $\theta = (\theta_1, \theta_2) \in \mathbb{R}^2$ and $\boldsymbol{\nabla}\boldsymbol{\nu} := \left[\frac{1}{2}(\partial_j\nu_i + \partial_i\nu_j)\right]_{i,j}$ (symmetrised gradient).

- Denoising using Huber-smoothed version of TGV^2 with an added H^1-regularisation incl. proof of existence and algorithms (De los Reyes et al., 2017).
- Denoising using Infimal Convolution Total Variation (De los Reyes et al., 2017).

Example of bi-level optimisation

Weighted sum of ℓ^p-regularisers: Given linear $K\colon X \to X$,

$$\mathcal{R}_\theta(f) := \frac{1}{p} \sum_{i=1}^{N} \theta_i \big\| K(f) \big\|_p^p \quad \text{with } \theta = (\theta_i)_i \in \mathbb{R}^N.$$

- $p = 1$ generalises total variation.
- Denoising incl. proof of existence for $p = 1, 2$ (Kunisch & Pock, 2013).
- Semi-smooth Newton algorithm can be used to solve the bilevel optimisation problem for $p = 1, 2$ (Kunisch & Pock, 2013).

Example of bi-level optimisation

Field of Experts model: Given $\rho \colon \mathbb{R} \to \mathbb{R}$ (potential function),

$$\mathcal{R}_\theta(f) := \sum_{i=1}^{N} w_i \left[\int_\Omega \rho\big((f * k_i)(x)\big)\, dx \right] \quad \text{with } \theta = (w_i, k_i)_i \in (\mathbb{R} \times X)^N.$$

- Filters $k_i \colon \Omega \to \mathbb{R}$ parametrised by finite dimensional parameters:
  - Standard finite difference approximations of first- and second-order derivatives
  - Higher-order linear operators obtained from dictionary atoms, like basis vectors of the discrete cosine transform (Kunisch & Pock, 2013).
- Denoising (Samuel & Tappen, 2009; Kunisch & Pock, 2013).
- Possible to learn regularisation terms and parameters from the training data using a deep neural network, leads to the 'Learned Experts' Assessment-based Reconstruction Network (LEARN)' method (H. Chen et al., 2018).

Example of bi-level optimisation

Nonconvex fields of experts model:

$$\mathcal{R}_\theta(f) := \sum_{i=1}^{N} w_i \left[\int_\Omega \rho_i\big((f * k_i)(x)\big)\, dx \right] \quad \text{with } \theta = (w_i, \rho_i, k_i).$$

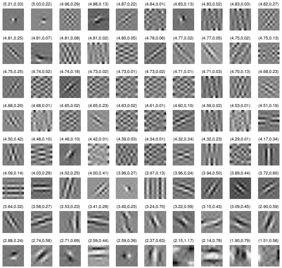
- Filters $k_i : \Omega \to \mathbb{R}$ and potential functions $\rho_i : \mathbb{R} \to \mathbb{R}$ parametrised by finite dimensional parameters.
- Denoising (Roth & Black, 2009; Y. Chen et al., 2014).
- Computing gradients of empirical risk w.r.t. θ can be done via implicit differentiation, very time consuming ...
- 2D denoising example (Y. Chen et al., 2014): $N = 80$, filters $k_i : \mathbb{R}^2 \to \mathbb{R}$ symmetric with 9×9 pixel support, and $\rho_i(t) = \lambda_i \log(1 + \beta_i t^2) \implies \theta \in \mathbb{R}^{6480}$.

  Training on test data with $m = 200$ images took two weeks!

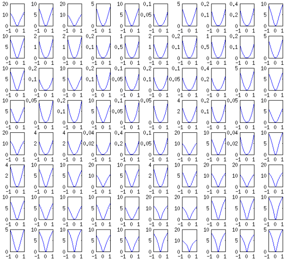
Example of bi-level optimisation

Nonconvex fields of experts model:

$$\mathcal{R}_\theta(f) := \sum_{i=1}^{N} w_i \left[\int_\Omega \rho_i\big((f * k_i)(x)\big)\,dx \right] \quad \text{with } \theta = (w_i, \rho_i, k_i).$$



Learned kernels $k_i\colon \Omega \to \mathbb{R}$.



Potential functions $\rho_i\colon \mathbb{R} \to \mathbb{R}$.

Example of bi-level optimisation

Sparse recovery (analysis formulation): $\mathcal{R}_\theta(f) := C\big(\rho_\theta(f)\big)$ with $\theta \subset X$.

- $C \colon \ell^2 \to \mathbb{R}$ is a fix proper lower-semicontinuous function, e.g., ℓ^1-norm.
- $\rho_\theta \colon X \to \ell^2$ analysis operator with θ being the dictionary.
- Denoising with $C =$ smooth version of ℓ^1-penalty incl. proof that $\theta \mapsto \mathcal{A}^\dagger_\theta(g)$ is differentiable:
$$\mathcal{A}^\dagger_\theta(g) \in \arg\min_f \Big[\mathcal{L}\big(\mathcal{A}(f), g\big) + \mathcal{R}_\theta(f)\Big]$$
incl. explicit expression for the derivative (Theorem 1) (Peyré & Fadili, 2011).
- It is more common to consider the synthesis (sparse coding) formulation, which we deal with in part related to 'Sparse models'.

Summary of bi-level optimisation

- Computationally demanding due to implicit differentiation:
  - For each θ, solve the inner problem $\mathcal{A}_\theta^\dagger(g) \in \arg\min_f \Big[\mathcal{L}\big(\mathcal{A}(f), g\big) + \mathcal{R}_\theta(f)\Big]$ exactly.
  - Invert the Hessian of $\theta \mapsto \mathcal{A}_\theta^\dagger(g)$.
- Alternative: Unroll T steps of an iterative algorithm (e.g. gradient descent):

$$
\begin{cases}
\theta^* \in \arg\min_\theta \Big[\dfrac{1}{m}\displaystyle\sum_{i=1}^{m} \ell_X\big(\mathcal{A}_{\theta,T}^\dagger(g_i), f_i\big)\Big] \\
f_\theta^{i+1} := f_\theta^i - \omega_i \nabla\Big[\mathcal{L}\big(\mathcal{A}(\,\cdot\,), g\big) + \mathcal{R}_\theta(\,\cdot\,)\Big](f_\theta^i) \qquad \text{for } i = 1, \dots, T-1 \\
\mathcal{A}_{\theta,T}^\dagger(g) := f_\theta^T
\end{cases}
$$

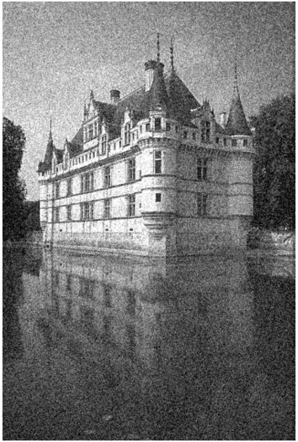
  - Computing gradient of objective can be done efficiently.
  - Taking only a few iterates (T small) already works very well.

Incremental gradient scheme

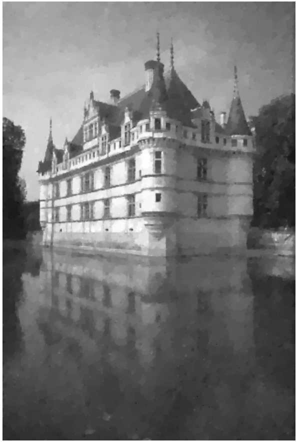
- Assume you can decompose data log-likelihood and regularisation functional into M components:
$$\mathcal{L}\big(\mathcal{A}(f), g\big) = \sum_{k=1}^{M} \mathcal{L}_k\big(\mathcal{A}(f), g\big) \quad \text{and} \quad \mathcal{R}_\theta(f) := \mathcal{R}_0(f) + \sum_{k=1}^{M} \mathcal{R}_{\theta_k}(f).$$
- $\mathcal{R}_0$ typically non-smooth, e.g., handles additional sparsity priors or constraints.
- Reconstruction method: Let $F_k(f; g, \theta_k) := \mathcal{L}_k\big(\mathcal{A}(f), g\big) + \mathcal{R}_{\theta_k}(f)$ and $i_k = \text{mod}(k, T) + 1$, perform T incremental proximal steps:
$$\begin{cases} f_\theta^{k+1} := \operatorname{prox}_{\omega_k \mathcal{R}_0}\big(f_\theta^k - \omega_k \nabla F_{i_k}(f_\theta^k; g, \theta_{i_k})\big) & \text{for } k = 1, \dots, T-1 \\ \mathcal{A}^\dagger_{\theta,T}(g) := f_\theta^T \end{cases}$$
- Used for denoising and MRI reconstruction from under-sampled k-space data (Kobler et al., 2017; Hammernik et al., 2018).
- Close connections to residual networks.

Example from 2D denoising

Signal: True image



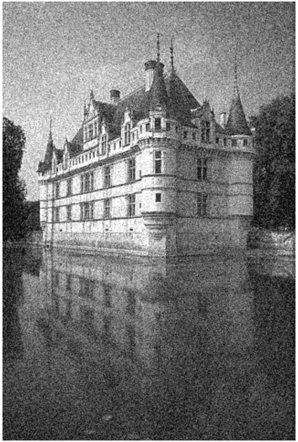
Data: Noisy image.



Total variation.

Example from 2D denoising

Signal: True image

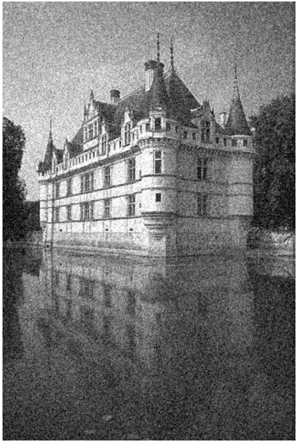


Data: Noisy image.

TGV^2.

Example from 2D denoising

Signal: True image

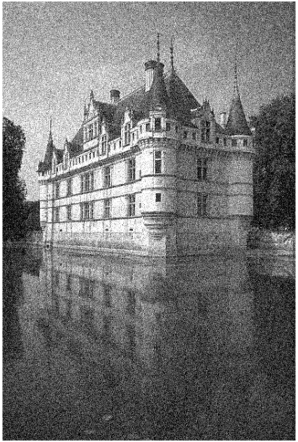


Data: Noisy image.

Sparse dictionary (with DCT).

Example from 2D denoising

Signal: True image

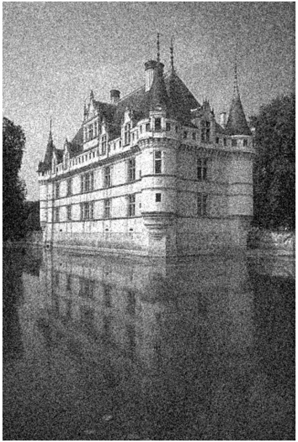


Data: Noisy image.

Nonconvex fields of experts.

Example from 2D denoising

Signal: True image



Data: Noisy image.

Incremental gradient scheme.

Sparse models

Basic notions

X separable Hilbert space (has countable ON-basis).

- Dictionary: A collection $\mathscr{D} := \{\phi_i\}_i \subset X$, elements are called atoms. Emerge from one of two sources (Lanusse et al., 2014; Bruckstein et al., 2009; Rubinstein et al., 2010; G. Chen & Needell, 2016):
  - Analytic: Based on a mathematical model.
  - Data-dependent: Derived from a set of realisations $f_i \in X$.
- Frame: $\mathscr{D} := \{\phi_i\}_i$ is a frame if there exists $C_1, C_2 > 0$ such that

$$C_1 \|f\|^2 \leq \sum_i \big|\langle f, \phi_i \rangle\big|^2 \leq C_2 \|f\|^2 \quad \text{for any } f \in X.$$

- Tight frame: Case when $C_1 = C_2 = 1$.
- Over-complete/redundant: The frame does *not* form a basis for X. Redundant dictionaries, e.g., translation invariant wavelets, often work better than non-redundant (Peyré & Fadili, 2011; Elad, 2010).

Analysis and synthesis

X separable Hilbert space, $\mathscr{D} := \{\phi_i\}_i \subset X$ fixed dictionary.

- Analysis operator: $\mathsf{E} \colon X \to \ell^2$ where $\mathsf{E}(f) := \big(\langle f, \phi_i \rangle\big)_i$
- Synthesis operator: $\mathsf{S} \colon \ell^2 \to X$ is the adjoint of the analysis operator, i.e.,

$$\mathsf{S}\big((\gamma_i)_i\big) := \sum_i \gamma_i \phi_i.$$

- Frame operator: $\mathsf{S} \circ \mathsf{E} \colon X \to X$, i.e.,

$$(\mathsf{S} \circ \mathsf{E})(f) = \sum_i \langle f, \phi_i \rangle \phi_i.$$

Notion of sparsity

X separable Hilbert space, $\mathscr{D} := \{\phi_i\}_i \subset X$ fixed dictionary.

- Sparsity: $f \in X$ is s-sparse w.r.t. $\mathscr{D}$ if

$$\left\|\mathrm{E}(f)\right\|_0 = \#\left\{i \mid \langle f, \phi_i \rangle \neq 0\right\} \leq s.$$

- Compressible: $f \in X$ is compressible w.r.t. $\mathscr{D}$ if the following power law decay holds:

$$\left|\widetilde{\mathrm{E}}(f)_k\right| \leq Ck^{-1/q} \quad \text{for some } C > 0 \text{ and } 0 < q < 1.$$

  $\widetilde{\mathrm{E}}(f) =$ a non-increasing rearrangement of the sequence $\mathrm{E}(f)$.

- Sparse signals are compressible.
  q small $\implies$ compressibility = sparsity.
- $\mathrm{E}_s(f) =$ vector consisting of the s largest (in magnitude) coefficients of the sequence $\mathrm{E}(f)$.

- Inverse problem: Recover $f^* \in X$ from $g = \mathcal{A}(f^*) + \delta g$.
- Data log likelihood: $\mathcal{L}\colon Y \times Y \to \mathbb{R}$.
- Prior model: $f^* \in X$ is compressible w.r.t. given dictionary $\mathscr{D} := \{\phi_i\}_i$.
- Reconstruction method (sparse recovery)
  - Synthesis (sparse coding):
$$\mathcal{A}^\dagger_\theta(g) := \mathsf{S}(\gamma^*) \quad \text{where} \quad \gamma^* \in \operatorname*{arg\,min}_{\gamma \in \ell^2} \Big[\mathcal{L}\Big(\mathcal{A}(\mathsf{S}(\gamma)), g\Big) + \theta \|\gamma\|_0\Big].$$
  - Analysis:
$$\mathcal{A}^\dagger_\theta(g) \in \operatorname*{arg\,min}_{f \in X} \Big[\mathcal{L}\big(\mathcal{A}(f), g\big) + \theta \big\|\mathsf{E}(f)\big\|_0\Big].$$
- $\mathscr{D}$ is an ON basis $\implies$ synthesis and analysis formulations are equivalent.

- Inverse problem: Recover $f^* \in X$ from $g = \mathcal{A}(f^*) + \delta g$.
- Data log likelihood: $\mathcal{L}\colon Y \times Y \to \mathbb{R}$.
- Prior model: $f^* \in X$ is compressible w.r.t. given dictionary $\mathscr{D} := \{\phi_i\}_i$.
- Reconstruction method (sparse recovery)
  - Synthesis (sparse coding):
$$\mathcal{A}_\theta^\dagger(g) := \mathsf{S}(\gamma^*) \quad \text{where} \quad \gamma^* \in \operatorname*{arg\,min}_{\gamma \in \ell^2} \Bigl[\mathcal{L}\Bigl(\mathcal{A}(\mathsf{S}(\gamma)), g \Bigr) + \theta \|\gamma\|_0 \Bigr].$$
  - Analysis:
$$\mathcal{A}_\theta^\dagger(g) \in \operatorname*{arg\,min}_{f \in X} \Bigl[\mathcal{L}\bigl(\mathcal{A}(f), g \bigr) + \theta \bigl\| \mathsf{E}(f) \bigr\|_0 \Bigr].$$
- $\mathscr{D}$ is an ON basis $\implies$ synthesis and analysis formulations are equivalent.
- Sparse recovery is NP-hard.

- Inverse problem: Recover $f^* \in X$ from $g = \mathcal{A}(f^*) + \delta g$.
- Data log likelihood: $\mathcal{L} \colon Y \times Y \to \mathbb{R}$.
- Prior model: $f^* \in X$ is compressible w.r.t. given dictionary $\mathscr{D} := \{\phi_i\}_i$.
- Reconstruction method (sparse recovery)
  - Synthesis (sparse coding):

$$\mathcal{A}^\dagger_\theta(g) := \mathsf{S}(\gamma^*) \quad \text{where} \quad \gamma^* \in \operatorname*{arg\,min}_{\gamma \in \ell^2} \Big[\mathcal{L}\Big(\mathcal{A}(\mathsf{S}(\gamma)), g \Big) + \theta \|\gamma\|_0 \Big].$$

  - Analysis:

$$\mathcal{A}^\dagger_\theta(g) \in \operatorname*{arg\,min}_{f \in X} \Big[\mathcal{L}\big(\mathcal{A}(f), g \big) + \theta \big\| \mathsf{E}(f) \big\|_0 \Big].$$

- $\mathscr{D}$ is an ON basis $\implies$ synthesis and analysis formulations are equivalent.
- Sparse recovery is NP-hard.
  - Greedy approach.
  - Convex relaxation $p > 0$, case $p = 1$ starting point for sparse signal processing (Elad, 2010; Foucart & Rauhut, 2013).

- Inverse problem: Recover $f^* \in X$ from $g = \mathcal{A}(f^*) + \delta g$.
- Data log likelihood: $\mathcal{L}\colon Y \times Y \to \mathbb{R}$.
- Prior model: $f^* \in X$ is compressible w.r.t. given dictionary $\mathscr{D} := \{\phi_i\}_i$.
- Reconstruction method (sparse recovery)
  - Synthesis (sparse coding):
$$\mathcal{A}_\theta^\dagger(g) := \mathsf{S}(\gamma^*) \quad \text{where} \quad \gamma^* \in \operatorname*{arg\,min}_{\gamma \in \ell^2} \Big[\mathcal{L}\Big(\mathcal{A}(\mathsf{S}(\gamma)), g\Big) + \theta \|\gamma\|_p\Big].$$
  - Analysis:
$$\mathcal{A}_\theta^\dagger(g) \in \operatorname*{arg\,min}_{f \in X} \Big[\mathcal{L}\big(\mathcal{A}(f), g\big) + \theta \big\|\mathsf{E}(f)\big\|_p\Big].$$
- $\mathscr{D}$ is an ON basis $\implies$ synthesis and analysis formulations are equivalent.
- Sparse recovery is NP-hard.
  - Greedy approach.
  - Convex relaxation $p > 0$, case $p = 1$ starting point for sparse signal processing (Elad, 2010; Foucart & Rauhut, 2013).

Theorem (Candès et al., 2006b)

Let $X = \mathbb{R}^n$, $Y = \mathbb{R}^m$, and assume $\mathcal{A}\colon X \to Y$ be a linear mapping whose matrix satisfies the restricted isometry property. If $g = \mathcal{A}(f^*) + \delta g$ with $\|\delta g\| \leq \delta$ and

$$\widehat{f}_\delta := \operatorname*{arg\,min}_{f \in X} \|f\|_1 \quad \text{subject to} \quad \big\|\mathcal{A}(f) - g\big\|_2 \leq \delta,$$

then

$$\big\|\widehat{f}_\delta - f^*\big\|_2 \leq C\left[\delta + \frac{\big\|f^* - f_s^*\big\|_2}{\sqrt{s}}\right].$$

f_s^* = vector consisting of the s largest (in magnitude) coefficients of f^*.

Restricted isometry property (RIP): $\mathcal{A}$ satisfies the following for sufficiently small $\epsilon_s > 0$:

$$(1 - \epsilon_s)\|f\|_2^2 \leq \big\|\mathcal{A}(f)\big\|_2^2 \leq (1 + \epsilon_s)\|f\|_2^2 \quad \text{for all } s\text{-sparse } f \in X.$$

RIP $\implies$ coherence (columns of $\mathcal{A}$ are ‘uncorrelated’).

Theorem (Candès et al., 2006b)

Let $X = \mathbb{R}^n$, $Y = \mathbb{R}^m$, and assume $\mathcal{A}\colon X \to Y$ be a linear mapping whose matrix satisfies the restricted isometry property. If $g = \mathcal{A}(f^*) + \delta g$ with $\|\delta g\| \leq \delta$ and

$$\widehat{f}_\delta := \arg\min_{f \in X} \|f\|_1 \quad \text{subject to} \quad \big\|\mathcal{A}(f) - g\big\|_2 \leq \delta,$$

then

$$\big\|\widehat{f}_\delta - f^*\big\|_2 \leq C\left[\delta + \frac{\big\|f^* - f^*_s\big\|_2}{\sqrt{s}}\right].$$

f^*_s = vector consisting of the s largest (in magnitude) coefficients of f^*.

Matrices satisfying RIP: Sub-Gaussian matrices, partial bounded orthogonal matrices (G. Chen & Needell, 2016).

Theorem (Candès et al., 2006b)

Let $X = \mathbb{R}^n$, $Y = \mathbb{R}^m$, and assume $\mathcal{A}\colon X \to Y$ be a linear mapping whose matrix satisfies the restricted isometry property. If $g = \mathcal{A}(f^*) + \delta g$ with $\|\delta g\| \leq \delta$ and

$$\widehat{f}_\delta := \operatorname*{arg\,min}_{f \in X} \|f\|_1 \quad \text{subject to} \quad \big\|\mathcal{A}(f) - g\big\|_2 \leq \delta,$$

then

$$\big\|\widehat{f}_\delta - f^*\big\|_2 \leq C \left[\delta + \frac{\big\|f^* - f^*_s\big\|_2}{\sqrt{s}}\right].$$

f^*_s = vector consisting of the s largest (in magnitude) coefficients of f^*.

- Reconstruction error is at most proportional to the norm of the noise in the data and the tail $f^* - f^*_s$ of the signal.
- Error bound is optimal (up to precise value of C) (Cohen et al., 2009).
- If f^* is s-sparse and $\delta = 0$ (no noise) $\implies$ f^* can be reconstructed exactly.

Theorem (Candès et al., 2006b)

Let $X = \mathbb{R}^n$, $Y = \mathbb{R}^m$, and assume $\mathcal{A}\colon X \to Y$ be a linear mapping whose matrix satisfies the restricted isometry property. If $g = \mathcal{A}(f^*) + \delta g$ with $\|\delta g\| \leq \delta$ and

$$\widehat{f}_\delta := \operatorname*{arg\,min}_{f \in X} \|f\|_1 \quad \text{subject to} \quad \big\|\mathcal{A}(f) - g\big\|_2 \leq \delta,$$

then

$$\big\|\widehat{f}_\delta - f^*\big\|_2 \leq C\left[\delta + \frac{\big\|f^* - f^*_s\big\|_2}{\sqrt{s}}\right].$$

f^*_s = vector consisting of the s largest (in magnitude) coefficients of f^*.

- If f^* is compressible, then

$$\big\|\widehat{f}_\delta - f^*\big\|_2 \leq C\Big(\delta + C' s^{1/2-1/q}\Big).$$

Sparse coding: Given dictionary $\{\phi_i\}_i$, compute sparse representation

$$\gamma^* \in \underset{\gamma \in \ell^2}{\arg\min} \left[\Big\| \mathcal{A}\Big(\sum_i \gamma_i \phi_i\Big) - g \Big\|_2^2 + \theta \|\gamma\|_0 \right].$$

- Greed approaches: Build up an approximation one step at a time by making locally optimal choices at each step.
  - Iterative (hard) thresholding (Blumensath & Davies, 2008; Foucart, 2016).

    $$\gamma^{i+1} = T_s\Big(\gamma^i - W^*\big(W(\gamma^i) - g\big)\Big) \quad \text{where } W(\gamma) := \mathcal{A}\Big(\sum_i \gamma_i \phi_i\Big).$$

    $T_s(\gamma) =$ sets all but the largest (in magnitude) s elements of γ to zero. Proximal-gradient method with proximal of the function that is 0 at 0 and 1 everywhere else.
  - Matching pursuit (MP) (S. G. Mallat & Zhang, 1993).
  - Orthogonal matching pursuit (OMP) (Tropp & Gilbert, 2007) and variants like StOMP (Donoho et al., 2012), ROMP (Needell & Vershynin, 2009), and CoSamp (Needell & Tropp, 2009).

Sparse coding: Given dictionary $\{\phi_i\}_i$, compute sparse representation

$$\gamma^* \in \underset{\gamma \in \ell^2}{\arg\min} \left[\left\| \mathcal{A}\left(\sum_i \gamma_i \phi_i \right) - g \right\|_2^2 + \theta \|\gamma\|_1 \right].$$

- Convex relaxation: Most common example is to replace ℓ_0-norm with ℓ_1-norm $\implies$ Basis pursuit (BP) (Candès et al., 2006b), also called Lasso in the statistics literature (Tibshirani, 1996).
  - Interior-point methods (Candès et al., 2006a; Kim et al., 2007).
  - Projected gradient methods (Figueiredo et al., 2007).
  - Iterative soft thresholding (forward-backward/proximal-gradient, proximal operator of ℓ_1 is sometimes called soft thresholding operator) (Fornasier & Rauhut, 2008).
  - Iterative thresholding (Daubechies et al., 2004).
  - Fast proximal gradient methods (FISTA and variants) (Bubeck, 2015).

Sparse coding: Given dictionary $\{\phi_i\}_i$, compute sparse representation

$$\gamma^* \in \operatorname*{arg\,min}_{\gamma \in \ell^2} \left[\left\| \mathcal{A}\Big(\sum_i \gamma_i \phi_i\Big) - g \right\|_2^2 + \theta \|\gamma\|_0 \right].$$

- Combinatorial algorithms: Acquire highly structured samples of the signal that support rapid reconstruction via group testing. This class includes Fourier sampling, chaining pursuit, and HHS pursuit, e.t.c. (Berinde et al., 2008).

Sparse coding: Given dictionary $\{\phi_i\}_i$, compute sparse representation

$$\gamma^* \in \operatorname*{arg\,min}_{\gamma \in \ell^2} \left[\left\| \mathcal{A}\Big(\sum_i \gamma_i \phi_i\Big) - g \right\|_2^2 + \theta \|\gamma\|_0 \right].$$

- Greedy methods will in general not give the same solution as convex relaxation. If the restricted nullspace property holds, then both approaches have the same solution.
- Convex relaxation: Succeed with a very small number of measurements, but they tend to be computationally burdensome.
- Combinatorial algorithms: Extremely fast (sublinear in the length of the target signal) but they require very specific structure of $\mathcal{A}$ and a large number of samples.
- Greedy methods: Intermediate in their running time and sampling efficiency.

- Patch-based local models: Split signal into patches (segments), process patches.
- Leading denoising methods are based on patch-based local models.
  - K-SVD: Sparse coding of image patches (Elad & Aharon, 2006a)
  - BM3D: Combines sparsity and self-similarity (Dabov et al., 2007)
  - EPLL: Gaussian mixture model of the image patches (Zoran & Weiss, 2011)
  - Deep convolutional neural networks (CNNs) (H. C. Burger et al., 2012)
  - NCSR: Non-local sparsity with centralised coefficients (Dong et al., 2013)
  - WNNM: Weighted nuclear norm regularisation of image patches (Gu et al., 2014)
  - SSC-GSM: Nonlocal sparsity with a Gaussian scale mixture (Dong et al., 2015)
- Sparse-Land model: Each patch is sparse w.r.t. some global dictionary, sparse coding applied patch-wise (Elad & Aharon, 2006b; Dong et al., 2011; Mairal & Ponce, 2014; Romano & Elad, 2015; Sulam & Elad, 2015).
- Computationally more feasible for dictionary learning.

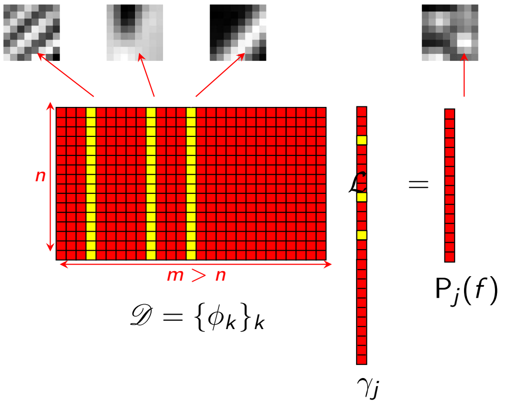
Sparsity of patches

- Dictionary: $\mathscr{D} = \{\phi_k\}_k$
- Patch: $\mathrm{P}_j(f) = f|_{\Omega_j}$
- Model: Patches sparse in $\mathscr{D}$.
- Patch-wise sparse coding:

$$\mathrm{P}_j(f) = \sum_k \gamma_{j,k} \phi_k$$

$\gamma_{j,k} \approx 0$ for most k.

Illustration in discrete setting



- Inverse problem: Recover $f^* \in X$ from $g = f^* + \delta g$.
- Data log likelihood: $\mathcal{L} \colon Y \times Y \to \mathbb{R}$.
- Prior model (Sparse-Land model): Given dictionary $\mathscr{D} \subset X$ and $\mathsf{P}_j \colon X \to X$ (patch extraction operator) for $j = 1, \dots, N$, assume

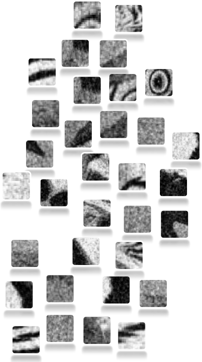
$$f^* = \sum_{j=1}^{N} \mathsf{P}_j(f^*) \quad \text{where } \mathsf{P}_j(f^*) \in X \text{ is compressible w.r.t. } \mathscr{D}.$$

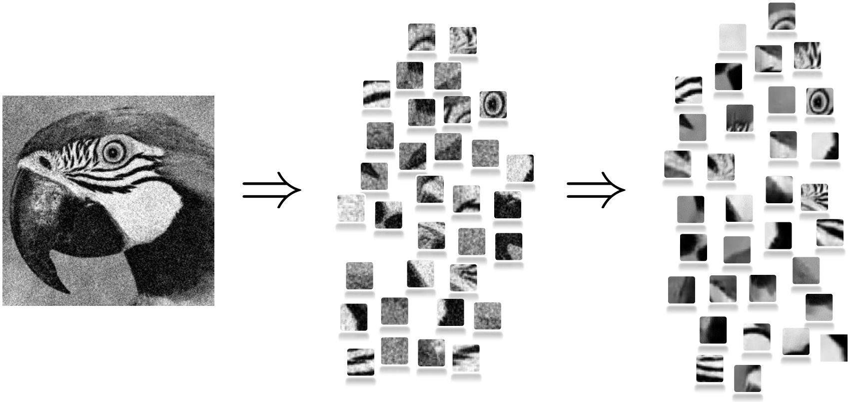
- Denoising:
  - Prior model can be applied to data $\implies$ $\mathsf{P}_j(g)$ compressible w.r.t. $\mathscr{D}$.
  - Denoised patch: Given by synthesis $\mathsf{S}(\gamma_j^*) \in X$ where $\gamma_j^* \in \ell^2$ solves sparse coding:

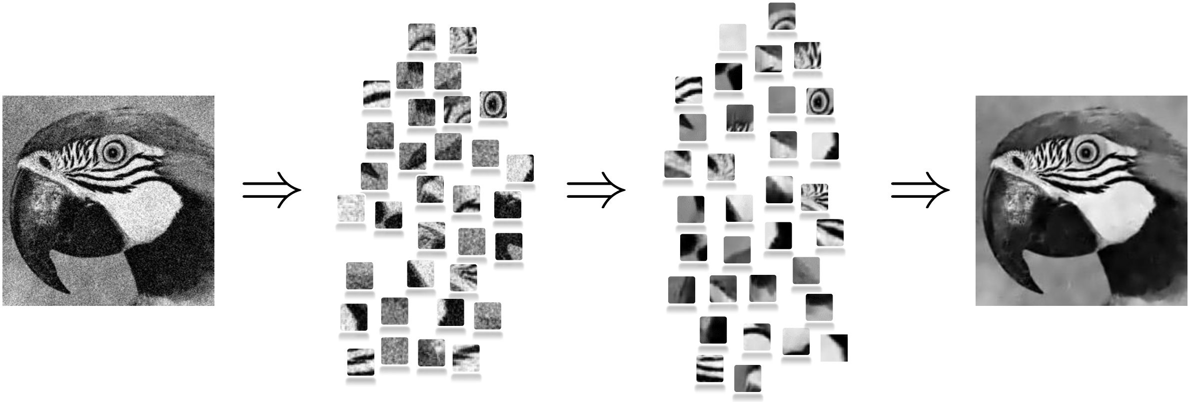
$$\gamma_j^* := \arg\min_{\gamma_j} \|\gamma_j\|_0 \quad \text{subject to } \left\|\mathsf{P}_j(g) - \mathsf{S}(\gamma_j)\right\|_2 \leq \epsilon.$$

  - Denoised image: $\widehat{f} = \sum_j \mathsf{S}(\gamma_j)$.

$\Longrightarrow$







Global dictionary based statistical iterative reconstruction (GDSIR)

- Inverse problem: Recover $f^* \in X$ from $g = \mathcal{A}(f^*) + \delta g$.
- Prior model: Sparse-Land model with N patches w.r.t. given dictionary $\mathscr{D} \subset X$ and associated synthesis operator $\mathsf{S} \colon \ell^2 \to X$.
- Reconstruction method: No sense to divide data into patches, instead

$$\min_{f \in X, \gamma_i \in \ell^2} \Big[\mathcal{L}\big(\mathcal{A}(f), g\big) + \mathcal{R}_\theta(f, \gamma_1, \dots, \gamma_N) \Big]$$

where

$$\mathcal{R}_\theta(f, \gamma_1, \dots, \gamma_N) := \sum_{j=1}^{N} \Big[\lambda_j \big\| \mathsf{P}_j(f) - \mathsf{S}(\gamma_j) \big\|_2^2 + \mu_j \|\gamma_j\|_p^p \Big]$$

with $\theta = (\lambda_j, \mu_j)_{j=1}^N \in (\mathbb{R}^2)^N$.

Global dictionary based statistical iterative reconstruction (GDSIR)

- Inverse problem: Recover $f^* \in X$ from $g = \mathcal{A}(f^*) + \delta g$.
- Prior model: Sparse-Land model with N patches w.r.t. given dictionary $\mathscr{D} \subset X$ and associated synthesis operator $\mathsf{S} \colon \ell^2 \to X$.
- Reconstruction method: Intertwined scheme (Bai et al., 2017):

$$\begin{cases} f^{i+1} := \operatorname*{arg\,min}_{f \in X} \Bigl[\mathcal{L}\bigl(\mathcal{A}(f), g\bigr) + \sum_{j=1}^{N} \lambda_j \Bigl[\bigl\| \mathsf{P}_j(f) - \mathsf{S}(\gamma_j^i) \bigr\|_2^2 \Bigr] \Bigr] \\ \gamma_j^{i+1} := \operatorname*{arg\,min}_{\gamma \in \ell^2} \Bigl[\lambda_j \bigl\| \mathsf{P}_j(f^{i+1}) - \mathsf{S}(\gamma_j) \bigr\|_2^2 + \mu_j \|\gamma_j\|_p^p \Bigr] \quad \text{for } j = 1, \ldots, N. \end{cases}$$

How to find the dictionary?

- Determine jointly while performing reconstruction (joint reconstruction & dictionary learning).
- Specify analytically.
- Determine from example data (dictionary learning).

Joint reconstruction & dictionary learning

Example of approach

Adaptive dictionary based statistical iterative reconstruction (ADSIR)

- Inverse problem: Recover $f^* \in X$ from $g = \mathcal{A}(f^*) + \delta g$.
- Prior model: Sparse-Land model as in GDSIR.
- Reconstruction method: Adds dictionary $\mathscr{D} = \{\phi_i\}_i \subset X$ as variable to GDSIR (Xu et al., 2012), see also (Chun et al., 2017).

$$\min_{\substack{f \in X \\ \gamma_i \in \ell^2, \mathscr{D} \subset X}} \Big[\mathcal{L}\big(\mathcal{A}(f), g\big) + \mathcal{R}_\theta(f, \gamma_1, \ldots, \gamma_N, \mathscr{D}) \Big]$$

where

$$\mathcal{R}_\theta(f, \gamma_1, \ldots, \gamma_N, \mathscr{D}) := \sum_{j=1}^{N} \Big[\lambda_j \big\| \mathsf{P}_j(f) - \mathsf{S}_{\mathscr{D}}(\gamma_j) \big\|_2^2 + \mu_j \|\gamma_j\|_p^p \Big]$$

with $\theta = \big((\lambda_j, \mu_j), \mathscr{D}\big) \in (\mathbb{R}^2)^N \times X$ and $\mathsf{S}_{\mathscr{D}} \colon \ell^2 \to X$ denotes synthesis operator associated with the dictionary $\mathscr{D}$.

- Alternating minimisation scheme is used to optimise the three variables.

How to find the dictionary?

- Determine jointly while performing reconstruction (joint reconstruction & dictionary learning).
- Specify analytically.
- Determine from example data (dictionary learning).

Dictionary learning

- Setting:
  - $\ell_X \colon X \times X \to \mathbb{R}$ loss function, e.g., 2- or 1-norm.
  - Unsupervised training data $f_1, \dots, f_N \in X$.
  - Dictionary $\mathscr{D} := \{\phi_k\}_k \subset X$.
  - Synthesis operator $\mathsf{S}_{\mathscr{D}} \colon \ell^2 \to X$ given as $\mathsf{S}_{\mathscr{D}}(\gamma) = \sum_k \gamma_k \phi_k$ for $\gamma \in \ell^2$.
- Dictionary learning (sparsity requirement):

$$\begin{cases} \displaystyle\operatorname*{arg\,min}_{\substack{\gamma_i \in \ell^2 \\ \mathscr{D} \subset X}} \sum_{i=1}^{N} \ell_X\bigl(f_i, \mathsf{S}_{\mathscr{D}}(\gamma_i)\bigr) \\ \|\gamma_i\|_0 \leq s \text{ for } i = 1, \dots, N. \end{cases} \qquad \text{for given sparsity level } s.$$

Dictionary learning

- Setting:
  - $\ell_X \colon X \times X \to \mathbb{R}$ loss function, e.g., 2- or 1-norm.
  - Unsupervised training data $f_1, \dots, f_N \in X$.
  - Dictionary $\mathscr{D} := \{\phi_k\}_k \subset X$.
  - Synthesis operator $\mathsf{S}_{\mathscr{D}} \colon \ell^2 \to X$ given as $\mathsf{S}_{\mathscr{D}}(\gamma) = \sum_k \gamma_k \phi_k$ for $\gamma \in \ell^2$.
- Dictionary learning (precision requirement):

$$\begin{cases} \displaystyle\operatorname*{arg\,min}_{\substack{\gamma_i \in \ell^2 \\ \mathscr{D} \subset X}} \sum_{i=1}^{N} \|\gamma_i\|_0 & \text{for given precision } \epsilon > 0. \\ \ell_X\big(f_i, \mathsf{S}_{\mathscr{D}}(\gamma_i)\big) \leq \epsilon \text{ for } i = 1, \dots, N. \end{cases}$$

Objective is total cost for representing signals in training data w.r.t. a dictionary.

Dictionary learning

- Setting:
  - $\ell_X \colon X \times X \to \mathbb{R}$ loss function, e.g., 2- or 1-norm.
  - Unsupervised training data $f_1, \ldots, f_N \in X$.
  - Dictionary $\mathscr{D} := \{\phi_k\}_k \subset X$.
  - Synthesis operator $\mathsf{S}_{\mathscr{D}} \colon \ell^2 \to X$ given as $\mathsf{S}_{\mathscr{D}}(\gamma) = \sum_k \gamma_k \phi_k$ for $\gamma \in \ell^2$.
- Dictionary learning (unified formulation):

$$\operatorname*{arg\,min}_{\substack{\gamma_i \in \ell^2 \\ \mathscr{D} \subset X}} \sum_{i=1}^{N} \Big[\ell_X\big(f_i, \mathsf{S}_{\mathscr{D}}(\gamma_i)\big) + \theta \|\gamma_i\|_0 \Big]$$

Dictionary learning

- Setting:
  - $\ell_X : X \times X \to \mathbb{R}$ loss function, e.g., 2- or 1-norm.
  - Unsupervised training data $f_1, \ldots, f_N \in X$.
  - Dictionary $\mathscr{D} := \{\phi_k\}_k \subset X$.
  - Synthesis operator $\mathsf{S}_{\mathscr{D}} : \ell^2 \to X$ given as $\mathsf{S}_{\mathscr{D}}(\gamma) = \sum_k \gamma_k \phi_k$ for $\gamma \in \ell^2$.
- Dictionary learning (unified formulation):

$$\operatorname*{arg\,min}_{\substack{\gamma_i \in \ell^2 \\ \mathscr{D} \subset X}} \sum_{i=1}^{N} \Big[\ell_X\big(f_i, \mathsf{S}_{\mathscr{D}}(\gamma_i)\big) + \theta \|\gamma_i\|_0 \Big]$$

- All formulations are NP-hard

Dictionary learning

- Setting:
  - $\ell_X \colon X \times X \to \mathbb{R}$ loss function, e.g., 2- or 1-norm.
  - Unsupervised training data $f_1, \dots, f_N \in X$.
  - Dictionary $\mathscr{D} := \{\phi_k\}_k \subset X$.
  - Synthesis operator $\mathsf{S}_{\mathscr{D}} \colon \ell^2 \to X$ given as $\mathsf{S}_{\mathscr{D}}(\gamma) = \sum_k \gamma_k \phi_k$ for $\gamma \in \ell^2$.
- Dictionary learning (unified formulation):

$$\operatorname*{arg\,min}_{\substack{\gamma_i \in \ell^2 \\ \mathscr{D} \subset X}} \sum_{i=1}^{N} \Big[\ell_X\big(f_i, \mathsf{S}_{\mathscr{D}}(\gamma_i)\big) + \theta \|\gamma_i\|_1 \Big]$$

- All formulations are NP-hard $\implies$ Convex relaxation.

Dictionary learning

- Setting:
  - $\ell_X \colon X \times X \to \mathbb{R}$ loss function, e.g., 2- or 1-norm.
  - Unsupervised training data $f_1, \ldots, f_N \in X$.
  - Dictionary $\mathscr{D} := \{\phi_k\}_k \subset X$.
  - Synthesis operator $\mathsf{S}_{\mathscr{D}} \colon \ell^2 \to X$ given as $\mathsf{S}_{\mathscr{D}}(\gamma) = \sum_k \gamma_k \phi_k$ for $\gamma \in \ell^2$.
- Dictionary learning (unified formulation):

$$\operatorname*{arg\,min}_{\substack{\gamma_i \in \ell^2 \\ \mathscr{D} \subset X}} \sum_{i=1}^{N} \Big[\ell_X\big(f_i, \mathsf{S}_{\mathscr{D}}(\gamma_i)\big) + \theta \|\gamma_i\|_0 \Big]$$

- All formulations are NP-hard $\implies$ Convex relaxation.
- Fix $\mathscr{D}$ $\implies$ sum in objective decouples $\implies$ sparse coding:

$$\gamma_i^* := \operatorname*{arg\,min}_{\gamma \in \ell^2} \Big[\ell_X\big(f_i, \mathsf{S}_{\mathscr{D}}(\gamma)\big) + \theta \|\gamma\|_1 \Big] \quad \text{for } i = 1, \ldots, N.$$

- Setting:
  - $X = \mathbb{R}^n$ and ℓ^2 replaced by $\mathbb{R}^m$ for some m.
  - Dictionary: $\mathscr{D} := \{\phi_k\}_{k=1}^m \subset \mathbb{R}^n$ represented by $(n \times m)$-matrix $\mathbf{D}$
    $\implies$ synthesis operator $\mathsf{S}_{\mathbf{D}} \colon \mathbb{R}^m \to \mathbb{R}^n$ with $\mathsf{S}_{\mathbf{D}}(\gamma) = \mathbf{D} \cdot \gamma$.
  - Dictionary size m can be larger than n to exploit redundancy.
    $\mathbf{D}$ is a basis $\implies$ unique solution (good) but limited expressiveness (bad).
    $\mathbf{D}$ overcomplete $\implies$ multiple solutions (bad) but greater expressiveness (good).
- Dictionary learning (unified formulation):

$$\min_{\substack{\gamma_i \in \mathbb{R}^m \\ \mathbf{D} \in \mathbb{R}^{n \times m}}} \sum_{i=1}^{N} \Big[\ell_X\big(f_i, \mathbf{D} \cdot \gamma_i\big) + \theta \|\gamma_i\|_1 \Big].$$

- Setting:
  - $X = \mathbb{R}^n$ and ℓ^2 replaced by $\mathbb{R}^m$ for some m.
  - Dictionary: $\mathscr{D} := \{\phi_k\}_{k=1}^m \subset \mathbb{R}^n$ represented by $(n \times m)$-matrix $\mathbf{D}$
    $\implies$ synthesis operator $\mathsf{S}_{\mathbf{D}} \colon \mathbb{R}^m \to \mathbb{R}^n$ with $\mathsf{S}_{\mathbf{D}}(\gamma) = \mathbf{D} \cdot \gamma$.
- Dictionary learning (unified formulation):

$$\min_{\substack{\gamma_i \in \mathbb{R}^m \\ \mathbf{D} \in \mathbb{R}^{n \times m}}} \sum_{i=1}^{N} \Big[\ell_X\big(f_i, \mathbf{D} \cdot \gamma_i\big) + \theta \|\gamma_i\|_1 \Big].$$

- Simultaneously learn dictionary $\mathbf{D}$ and sparse representation $\mathbf{\Gamma} := [\gamma_1 \dots \gamma_N]$.

Dictionary learning
Finite dimensional setting

- Setting:
  - $X = \mathbb{R}^n$ and ℓ^2 replaced by $\mathbb{R}^m$ for some m.
  - Dictionary: $\mathscr{D} := \{\phi_k\}_{k=1}^m \subset \mathbb{R}^n$ represented by $(n \times m)$-matrix $\mathbf{D}$
    $\implies$ synthesis operator $\mathsf{S}_{\mathbf{D}} \colon \mathbb{R}^m \to \mathbb{R}^n$ with $\mathsf{S}_{\mathbf{D}}(\gamma) = \mathbf{D} \cdot \gamma$.
- Dictionary learning (unified formulation):

$$\min_{\substack{\gamma_i \in \mathbb{R}^m \\ \mathbf{D} \in \mathbb{R}^{n \times m}}} \sum_{i=1}^{N} \Big[\ell_X\big(f_i, \mathbf{D} \cdot \gamma_i\big) + \theta \|\gamma_i\|_1 \Big].$$

- Simultaneously learn dictionary $\mathbf{D}$ and sparse representation $\mathbf{\Gamma} := [\gamma_1 \ldots \gamma_N]$.
- $\mathbf{D}$ satisfies RIP $\implies$ relaxation preserves sparse solution (Candès et al., 2006b).
- Separately convex in $\mathbf{D}$ and $\mathbf{\Gamma}$, but not jointly convex
  $\implies$ intertwined iterates that alternatingly update $\mathbf{D}$ and $\mathbf{\Gamma}$.
- Fixed $\mathbf{D}$ $\implies$ sparse coding problem.

Dictionary learning

Problem: Simultaneously learn dictionary $\mathbf{D}$ and sparse representation γ_i's using L^2-loss:

$$\min_{\substack{\gamma_i \in \mathbb{R}^m \\ \mathbf{D} \in \mathbb{R}^{n \times m}}} \sum_{i=1}^{N} \left[\|f_i - \mathbf{D} \cdot \gamma_i\|_2^2 + \theta \|\gamma_i\|_1 \right].$$

- Intertwined alternated updating of (matrices) $\mathbf{D}$ and $\mathbf{\Gamma} := [\gamma_1 \dots \gamma_N]$.
- State-of-the-art dictionary learning algorithms (Rubinstein et al., 2010):
  - K-SVD (Aharon et al., 2006): Two-stage iterative process.
  - Geometric multi-resolution analysis (GRMA) (Allard et al., 2012).
  - Online dictionary learning (Mairal et al., 2010).
- Most work done in the context of denoising.

Dictionary learning

Problem: Simultaneously learn dictionary $\mathbf{D}$ and sparse representation γ_i's using L^2-loss:

$$\min_{\substack{\gamma_i \in \mathbb{R}^m \\ \mathbf{D} \in \mathbb{R}^{n \times m}}} \sum_{i=1}^{N} \left[\| f_i - \mathbf{D} \cdot \gamma_i \|_2^2 + \theta \| \gamma_i \|_1 \right].$$

K-SVD (Aharon et al., 2006): Two-stage iterative process.

- Sparse coding stage: Solve a sparse coding problem to compute a sparse representation with a priori bound on sparsity.
- Codebook update stage: Sequentially changes dictionary atoms (columns of $\mathbf{D}$) and update relevant γ_i's (coefficients of the sparse representation).

Dictionary learning

Problem: Simultaneously learn dictionary $\mathbf{D}$ and sparse representation γ_i's using L^2-loss:

$$\min_{\substack{\gamma_i \in \mathbb{R}^m \\ \mathbf{D} \in \mathbb{R}^{n \times m}}} \sum_{i=1}^{N} \Big[\| f_i - \mathbf{D} \cdot \gamma_i \|_2^2 + \theta \|\gamma_i\|_1 \Big].$$

Geometric multi-resolution analysis (GRMA) (Allard et al., 2012):

- Training data are noisy samples from a probability distribution on n_0-dimensional manifold $M \subset R^n$ where $n_0 \ll n$.
- Analyse M by techniques from geometric measure theory (Jones, 1990; David & Semmes, 1993) and multi-scale approximation (Binev & DeVore, 2004; Binev et al., 2005).
- Resulting dictionary (geometric wavelets) is structured in a multi-scale fashion with synthesis and analysis operators that can be computed fast.

Dictionary learning

Problem: Simultaneously learn dictionary $\mathbf{D}$ and sparse representation γ_i's using L^2-loss:

$$\min_{\substack{\gamma_i \in \mathbb{R}^m \\ \mathbf{D} \in \mathbb{R}^{n \times m}}} \sum_{i=1}^{N} \left[\|f_i - \mathbf{D} \cdot \gamma_i\|_2^2 + \theta \|\gamma_i\|_1 \right].$$

Online dictionary learning (Mairal et al., 2010):

- Randomly sample the training set.
- Use at each iteration only one sample to update the dictionary.
- Shown to be significantly faster than batch algorithms while achieving similar results.

Convolutional dictionaries

- Issues with the Sparse-Land model:
  1. Performing sparse coding over all the patches tends to be a slow process.

     Can be addressed using learned iterative scheme, e.g., LISTA which learns a finite number of unrolled ISTA iterates using unsupervised training data as to match ISTA solutions (Gregor & LeCun, 2010).

  2. Learning a dictionary over each patch independently cannot account for global information, e.g., shift-invariance in images.

     Need computational feasible approach that introduces further structure and invariances on dictionary, e.g., shift-invariance and making each atom dependent on whole signal instead of patches.

Convolutional dictionaries

- Issues with the Sparse-Land model:
  1. Performing sparse coding over all the patches tends to be a slow process.

     Can be addressed using learned iterative scheme, e.g., LISTA which learns a finite number of unrolled ISTA iterates using unsupervised training data as to match ISTA solutions (Gregor & LeCun, 2010).

  2. Learning a dictionary over each patch independently cannot account for global information, e.g., shift-invariance in images.

     Need computational feasible approach that introduces further structure and invariances on dictionary, e.g., shift-invariance and making each atom dependent on whole signal instead of patches.
- Convolutional dictionaries: Atoms given by convolutional kernels and act on signal features by convolutions, i.e., $\mathbf{D}$ is a concatenation of Toeplitz matrices (union of banded and circulant matrices).

  $\Longrightarrow$ Computationally feasible shift-invariant dictionary where atoms depend on entire signal.

- Inverse problem: Recover $f^* \in X$ from $g = \mathcal{A}(f^*) + \delta g$.
- Data log likelihood: $\mathcal{L}\colon Y \times Y \to \mathbb{R}$.
- Prior model: $f^* \in X$ is compressible w.r.t. convolution dictionary $\mathscr{D} := \{\phi_i\}_i \subset X$.
- Convolutional Sparse Coding (CSC): Sparse coding (synthesis) using convolutional dictionaries (atoms act by convolutions):

$$\mathcal{A}_\theta^\dagger(g) := \sum_i \gamma_i^* * \phi_i \quad \text{where} \quad \gamma_i^* \in \operatorname*{arg\,min}_{\gamma_i \in X} \Big[\mathcal{L}\Big(\mathcal{A}\Big(\sum_i \gamma_i * \phi_i\Big), g\Big) + \theta \sum_i \|\gamma\|_0 \Big].$$

- Methods for denoising by CSC use convex relaxation followed by ADMM in frequency space (Bristow et al., 2013), along with variants of it. See als (Sreter & Giryes, 2017) for using LISTA in this context.
- Analysed in the context of denoising (Bristow et al., 2013; Wohlberg, 2014; Gu et al., 2015; Papyan et al., 2016b, 2016a; Garcia-Cardona & Wohlberg, 2017).
- Theoretical properties for denoising analysed in (Papyan et al., 2016b, 2016a).

- Unsupervised training data: $f_1, \dots, f_N \in X$.
- Loss function $\ell_X \colon X \to X$.
- Convolutional dictionary learning:

$$\min_{\phi_i, \gamma_{j,i} \in X} \left[\sum_{j=1}^{N} \ell_X\Big(f_j, \sum_i \gamma_{j,i} * \phi_i \Big) + \theta \sum_{j=1}^{N} \sum_i \|\gamma_{j,i}\|_1 \right] \quad \text{and } \|\phi_i\|_2 = 1.$$

- Convex relaxation and L^2-loss: Solved using ADMM type of scheme (Garcia-Cardona & Wohlberg, 2017).
- Extension to supervised data setting: Learn discriminative dictionaries instead of purely reconstructive ones by introducing a supervised regularisation term into the usual CSC objective that encourages the final dictionary elements to be discriminative (Affara et al., 2018).

Multi-Layer Convolutional Sparse Model

Multi-Layer Convolutional Sparse Model (ML-CSC) (Sulam et al., 2017)

- L convolution dictionaries $\mathscr{D}_1, \ldots, \mathscr{D}_L \subset X$.
- Prior model:
  - $f^* \in X$ is compressible w.r.t. convolution dictionary $\mathscr{D}_1 := \{\phi_{1,i}\}_i \subset X$.
  - Atoms $\phi_{k,i} \in \mathscr{D}_k$ are compressible w.r.t. convolution dictionary $\mathscr{D}_{k+1}$ for $k = 1, \ldots, L-1$.
- Special case of a Convolutional Sparse Model where intermediate representations have specific structure (Sulam et al., 2017, Lemma 1).
- Building on theory for CSC, (Sulam et al., 2017) provides a theoretical study of this novel model and its associated pursuits for dictionary learning and sparse coding (for denoising).
  $\implies$ Layered thresholding algorithm and the layered basis pursuit which share many similarities with deep CNNs.

Theoretical analysis of ML-CSC (Papyan et al., 2017).

- ML-CSC yields a Bayesian model implicitly imposed on f^* when deploying a CNN $\implies$ Characterise signals belonging to the model behind a deep CNN.
- Does not assume any specific property of network's parameters (apart from broad coherence).
  - (Bruna & Mallat, 2013; S. Mallat, 2016) assumes filters are Wavelets.
  - (Giryes et al., 2015) assumes random weights.

Theoretical analysis of ML-CSC (Papyan et al., 2017).

- Deep CNN $\iff$ layered thresholding algorithm.
- Offers a mathematical analysis of the CNN architecture:
  - Theorem 4: The CNN is guaranteed to recover an estimate of the underlying representations of an input signal, assuming these are sparse in a local sense.
  - Theorem 8 & 10: Adding norm-bounded noise to the signal results in a bounded perturbation in the output $\implies$ stability of the CNN in recovering the underlying representations.
- ML-CSC can be used to propose an alternative to the commonly used forward pass algorithm in CNN. This is related to both deconvolutional (Zeiler et al., 2010; Pu et al., 2016) and recurrent networks (Bengio et al., 1994).
- Many of the results also hold for fully connected networks.

Deep Dictionary Learning

- Two popular representation learning paradigms: Dictionary learning and deep learning.
  - Dictionary learning focuses on learning ‘basis’ and ‘features’ by matrix factorisation.
  - Deep learning focuses on extracting features via learning ‘weights’ or ‘filter’ in a greedy layer by layer fashion.
- Deep dictionary learning: Deeper architectures are built using the layers of dictionary learning (Tariyal et al., 2016).
- Competitive against other deep learning approaches, such as stacked autoencoder, deep belief network, and convolutional neural network, regarding classification and clustering accuracies.

Other

Prior with black box denoiser

- There is an abundance of high-performing image denoising algorithms, would be nice to integrate these in reconstrruction.
- Design regularisation that can incorporate a black box denoiser?
- Plug-and-Play Prior (P^3) method (Venkatakrishnan et al., 2013)
  - Implicit prior for regularising general inverse problems.
  - Based on using the ADMM optimisation scheme: The overall problem decomposes into a sequence of image denoising tasks, coupled with simpler L^2-regularised inverse problems that are much easier to handle.
  - Regularisation only implicitly implied by the denoising algorithm $\implies$ No clear definition of the objective function (unclear if there is an underlying objective function).
  - Stability issues, parameter tuning of the ADMM scheme is a delicate matter.
  - Intimately coupled with the ADMM algorithm, does not offer easy and flexible ways of replacing the iterative procedure.

Prior with black box denoiser

- There is an abundance of high-performing image denoising algorithms, would be nice to integrate these in reconstrruction.
- Design regularisation that can incorporate a black box denoiser?
- Regularisation by Denoising (RED) (Romano et al., 2017)
  - Reconstruction method:

$$\min_{f \in X} \Big[\mathcal{L}\big(\mathcal{A}(f), g\big) + \mathcal{R}_\theta(f) \Big] \quad \text{with} \quad \mathcal{R}_\theta(f) := \theta \big\langle f, f - \Lambda(f) \big\rangle$$

    and where $\Lambda \colon X \to X$ is denoiser.
  - Denoiser needs to fulfil some weak conditions (local homogeneity and strong passivity). $\implies$ Can efficiently compute gradient of denoiser
  - Many Denoising algorithms, such as the NLM, kernel-regression, K-SVD, fulfil necessary assumptions.
  - More general, simpler and stabler alternative to the P^3 method.

References I

Affara, L., Ghanem, B., & Wonka, P. (2018). Supervised convolutional sparse coding. *arXiv*, *1804.02678*.

Aharon, M., Elad, M., & Bruckstein, A. M. (2006). K-SVD: An algorithm for designing of over-complete dictionaries for sparse representation. *IEEE Transactions in Signal Processing*, *54*(11), 4311–4322.

Allard, W. K., Chen, G., & Maggioni, M. (2012). Multi-scale geometric methods for data sets II: Geometric multi-resolution analysis. *Applied Computational and Harmonic Analysis*, *32*(3), 435–462.

Bai, T., Yan, H., Jia, X., Jiang, S., Wang, G., & Mou, X. (2017). Volumetric computed tomography reconstruction with dictionary learning. In *14th International Meeting on Fully Three-Dimensional Image Reconstruction in Radiology and Nuclear Medicine (Fully3D 2017)*.

References II

Bakushinskii, A. B. (1984). Remarks on choosing a regularization parameter using the quasi-optimality and ratio criterion. *USSR Computational Mathematics and Mathematical Physics*, *24*, 181–182.

Bakushinsky, A. B., & Kokurin, M. Y. (2004). *Iterative methods for approximate solution of inverse problems*. Springer Verlag.

Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, *5*(2), 157–166.

Berinde, R., Gilbert, A. C., Indyk, P., Karloff, H., & Strauss, M. J. (2008). Combining geometry and combinatorics: A unified approach to sparse signal recovery. In *2008 46th Annual Allerton Conference on Communication, Control, and Computing* (pp. 798–805).

Bertero, M., & Boccacci, P. (1998). *Introduction to inverse problems in imaging*. Institute of Physics Publishing.

References III

Bertero, M., Lantéri, H., & Zanni, L. (2008). Iterative image reconstruction: a point of view. In Y. Censor, M. Jiang, & A. K. Louis (Eds.), *Proceedings of the interdisciplinary workshop on mathematical methods in biomedical imaging and intensity-modulated radiation (imrt), pisa, italy* (pp. 37–63).

Binev, P., Cohen, A., Dahmen, W., DeVore, R., & Temlyakov, V. (2005). Universal algorithms for learning theory part I: Piecewise constant functions. *Journal of Machine Learning Research*, *6*, 1297–1321.

Binev, P., & DeVore, R. (2004). Fast computation in adaptive tree approximation. *Numerische Mathematik*, *19*(2), 193–217.

Blumensath, T., & Davies, M. E. (2008). Iterative thresholding for sparse approximations. *Journal of Fourier Analysis and Applications*, *14*(5–6), 629–654.

Bredies, K., Kunisch, K., & Pock, T. (2010). Total generalized variation. *SIAM Journal on Imaging Sciences*, *3*(3), 492–526.

References IV

Bristow, H., Eriksson, A., & Lucey, S. (2013). Fast convolutional sparse coding. In *2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 391–398).

Bruckstein, A. M., Donoho, D. L., & Elad, M. (2009). From sparse solutions of systems of equations to sparse modeling of signals and images. *SIAM Review*, *51*(1), 34–18.

Bruna, J., & Mallat, S. (2013). Invariant scattering convolution networks. *IEEE transactions on pattern analysis and machine intelligence*, *35*(8), 1872–1886.

Bubeck, S. (2015). Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, *8*(3–4), 231–357.

Burger, H. C., Schuler, C. J., & Harmeling, S. (2012). Image denoising: Can plain neural networks compete with BM3D? In *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2392–2399).

References V

Burger, M., Kaltenbacher, B., & Neubauer, A. (2015). Iterative solution methods. In *Handbook of mathematical methods in imaging* (2nd ed., pp. 431–470). Springer Verlag.

Byrne, C. L. (2008). *Applied iterative methods*. A K Peters-CRC Press.

Byrne, C. L., & Eggermont, P. P. B. (2015). EM algorithms. In *Handbook of mathematical methods in imaging* (2nd ed., pp. 305–388). Springer Verlag.

Candès, E. L., Romberg, J. K., & Tao, T. (2006a). Robust uncertainty principles: Exact signal reconstruction from highly incomplete Fourier information. *IEEE Transactions on Information Theory*, *52*(2), 489–509.

Candès, E. L., Romberg, J. K., & Tao, T. (2006b). Stable signal recovery from incomplete and inaccurate measurements. *Communications in Pure and Applied Mathematics*, *59*(8), 1207–1223.

References VI

Caselles, V., Chambolle, A., & Novaga, M. (2015). Total variation in imaging. In *Handbook of mathematical methods in imaging* (2nd ed., pp. 1455–1499). Springer Verlag.

Chen, G., & Needell, D. (2016). Compressed sensing and dictionary learning. *Proceedings of Symposia in Applied Mathematics*, *73*, 201–241.

Chen, H., Zhang, Y., Chen, Y., Zhang, J., Zhang, W., Sun, H., . . . Wang, G. (2018). LEARN: learned experts' assessment-based reconstruction network for sparse-data CT. *IEEE Transactions on Medical Imaging*. (Accepted for publication)

Chen, Y., Ranftl, R., & Pock, T. (2014). A bi-level view of inpainting - based image compression. *arXiv*, *1401.4112v2*.

References VII

Chun, I. Y., Zheng, X., Long, Y., & Fessler, J. A. (2017). Sparse-view x-ray CT reconstruction using ℓ_1 regularization with learned sparsifying transform. In *14th International Meeting on Fully Three-Dimensional Image Reconstruction in Radiology and Nuclear Medicine (Fully3D 2017).*

Cohen, A., Dahmen, W., & DeVore, R. (2009). Compressed sensing and best *k*-term approximation. *Journal of the American Mathematical Society*, *22*(1), 211–231.

Dabov, K., Foi, A., Katkovnik, V., & Egiazarian, K. (2007). Image denoising by sparse 3-D transform-domain collaborative filtering. *IEEE Transactions on Image Processing*, *16*(8), 2080–2095.

Daubechies, I., Defrise, M., & De Mol, C. (2004). An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications in pure and applied mathematics*, *57*(11), 1413–1457.

David, G., & Semmes, S. (1993). *Analysis of and on uniformly rectifiable sets* (Vol. 38). Providence, RI: American Mathematical Society.

References VIII

De los Reyes, J. C., Schönlieb, C.-B., & Valkonen, T. (2017). Bilevel parameter learning for higher-order total variation regularisation models. *Journal of Mathematical Imaging and Vision*, *57*(1), 1–25.

Dong, W., Shi, G., Ma, Y., & Li, X. (2015). Image restoration via simultaneous sparse coding: Where structured sparsity meets Gaussian scale mixture. *International Journal of Computer Vision*, *114*(2–3), 217–232.

Dong, W., Zhang, L., Shi, G., & Li, X. (2013). Nonlocally centralized sparse representation for image restoration. *IEEE Transactions on Image Processing*, *22*(4), 1620–1630.

Dong, W., Zhang, L., Shi, G., & Wu, X. (2011). Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization. *IEEE Transactions on Image Processing*, *20*(7), 1838–1857.

Donoho, D. L., Tsaig, Y., Drori, I., & Starck, J.-L. (2012). Sparse solution of underdetermined systems of linear equations by stagewise orthogonal matching pursuit. *IEEE Transactions on Information Theory*, *58*(2), 1094–1121.

Elad, M. (2010). *Sparse and redundant representations: From theory to applications in signal and image processing*. Springer Verlag.

Elad, M., & Aharon, M. (2006a). Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image Processing*, *15*(12), 3736–3745.

Elad, M., & Aharon, M. (2006b). Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image Processing*, *15*(12), 3736–3745.

Engl, H. W., Hanke, M., & Neubauer, A. (2000). *Regularization of inverse problems* (No. 375). Kluwer Academic Publishers.

References X

Figueiredo, M. A. T., Nowak, R. D., & Wright, S. J. (2007). Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems. *IEEE Journal of Selected Topics in Signal Processing: Special Issue on Convex Optimization Methods for Signal Processing*, *1*(4), 586–598.

Fornasier, M., & Rauhut, H. (2008). Iterative thresholding algorithms. *Applied and Computational Harmonic Analysis*, *25*(2), 187–208.

Foucart, S. (2016). Dictionary-sparse recovery via thresholding-based algorithms. *Journal of Functional Analsysis and Its Applications*, *22*, 6–19.

Foucart, S., & Rauhut, H. (2013). *A mathematical introduction to compressive sensing*. Springer Verlag.

Garcia-Cardona, C., & Wohlberg, B. (2017). Convolutional dictionary learning. *arXiv*, *1709.02893*.

References XI

Giryes, R., Sapiro, G., & Bronstein, A. M. (2015). Deep neural networks with random gaussian weights: A universal classification strategy? *IEEE Transactions on Signal Processing*, *64*(13), 3444–3457.

Golub, G. H., Heat, M., & Wahba, G. (1979). Generalized cross validation as a method for choosing a good ridge parameter. *Technometrics*, *21*, 215–223.

Gregor, K., & LeCun, Y. (2010). Learning fast approximations of sparse coding. In *Proceedings of the 27th International Conference on International Conference on Machine Learning (ICML'10)* (pp. 399–406).

Gu, S., Zhang, L., Zuo, W., & Feng, X. (2014). Weighted nuclear norm minimization with application to image denoising. In *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2862–2869).

Gu, S., Zuo, W., Xie, Q., Meng, D., Feng, X., & Zhang, L. (2015). Convolutional sparse coding for image super-resolution. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 1823–1831).

References XII

Haber, E., & Tenorio, L. (2003). Learning regularization functionals. *Inverse Problems*, *19*, 611–626.

Hammernik, K., Klatzer, E., T. ans Kobler, Recht, M. P., Sodickson, D. K., Pock, T., & Knoll, F. (2018). Learning a variational network for reconstruction of accelerated MRI data. *Magnetic Resonance in Medicine*, *79*(6), 3055–3071.

Hansen, P. C. (1992). Analysis of discrete ill-posed problems by means of the L-curve. *SIAM Review*, *34*, 561–580.

Hansen, P.-C. (1997). *Rank-deficient and discrete ill-posed problems: Numerical aspects of linear inversion* (Vol. 4). SIAM.

Jones, P. W. (1990). Rectifiable sets and the traveling salesman problem. *Inventiones Mathematicae*, *102*(1), 1–15.

Kaltenbacher, B., Neubauer, A., & Scherzer, O. (2008). *Iterative regularization methods for nonlinear ill-posed problems* (Vol. 6). Walter de Gruyter.

References XIII

Kim, S.-J., Koh, K., Lustig, M., Boyd, S., & Gorinevsky, D. (2007). An interior-point method for large-scale ℓ_1-regularized least squares. *IEEE Journal of Selected Topics in Signal Processing*, *1*(4), 606–617.

Kobler, E., Klatzer, T., Hammernik, K., & Pock, T. (2017). Variational networks: connecting variational methods and deep learning. In *Proceedings of the German Conference on Pattern Recognition (GCPR)* (pp. 281–293).

Krishnan, V. P., & Quinto, E. T. (2015). Microlocal analysis in tomography. In *Handbook of mathematical methods in imaging* (2nd ed., pp. 847–902). Springer Verlag.

Kunisch, K., & Pock, T. (2013). A bilevel optimization approach for parameter learning in variational models. *SIAM Journal on Imaging Sciences*, *6*(2), 938–983.

Lanusse, F., Starck, J.-L., Woiselle, A., & Fadili, J. M. (2014). 3-D sparse representations. *Advances in Imaging and Electron Physics*, *183*, 99–204.

Louis, A. K. (1996). Approximate inverse for linear and some nonlinear problems. *Inverse Problems*, *12*(2), 175–190.

Mairal, J., Bach, F., Ponce, J., & Sapiro, G. (2010). Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, *11*, 19–60.

Mairal, J., & Ponce, J. (2014). Sparse modeling for image and vision processing. *arXiv*, *1411.3230v2*.

Mallat, S. (2016). Understanding deep convolutional networks. *Philosophical Transactions of the Royal Society of London A: athematical, Physical & Engineering Sciences*, *374*, 20150203.

Mallat, S. G., & Zhang, Z. (1993). Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, *41*(12), 3397–3415.

Miller, K. (1970). Least squares methods for ill-posed problems with a prescribed bound. *SIAM Journal on Mathematical Analysis*, *1*(1), 52–74.

References XV

Morozov, V. A. (1966). On the solution of functional equations by the method of regularization. *Soviet mathematics – Doklady*, *7*, 414–417.

Natterer, F., & Wübbeling, F. (2001). *Mathematical methods in image reconstruction* (Vol. 5). SIAM.

Needell, D., & Tropp, J. A. (2009). CoSaMP: iterative signal recovery from incomplete and inaccurate samples. *Applied Computational and Harmonic Analysis*, *26*(3), 301–321.

Needell, D., & Vershynin, R. (2009). Uniform uncertainty principle and signal recovery via regularized orthogonal matching pursuit. *Foundations of Computational Mathematics*, *9*(3), 317–334.

Papyan, V., Romano, Y., & Elad, M. (2017). Convolutional neural networks analyzed via convolutional sparse coding. *Journal of Machine Learning Research*, *18*, 1–52.

References XVI

Papyan, V., Sulam, J., & Elad, M. (2016a). Working locally thinking globally - part II: Stability and algorithms for convolutional sparse coding. *arXiv*, *1607.02009*.

Papyan, V., Sulam, J., & Elad, M. (2016b). Working locally thinking globally - part I: Theoretical guarantees for convolutional sparse coding. *arXiv*, *1607.02005*.

Peyré, G., & Fadili, J. (2011). Learning analysis sparsity priors. In *Proceedings of Sampta'11.* Retrieved from http://hal.archives-ouvertes.fr/hal-00542016/

Pu, Y., Yuan, X., Stevens, A., Li, C., & Carin, L. (2016). A deep generative deconvolutional image model. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics* (pp. 741–750).

Quinto, E. T. (1993). Singularities of the X-ray transform and limited data tomography in $\mathbb{R}^2$ and $\mathbb{R}^3$. *SIAM Journal on Mathematical Analysis*, *24*, 1215-1225.

References XVII

Romano, Y., & Elad, M. (2015). Patch-disagreement as a way to improve K-SVD denoising. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1280–1284).

Romano, Y., Elad, M., & Milanfar, P. (2017). The little engine that could: Regularization by denoising (RED). *SIAM Journal on Imaging Sciences*, *10*(4), 1804–1844.

Roth, S., & Black, M. J. (2009). Fields of experts. *International Journal of Computer Vision*, *82*, 205–229.

Rubinstein, R., Bruckstein, A. M., & Elad., M. (2010). Dictionaries for sparse representation modeling. *Proceedings of the IEEE*, *98*(1045–1057).

Samuel, K. G. G., & Tappen, M. F. (2009). Learning optimized MAP estimates in continuously-valued MRF models. In *IEEE Conference on Computer Vision and Pattern Recognition, 2009. CVPR 2009* (pp. 477–484).

References XVIII

Scherzer, O., Grasmair, M., Grossauer, H., Haltmeier, M., & Lenzen, F. (2009). *Variational Methods in Imaging* (Vol. 167). New York: Springer-Verlag.

Schuster, T. (2007). *The Method of Approximate Inverse: Theory and Applications* (Vol. 1906). Heidelberg: Springer Verlag.

Sreter, H., & Giryes, R. (2017). Learned convolutional sparse coding. *arXiv*, *1711.00328*.

Sulam, J., & Elad, M. (2015). Expected patch log likelihood with a sparse prior. In *Energy Minimization Methods in Computer Vision and Pattern Recognition: Proceedings of the 10th International Conference, EMMCVPR 2015* (pp. 99–111).

Sulam, J., Papyan, V., Romano, Y., & Elad, M. (2017). Multi-layer convolutional sparse modeling: Pursuit and dictionary learning. *arXiv*, *1708.08705*.

Tariyal, S., Majumdar, A., Singh, R., & Vatsa, M. (2016). Deep dictionary learning. *IEEE Access*, *4*(10096–10109).

References XIX

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267–288.

Tropp, J. A., & Gilbert, A. C. (2007). Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on Information Theory*, *53*(12), 4655–4666.

Venkatakrishnan, S. V., Bouman, C., & Wohlberg, B. (2013). Plug-and-play priors for model based reconstruction. In *2013 IEEE Global Conference on Signal and Information Processing (GlobalSIP)* (pp. 945–948).

Wohlberg, B. (2014). Efficient convolutional sparse coding. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 7173–7177).

Xu, Q., Yu, H., Mou, X., Zhang, L., Hsieh, J., & Wang, G. (2012). Low-dose x-ray CT reconstruction via dictionary learning. *IEEE Transactions in Medical Imaging*, *31*(9), 1682–1697.

References XX

Zeiler, M. D., Krishnan, D., Taylor, G. W., & Fergus, R. (2010). Deconvolutional networks. In *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2528–2535).

Zoran, D., & Weiss, Y. (2011). From learning models of natural image patches to whole image restoration. In *2011 IEEE International Conference on Computer Vision (ICCV)* (pp. 479–486).