The faster proximal algorithm, the better unfolded deep learning architecture ? The study case of image denoising

Hoang Trieu Vy Le, Nelly Pustelnik, Marion Foare

The faster proximal algorithm, the better unfolded deep learning architecture ? The study case of image denoising.

Hoang Trieu Vy Le[(1)], Nelly Pustelnik[(1,2)], Marion Foare[(3,4)]

[(1)] Univ Lyon, Ens de Lyon, Univ Claude Bernard, CNRS, Laboratoire de Physique, Lyon, France.
[(2)] ISPGroup/ICTEAM, UCLouvain, Belgium.
[(3)] Univ Lyon, Ens de Lyon, Univ Lyon 1, CNRS, INRIA, LIP, Lyon, France.
[(4)] CPE Lyon,Villeurbanne, France.

***Abstract*—Deep learning has revolutionized many image processing tasks such as classification or segmentation and, more recently, gives very promising results for solving inverse problems. It however remains a gap between the deep learning black-box approaches and the more recently unrolled deep learning techniques proposed to bring the physics of the model and standard solving techniques into the network design. In order to understand more precisely the mechanisms, we place ourselves in the framework of the simple study of image denoising and we study four networks designed from unrolled forward-backward iterations in the dual, FISTA in the dual, Chambolle-Pock, and Chambolle-Pock exploiting the strong convexity. Performance and stability obtained with each of these networks will be detailed. A comparison within these approaches, standard penalized likelihood approaches, and the state-of-the-art black-box approach DnCNN is also provided.**

***Index Terms*—Image denoising, unfolded proximal algorithms, accelerated methods, deep learning**

I. Introduction

Deep learning has renewed many fields of image processing tasks such as classification, segmentation and, more recently, inverse problems solving. For several years, there was a gap between standard image processing and neural network procedures as the first one was guided by the physics of the data acquisition and prior knowledge about the object to analyze or recover while the second one was considered as a very efficient prior-free black-box procedure.

Recent works make the bridge between standard image processing techniques and deep neural networks architectures helping in the understanding and the analysis of this complex highly nonlinear tool.

On the one hand, several works have established relations between deep strategies and wavelet transform literature, thus inheriting wavelet properties. Wavelet scattering transforms that cascade wavelet transform with nonlinear modulus and averaging operators, benefit from a mathematical framework explaining important properties of deep convolution networks [1], [2] while in [3] the authors prove that affine systems, including several wavelets or frame transforms, can be approximated by deep neural networks with minimal connectivity and memory requirements.

We can also refer to the link between neural networks and nonlocal denoising filtering techniques (e.g., nonlocal means [4] or BM3D [5]) through their neural tangent kernel [6], allowing to avoid the costly training step.

On the other hand, the deep unfolded approaches (cf. LISTA for the pioneering work [7]) have helped connect penalized likelihood approaches to neural networks. The two strategies feed off each other providing then a better understanding of neural network architecture and a formal framework toward the design of robust neural networks based on explicit links between activation functions and proximity operators [8].

Focus of unfolded architectures – The unfolded deep architectures rely both on an objective function to minimize (e.g., F) and an algorithmic procedure adapted to the properties of the objective function. For almost any given image processing problem, there exists a massive literature dedicated to the choice of the objective function F and to algorithmic schemes to design a sequence of iterates $(\mathrm{x}_k)_{k\in\mathbb{N}}$ that converges efficiently to a minimizer $\widehat{\mathrm{x}}$ of F. A standard criterion to measure the efficiency of an algorithm is the convergence rate. For instance, handling with strong convexity property into Chambolle-Pock iteration leads to $F(\mathrm{x}_k)-F(\widehat{\mathrm{x}}) \leq \zeta/k^2$ while the standard iterations are reduced to $F(\mathrm{x}_k) - F(\widehat{\mathrm{x}}) \leq \zeta/k$ where ζ is a constant [9].

However, to the best of our knowledge, when considering unfolded architectures, the impact of accelerated schemes on learning performance has not been studied yet and is the core of our contribution. To better understand the mechanisms with or without employing accelerated schemes and also to facilitate the presentation of the proposed unfolded networks, we handle this question in the context of image denoising focusing on two standard algorithms of the image processing literature: Iterative Soft Thresholding Algorithm (ISTA) [10]–[12] and Chambolle-Pock (CP) [9].

Contributions – The contributions of this work are: *(i)* the design of end-to-end trainable unfolded deep architectures

This work is supported by the ANR (Agence Nationale de la Recherche) from France ANR-19-CE48-0009 Multisc'In and Labex MI-LYON / ANR-10-LABX-0070. We are also grateful for the supports by CBP (Centre Blaise Pascal-ENS lyon) who provide necessary equipments for our numerical experiments.

for image denoising based on two standard iterative algorithms (ISTA and CP) and their associated accelerated version (FISTA [13], [14] and CP with strong convexity assumption [9, Algorithm 2]), *(ii)* the formulation of `Deep-(F)ISTA-GD` and `Deep-(Sc)CP-GD` as two neural networks expressed as the combination of weight and activation functions with closed-form expressions. Each of them includes an extra parameter allowing to consider the basic algorithm or its accelerated version, *(iii)* a study of the performance and robustness of the proposed networks and comparison to standard DnCNN regarding the learning cost and the stability.

The remainder of this paper is organized as follows. Section II provides a brief recall of the considered iterative algorithms and introduces the unfolding procedure of these algorithms. The robustness of the proposed networks is also discussed. Section III is dedicated to the numerical experiments and comparisons with state-of-the-art methods.

II. Unfolded Deep learning architecture

A. Image denoising and associated objective function

Our study focusses on the standard image denoising task, which consists to recover the original image $\overline{\mathrm{x}} \in \mathbb{R}^N$ from its degraded version $\mathrm{z} = \overline{\mathrm{x}} + \mathrm{n} \in \mathbb{R}^N$, where $\mathrm{n} \sim \mathcal{N}(0, \delta^2 \mathbb{I}_N)$ denotes an additive white Gaussian noise with a standard deviation $\delta > 0$. Thus, considering a prior imposing sparsity after some linear transform, and according to Bayesian interpretation, maximizing the a posteriori distribution reads:

$$\widehat{\mathrm{x}} = \underset{\mathrm{x}\in\mathbb{R}^N}{\operatorname{argmin}}\ F(\mathrm{x}) := \frac{1}{2}\|\mathrm{x}-\mathrm{z}\|_2^2 + g(\mathrm{Dx}), \tag{1}$$

where $\mathrm{D} \in \mathbb{R}^{|\mathbb{F}|\times N}$ denotes a linear operator that converts the image in $\mathbb{R}^N$ to a feature space $\mathbb{R}^{|\mathbb{F}|}$ with $|\mathbb{F}| > N$, and g denotes a proper, lower-semicontinous, convex function from $\mathbb{R}^{|\mathbb{F}|}$ to $(-\infty, +\infty]$. A standard choice for g is a ℓ_1-norm or a hybrid $\ell_{1,2}$-norm to favor the sparsity of the features. In the standard penalized likelihood approaches, a regularization parameter appears in front of g and the variance of the noise can be added in front of the data fidelity-term. In this work, these parameters are merged within D.

B. Proximal algorithms: (F)ISTA and (Sc)CP

Despite the simplicity of the minimization problem (1), its resolution requires to handle with iterative schemes as no closed form expression exists in a general framework (see a contrario when D is orthonormal, e.g., a wavelet transform). In this work, we focus on two algorithmic schemes: ISTA and CP, as both offer accelerated procedures: FISTA or CP involving strongly convexity (ScCP).

(F)ISTA in the dual – (F)ISTA requires to handle the dual formulation of problem (1) to have closed form expression steps, leading to $\widehat{\mathrm{x}} = \mathrm{z} - \mathrm{D}^\top \widehat{\mathrm{u}}$ where [1]:

$$\widehat{\mathrm{u}} \in \underset{\mathrm{u}\in\mathbb{R}^{|\mathbb{F}|}}{\operatorname{Argmin}}\ \widetilde{F}(\mathrm{u}) := \frac{1}{2}\|\mathrm{D}^\top \mathrm{u} - \mathrm{z}\|_2^2 + g^*(\mathrm{u}) \tag{2}$$

[1]The constant terms depending solely on z as been removed.

with g^* the Fenchel conjugate of g (e.g. $g = \|\cdot\|_1$ then $g^* = \iota_{\|\cdot\|_\infty \le 1}$ the indicator function of the ℓ_∞-ball). The resulting FISTA iterations [13] [14] read, for every iteration k,

$$\left\lfloor \begin{array}{ll} \mathrm{u}_{k+1} & = \operatorname{prox}_{\tau_k g^*}\Big((\mathrm{Id} - \tau \mathrm{D}\mathrm{D}^\top)\mathrm{y}_k + \tau \mathrm{D}\mathrm{z}\Big) \\ \mathrm{y}_{k+1} & = (1+\alpha_k)\mathrm{u}_{k+1} - \alpha_k \mathrm{u}_k \end{array} \right. \tag{3}$$

where $\mathrm{u}_1 \in \mathbb{R}^{|\mathbb{F}|}$, and $\mathrm{y}_1 \in \mathbb{R}^{|\mathbb{F}|}$. The sequence $(\mathrm{u}_k)_{k\in\mathbb{N}}$ converges to $\widehat{\mathrm{u}}$ when $\alpha_k = \frac{t_k - 1}{t_{k+1}}$ and $t_{k+1} = \frac{k+a-1}{a}$, $a > 2$, $\tau < \frac{1}{\|\mathrm{D}\|^2}$ and $\widetilde{F}(\mathrm{u}_k) - \widetilde{F}(\widehat{\mathrm{u}}) \le \frac{\zeta}{k^2}$. When $\alpha_k \equiv 0$, these iterations reduce to ISTA. The convergence of the iterates is proved when $\tau < \frac{2}{\|\mathrm{D}\|^2}$ for this limit case, and $\widetilde{F}(\mathrm{u}_k) - \widetilde{F}(\widehat{\mathrm{u}}) \le \frac{\zeta}{k}$.

(Sc)CP – CP iterations can be directly applied to the minimization problem (1). The data-term being strongly convex of parameter $\gamma = 1$, the accelerated CP (ScCP: for Strongly convex CP) [9, Algorithm 2] can be employed, leading to, for every iteration k,

$$\left\lfloor \begin{array}{ll} \mathrm{u}_{k+1} & = \operatorname{prox}_{\tau_k g^*}\Big(\mathrm{u}_k + \tau_k \mathrm{D}\Big((1+\alpha_k)\,\mathrm{x}_k - \alpha_k \mathrm{x}_{k-1}\Big)\Big) \\ \mathrm{x}_{k+1} & = \frac{\sigma_k}{1+\sigma_k}\mathrm{z} + \frac{1}{1+\sigma_k}\mathrm{x}_k - \frac{\sigma_k}{1+\sigma_k}\mathrm{D}^\top \mathrm{u}_{k+1} \end{array} \right. \tag{4}$$

where $\alpha_k = \frac{1}{\sqrt{1+2\gamma\sigma_k}}$, $\sigma_{k+1} = \alpha_k \sigma_k$, $\tau_{k+1} = \frac{\tau_k}{\alpha_k}$. The update of x_{k+1} comes from the proximity operator expression of the ℓ_2-norm square: $\operatorname{prox}_{\frac{\sigma_k}{2}\|\cdot-\mathrm{z}\|_2^2}\Big(\mathrm{x}_k - \sigma_k \mathrm{D}^\top \mathrm{u}_{k+1}\Big) = \frac{\sigma_k \mathrm{z} + \mathrm{x}_k - \sigma_k \mathrm{D}^\top \mathrm{u}_{k+1}}{1+\sigma_k}$. This general framework fits both the ScCP iterations [9, Algorithm 2] and the standard CP iterations [9, Algorithm 1] when $\gamma = 0$, $\sigma_k \equiv \sigma$, $\tau_k \equiv \tau$ and assuming $\sigma\tau\|D\|^2 < 1$. The sequence $(\mathrm{x}_k)_{k\in\mathbb{N}}$ converges to the minimizer of (1) and convergence rates have already been provided in the introduction.

C. Unfolded architecture for (F)ISTA and (Sc)CP

From a training set $\mathcal{S} = \{(\overline{\mathrm{x}}_s, \mathrm{z}_s) | s = 1, \dots, I\}$ where $\overline{\mathrm{x}}_s$ denotes a clean image, and z_s the associated noisy one, the goal of a deep denoiser $f_{\widehat{\Theta}}$ is to learn parameters $\widehat{\Theta}$, to minimize the following standard empirical loss:

$$\widehat{\Theta} \in \underset{\Theta}{\operatorname{Argmin}}\ E(\Theta) := \frac{1}{I}\sum_{s=1}^{I} \|\overline{\mathrm{x}}_s - f_\Theta(\mathrm{z}_s)\|_2^2. \tag{5}$$

Standard networks with K layers can be written as, $\forall \mathrm{z} \in \mathbb{R}^N$,

$$f_\Theta(\mathrm{z}) = \eta^{[K]}\Bigg(\cdots \eta^{[1]}\Big(W^{[1]}\mathrm{z} + b^{[1]}\Big) \cdots + b^{[K]}\Bigg), \tag{6}$$

where, for every $k \in \{1, \dots, K\}$, in the generic learning framework (e.g DnCNN [15]), $W^{[k]}$ denotes a linear transform such as convolutions or pooling to reduce the number of parameters, associated with a bias $b^{[k]}$, and a nonlinear activation function $\eta^{[k]}$ (e.g ReLu, sigmoid, HardTanh ...).

We propose two networks taking the form of (6): `Deep-(F)ISTA-GD` (Network 1) and `Deep-(Sc)CP-GD`

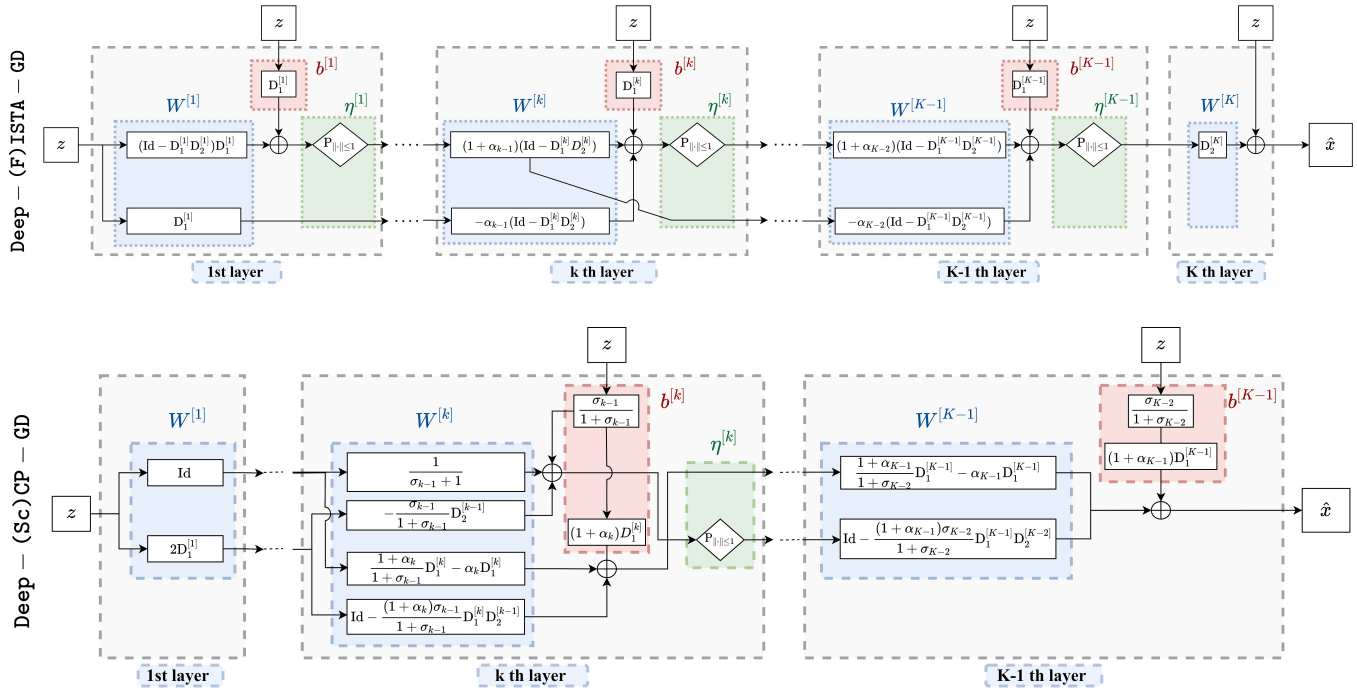

Fig. 1. Architecture of the proposed `Deep-(F)ISTA-GD` (top) and `Deep-(Sc)CP-GD` (bottom).

(Network 2). Illustrations of the two networks are provided in Figure 1.

Network 1. `Deep-(F)ISTA-GD` *for Gaussian Denoising has the following architecture, for every* $k \in \{2, \dots, K-1\}$*:*

$$\begin{cases} W^{[1]} = \begin{bmatrix} \mathrm{D}_1^{[1]} \\ (\mathrm{Id}_{|\mathbb{F}|} - \mathrm{D}_1^{[1]}\mathrm{D}_2^{[1]})\mathrm{D}_1^{[1]} \end{bmatrix}, \\ b^{[1]} = \begin{bmatrix} 0 \\ \mathrm{D}_1^{[1]} z_l \end{bmatrix}, \eta^{[1]} = \left\{ \begin{array}{c} \mathrm{Id}_{|\mathbb{F}|} \\ \mathrm{HardTanh}_\lambda \end{array} \right\}, \\ W^{[k]} = \begin{bmatrix} 0 & \mathrm{Id}_{|\mathbb{F}|} \\ -\alpha_{k-1}(\mathrm{Id}_{|\mathbb{F}|} - \mathrm{D}_1^{[k]}\mathrm{D}_2^{[k]}) & (1+\alpha_{k-1})(\mathrm{Id}_{|\mathbb{F}|} - \mathrm{D}_1^{[k]}\mathrm{D}_2^{[k]}) \end{bmatrix}, \\ b^{[k]} = \begin{bmatrix} 0 \\ \mathrm{D}_1^{[k]} z_l \end{bmatrix}, \eta^{[k]} = \left\{ \begin{array}{c} \mathrm{Id}_{|\mathbb{F}|} \\ \mathrm{HardTanh}_\lambda \end{array} \right\}, \\ W^{[K]} = \begin{bmatrix} 0 & -\mathrm{D}_2^{[k]} \end{bmatrix}, b^{[K]} = z_l, \eta^{[K]} = \mathrm{Id}_N \,. \end{cases}$$

Network 2. `Deep-(Sc)CP-GD` *for Gaussian Denoising has the following architecture, for every* $k \in \{2, \dots, K-1\}$*:*

$$\begin{cases} W^{[1]} = \begin{bmatrix} \mathrm{Id}_N \\ 2D_1^{[1]} \end{bmatrix}, b^{[1]} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \eta^{[1]} = \left\{ \begin{array}{c} \mathrm{Id}_N \\ \mathrm{HardTanh}_\lambda \end{array} \right\}, \\ W^{[k]} = \begin{bmatrix} \frac{1}{1+\sigma_{k-1}} & -\frac{\sigma_{k-1}}{1+\sigma_{k-1}}\mathrm{D}_2^{[k-1]} \\ \frac{1+\alpha_k}{1+\sigma_{k-1}} D_1^{[k]} - \alpha_k \mathrm{D}_1^{[k]} & \mathrm{Id}_{|\mathbb{F}|} - \frac{(1+\alpha_k)\sigma_{k-1}}{1+\sigma_{k-1}}\mathrm{D}_1^{[k]}\mathrm{D}_2^{[k-1]} \end{bmatrix}, \\ b^{[k]} = \begin{bmatrix} \frac{\sigma_{k-1}}{1+\sigma_{k-1}}\mathrm{z} \\ \frac{(1+\alpha_k)\sigma_{k-1}}{1+\sigma_{k-1}}\mathrm{D}_1^{[k]}\mathrm{z} \end{bmatrix}, \eta^{[k]} = \left\{ \begin{array}{c} \mathrm{Id}_N \\ \mathrm{HardTanh}_\lambda \end{array} \right\}, \\ W^{[K]} = \begin{bmatrix} \mathrm{Id}_N & 0 \end{bmatrix}, b^{[K]} = 0, \eta^{[K]} = \mathrm{Id}_N \,. \end{cases}$$

Proposition 2 (resp. Proposition 3) establishes the relation between Network 1 (resp. Network 2) and (F)ISTA (resp. (Sc)CP algorithm) described in section II-B. These relations are established when $g = \|\cdot\|_1$ noticing that the proximity operator of the conjugate of the ℓ_1-norm fits the HardTanh activation, a standard activation function used in deep denoiser.

Proposition 1. *The proximity operator of the conjugate of the* ℓ_1*-norm scaled by parameter* $\lambda > 0$ *fits the HardTanh activation function, i.e., for every* $\mathrm{x} = (\mathrm{x}_i)_{1 \le i \le N}$*:*

$$\mathrm{P}_{\lambda\|\cdot\|_1^*}(\mathrm{x}) = \mathrm{HardTanh}_\lambda(\mathrm{x}) = (\mathrm{p}_i)_{1 \le i \le N}$$

where

$$\mathrm{p}_i = \begin{cases} -\lambda & \textit{if} \quad \mathrm{p}_i < -\lambda, \\ \lambda & \textit{if} \quad \mathrm{p}_i > \lambda, \\ \mathrm{p}_i & \textit{otherwise}. \end{cases}$$

Proposition 2. *We set, for every* $k \in \{1, \dots, K\}, D_1^{[k]} \in \mathbb{R}^{|\mathbb{F}| \times N}, D_2^k \in \mathbb{R}^{N \times |\mathbb{F}|}$*,* $W^{[k]}$*,* $b^{[k]}$*and* $\eta^{[k]}$ *provided by Network 1. If* $g = \|\cdot\|_1$*,* $\mathrm{D}_1^{[k]} = \tau_k \mathrm{D}$ *and* $\mathrm{D}_2^{[k]} = \mathrm{D}^\top$*,* $\mathrm{u}_0 = \mathrm{u}_1 = \mathrm{D}_1^{[1]} z_l$*,* $\mathrm{y}_1 = (\mathrm{Id} - \mathrm{D}_1^{[1]}\mathrm{D}_2^{[1]})\mathrm{D}_1^{[1]} z_l$*, then* `Deep-(F)ISTA-GD` *network fits the generic* `(F)ISTA` *scheme* (3)*.*

A similar proposition holds for Deep-(Sc)CP-GD.

Proposition 3. *We set, for every* $k \in \{1, \dots, K\}, \mathrm{D}_1^{[k]} \in \mathbb{R}^{|\mathbb{F}| \times N}, \mathrm{D}_2^{[k]} \in \mathbb{R}^{N \times |\mathbb{F}|}$*,* $W^{[k]}$*,* $b^{[k]}$*, and* $\eta^{[k]}$ *provided by Network 2. If* $g = \|\cdot\|_1$*,* $\mathrm{D}_1^{[k]} = \tau_k \mathrm{D}$ *and* $\mathrm{D}_2^{[k]} = \mathrm{D}^\top$*,* $\mathrm{u}_1 = z_l$*,* $\alpha_1 = 1$*,* $\mathrm{y}_1 = \mathrm{D}_1^{[1]} z_l$*, then the* `Deep-(Sc)CP-GD` *network fits the generic* `(Sc)CP` *scheme* (4)*.*

D. Robustness of the newtork

Following [8], Network 1 and Network 2 have a Lipschitz behavior with Lipschitz constant $\chi = \prod_{k=1}^{K} \|W^{[k]}\|$. Thus,

given an input z and a perturbation ϵ, we can majorize the perturbation on the output via the inequality

$$\|f_\Theta(\mathrm{z}+\epsilon) - f_\Theta(\mathrm{z})\| \leq \chi\|\epsilon\|.$$

χ can be used as a certificate of the robustness of the network provided that it is tightly estimated. Tighter bound exists but at the price of much more complex computations.

III. Experiments

This section provides numerical comparisons between `Deep-ISTA-GD`, `Deep-FISTA-GD`, `Deep-CP-GD` and `Deep-ScCP-GD`, and aims to illustrate the impact of handling with unfolded deep accelerated schemes. Comparisons with state-of-the-art methods are also provided.

A. Experimental setting

Training and testing datasets – We combined RGB-images obtained from BSD500 and BSD300 [16] to construct our dataset including 300 training images of size 180×180 and 300 testing images of size 320×320 both being cropped from the database, as suggested in [15]. In our experiments, no data augmentation like flip or rotation has been done. The images are degraded with a white Gaussian noise with standard deviation $\delta = \{50, 100\}$.

Strategy to compare – We compare four configurations of the unfolded strategies described in Section II: `Deep-ISTA-GD` with parameter $\Theta = \{\mathrm{D}_1^{[k]}, \mathrm{D}_2^{[k]}, \alpha^{[k]} \equiv 0\}$, `Deep-FISTA-GD` with parameter $\Theta = \{\mathrm{D}_1^{[k]}, \mathrm{D}_2^{[k]}, \alpha^{[k]}\}$, `Deep-CP-GD` with parameter $\Theta = \{\mathrm{D}_1^{[k]}, \mathrm{D}_2^{[k]}, \sigma^{[k]}, \alpha^{[k]} \equiv 1\}$ suited to the case without handling with strong convexity (i.e., $\gamma = 0$), `Deep-ScCP-GD` with parameter $\Theta = \{\mathrm{D}_1^{[k]}, \mathrm{D}_2^{[k]}, \sigma^{[k]}, \alpha^{[k]} = \frac{1}{\sqrt{1+2\sigma^{[k]}}}\}$ associated with the choice $\gamma = 1$.

Parameter setting – We train all of our networks on Pytorch with ADAM optimizer [17] with 500 epochs and a learning rate set to 10^{-4}. The batch size is set to 4 in Section III B and set to 10 for Sections III C-D. The impact of the network depth K and the feature size $|\mathbb{F}|$ will be evaluated.

B. Standard versus accelerated unfolded schemes

In Figure 2, we provide a comparison between `Deep-ISTA-GD`, `Deep-FISTA-GD`, `Deep-CP-GD`, and `Deep-ScCP-GD` for different choices of layer number $K = \{5, 13, 21, 29, 37\}$ and feature size $|\mathbb{F}| = \{13, 21, 29, 37, 45\}$. The maps reported in Figure 2[top] displays the performance in terms of average PSNR over the testing dataset w.r.t $(K, |\mathbb{F}|)$. The map reported in Figure 2[bottom] displays the value of χ measuring the robustness for each network, where for every k, $\|W^{[k]}\|$ is computed with the power method.

We observe on the top row of Fig. 2 that `Deep-FISTA-GD` achieves better PSNR that `Deep-ISTA-GD` and that `Deep-ScCP-GD` achieves better PSNR that `Deep-CP-GD` for every set of parameters $(|\mathbb{F}|, K)$, allowing to conclude that accelerated schemes perform better. Additionally, `Deep-ScCP-GD` appears to be the most efficient method among the four for each range of parameters.

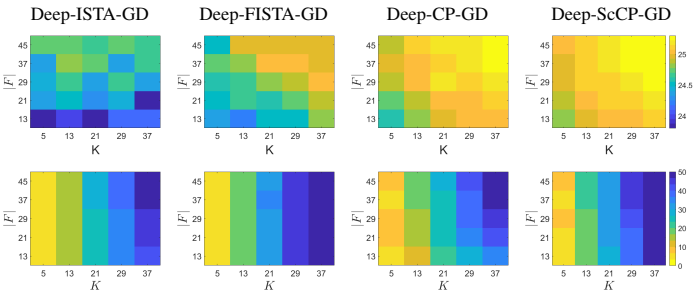

Fig. 2. $1st$ row: PSNR and $2nd$ row: χ (exponential scale) on a grid comparison between different choices of depth K and number of features $|\mathbb{F}|$ for each model.

In Fig. 2[bottom], we can observe that all networks have similar behaviours. The deeper is the network, the larger is the parameter χ. This tends to conclude that the deeper is the network the less robust is the network, which might be explained by the huge amount of parameters to learn for deeper networks.

C. Learning behaviour and comparison with DnCNN

In this section, the proposed networks are trained with $K = 13$ and $|\mathbb{F}| = 21$ being a good compromise between performance and robustness. We compare the proposed networks to DnCNN that appears to be one of the most efficient according to [15]. DnCNN has been re-implemented and re-trained with $K = 9$ and $|\mathbb{F}| = 13$. K and $|\mathbb{F}|$ has been selected to obtain a similar number of parameters to train for all networks (i.e. ~ 14800 parameters). The batch size is set to 10, leading to a gain of almost 1 dB for all methods compared to a training with batch size of size 4.

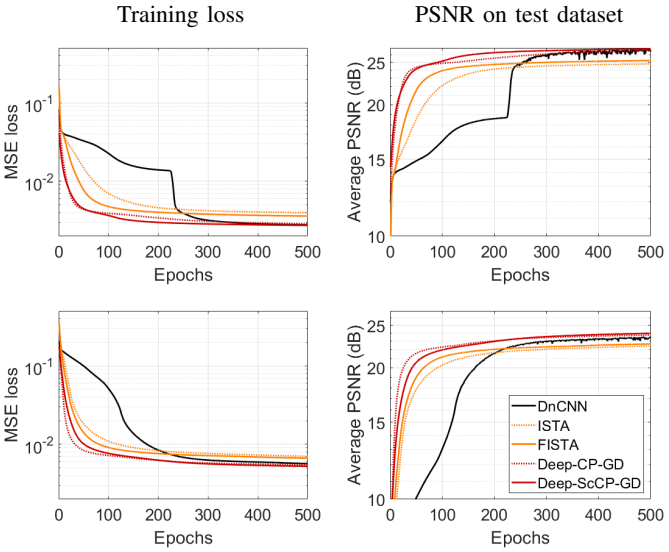

Fig. 3. Comparisons between `Deep-ISTA-GD` , `Deep-FISTA-GD` , `Deep-CP-GD` , `Deep-ScCP-GD`, and `DnCNN` in terms of Training loss and Averaged PSNR on the testing dataset. (top) $\delta = 50$ (bottom) $\delta = 100$.

In Fig.3, we observe that the training losses converge faster for accelerated-based unfolded schemes and that `Deep-ScCP-GD` reaches a higher PSNR faster than DnCNN. In terms of performance, from the PSNR plot, when $\delta = 50$,

only Deep−ScCP−GD provides better (even close) results than DnCNN and we can observe that Deep−ScCP−GD is much more stable than DnCNN (cf. oscillations in the black plot). For a larger value of δ, the benefit of Deep−ScCP−GD compared to DnCNN is observed both in terms of PSNR and stability. In the next section, we only run experiments with Deep−ScCP−GD.

D. Visual comparisons

In Figure 4,we provide visual comparisons between our proposed Deep−ScCP−GD and state-of-the-art methods: TV [18], NL-TV [19], DnCNN [15]. The hyperparameters in TV and NL-TV are trained on the training dataset with a multigrid search-parameter strategy described in [20]. TV and NL-TV fits the minimization of (1) when $g = \ell_{1,2}$-norm and for different choices of $|\mathbb{F}|$. The estimate $\widehat{\mathrm{x}}$ is obtained with FISTA. We can observe the good performance of the proposed Deep−ScCP−GD compared to state-of-the-art physics-based methods (TV, NL-TV) and black-box strategy (DnCNN).

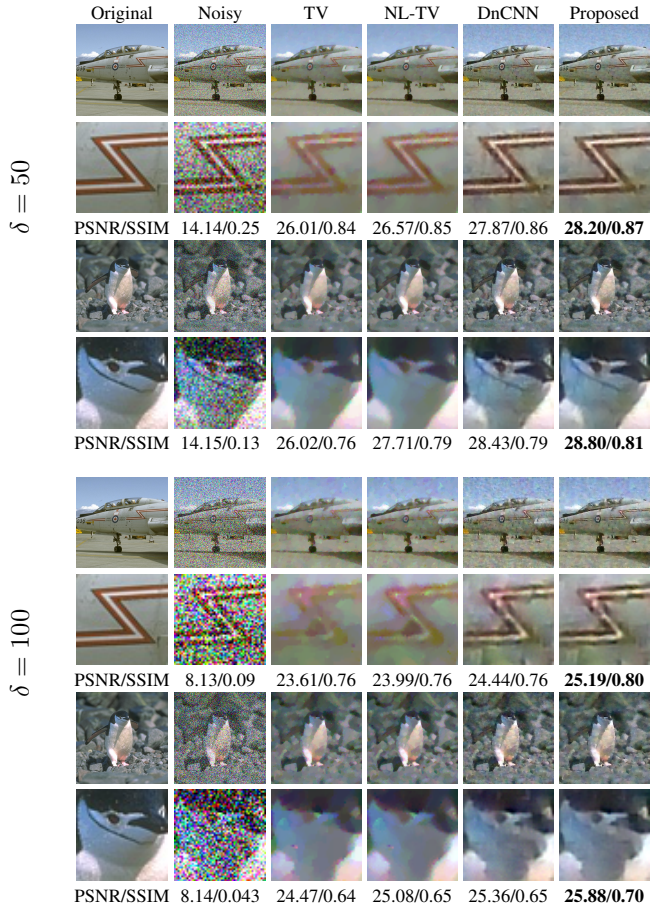

Fig. 4. Comparison between TV [18], NL-TV [4], DnCNN [15] and our proposed network Deep−ScCP−GD on different RGB images degraded with a white Gaussian noise of standard deviation δ.

IV. Conclusion

In this work, we proposed two deep unfolded networks for gaussian denoising that can be activated from standard or accelerated schemes. All networks are related to standard proximal algorithms: ISTA, FISTA, CP, and ScCP. Our numerical experiments illustrate the benefit to unroll accelerated schemes when possible. Additionally, the proposed scheme Deep−ScCP−GD is the most efficient from the PSNR performance and stability, and obtain better results than state-of-the-art methods. Further experiments on the impact of accelerated schemes for solving more complex inverse problems would be investigated in future works.

References

[1] J. Bruna and S. Mallat, "Invariant scattering convolution networks," *IEEE Trans. Pattern Anal. Match. Int.*, vol. 35, no. 8, pp. 1872–1886, Aug. 2013.

[2] E. Oyallon, S. Mallat, and L. Sifre, "Generic deep networks with wavelet scatteringe oyallon, s mallat, l sifreinternational conference on learning representations," in *Proc. of Int. Conf. Learn. Represent.*, Banff, Canada, Apr. 14-16 2014.

[3] H. Bolcskei, P. Grohs, G. Kutyniok, and P. Petersen, "Optimal approximation with sparsely connected deep neural networks," *SIAM J. Math. Data Sci.*, vol. 1, no. 1, pp. 8–45, 2019.

[4] P. Milanfar, "A tour of modern image filtering: New insights and methods, both practical and theoretical," *IEEE Signal Process. Mag.*, vol. 30, no. 1, pp. 106–128, 2012.

[5] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3D transform-domain collaborative filtering," *IEEE Trans. Image Process.*, vol. 16, no. 8, pp. 2080 – 2095, Aug. 2007.

[6] J. Tachella, J. Tang, and M. Davies, "The neural tangent link between CNN denoisers and non-local filters," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2021, pp. 8618–8627.

[7] K. Gregor and Y. Le Cun, "Learning fast approximations of sparse coding," in *Proc. International Conference on Machine Learning*, Haifa, Israel, Jun. 21-24 2010, pp. 399–406.

[8] P. L. Combettes and J.-C. Pesquet, "Lipschitz certificates for layered network structures driven by averaged activation operators," *SIAM J. Math. Data Sci.*, vol. 2, no. 2, pp. 529–557, 2020.

[9] A. Chambolle and T. Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging," *J. Math. Imag. Vis.*, vol. 40, no. 1, pp. 120–145, 2011.

[10] M. A. T. Figueiredo and R. D. Nowak, "An EM algorithm for wavelet-based image restoration," *IEEE Trans. Image Process.*, vol. 12, no. 8, pp. 906–916, Aug. 2003.

[11] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Comm. Pure Applied Math.*, vol. 57, no. 11, pp. 1413–1457, Nov. 2004.

[12] P.L. Combettes and V.R Wajs, "Signal recovery by proximal forward-backward splitting," *Multiscale model. simul.*, vol. 4, no. 4, pp. 1168–1200, 2005.

[13] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imaging Sci.*, vol. 2, no. 1, pp. 183–202, 2009.

[14] A. Chambolle and C. Dossal, "On the Convergence of the Iterates of the Fast Iterative Shrinkage/Thresholding Algorithm," *J. Optim. Theory Appl.*, vol. 166, no. 3, pp. 968–982, 2015.

[15] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142–3155, 2017.

[16] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proc. 8th Int Conf. Computer Vision*, July 2001, vol. 2, pp. 416–423.

[17] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[18] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D: nonlinear phenomena*, vol. 60, no. 1-4, pp. 259–268, 1992.

[19] G. Chierchia, N. Pustelnik, B. Pesquet-Popescu, and J.-C. Pesquet, "A nonlocal structure tensor-based approach for multicomponent image recovery problems," *IEEE Trans. Image Process.*, vol. 23, no. 12, pp. 5531–5544, Dec. 2014.

[20] H.T.V Le, M. Foare, and N. Pustelnik, "A proximal based strategy for solving Discrete Mumford-Shah and Ambrosio-Tortorelli models," *IEEE Signal Process. Lett.*, 2022, to appear.