Bayesian Deep Deconvolutional Neural Networks

A. Bhowmik* **A. Adiga*** **C. S. Seelamantula*** **F. Hauser†** **J. Jacak†** **B. Heise†**
*Department of Electrical Engineering, Indian Institute of Science, Bangalore, India
†Institute of Applied Physics, Johannes Kepler University, Linz, Austria
aritra0593@gmail.com, {aaniruddha, chandrasekhar}@iisc.ac.in
{Fabian.Hauser, Jaroslaw.Jacak}@fh-linz.at, bettina.heise@recendt.at

Abstract

We address the problem of sparse spike deconvolution from noisy measurements within a Bayesian paradigm that incorporates sparsity promoting priors about the ground truth. The optimization problem arising out of this formulation warrants an iterative solution, which we accomplish using a deep neural network (DNN). The architecture of the DNN is such that the weights and biases in each layer are fixed and determined by the blur matrix and the noisy measurements, and the sparsity promoting prior determines the activation function in each layer. In scenarios where the priors are not available exactly, but adequate training data is available, the formulation can be adapted to learn the priors by parameterizing the activation function using a linear expansion of threshold (LET) functions. As a proof of concept, we demonstrate successful spike deconvolution on synthetic dataset and compare our results with the fast iterative shrinkage-thresholding algorithm (FISTA). We also show an application of the proposed method for performing image reconstruction in super-resolution localization microscopy.

1 Introduction

Sparse spike deconvolution is the problem of determining the point-source excitation $\mathbf{X} \in \mathbb{R}^{M\times N}$ from noisy, blurred measurements $\mathbf{Y} \in \mathbb{R}^{M\times N}$ modeled as

$$\mathbf{Y} = \mathbf{H} * \mathbf{X} + \mathbf{W}, \tag{1}$$

where $*$ represents 2-D convolution, $\mathbf{H} \in \mathbb{R}^{M\times N}$ is the point-spread function (PSF), and $\mathbf{W} \in \mathbb{R}^{M\times N}$ is the additive white Gaussian noise (AWGN). This model is frequently encountered in localization microscopy [1–3], astronomical imaging [4], deflectometry [5], etc. Typically, $\mathbf{H}$ is non-invertible, which makes the linear inverse problem in (1) ill-posed. However, one could circumvent this problem by incorporating priors on $\mathbf{X}$, and reformulating the reconstruction within a Bayesian framework.

1.1 A Maximum A posteriori (MAP) Formulation

The point-source excitation $\mathbf{X}$ is assumed to contain i.i.d. entries and following the distribution $g(\mathbf{X})$. Typically, g promotes sparsity in $\mathbf{X}$. Further, the PSF is assumed to be separable, that is, $\mathbf{H} = \mathbf{h}_r\mathbf{h}_c^\top$, where $\mathbf{h}_r$ and $\mathbf{h}_c$ are the blur kernels along the x and y axes, respectively. Hence, the 2-D convolution operation in (1) can be expressed as $\mathbf{H} * \mathbf{X} = \mathbf{H}_c\mathbf{X}\mathbf{H}_r^\top$, where $\mathbf{H}_c$ and $\mathbf{H}_r$ are Toeplitz matrices constructed from $\mathbf{h}_c$ and $\mathbf{h}_r$, respectively [6]. Denoting the likelihood of the observations by f, we consider the maximum a posteriori (MAP) estimate:

$$\mathbf{X}_{\text{MAP}} = \arg\max_{\mathbf{X}} f(\mathbf{Y}/\mathbf{X};\mathbf{H})g(\mathbf{X}),$$

Second workshop on Bayesian Deep Learning (NIPS 2017), Long Beach, CA, USA.

which turns out to be a solution to an optimization problem of the form:

$$\mathbf{X}_{\text{MAP}} = \arg\min_{\mathbf{X}} \frac{1}{2} \left\| \mathbf{Y} - \mathbf{H}_c \mathbf{X} \mathbf{H}_r^\top \right\|_{\text{F}}^2 + \lambda \mathcal{G}\left(\mathbf{X}\right), \tag{2}$$

where $\mathcal{G}(\mathbf{X}) = \log g(\mathbf{X})$ acts as the regularizer, λ encapsulates the parameters of the distribution, and the subscript F denotes the Frobenius norm.

2 A Deep Deconvolutional Neural Network

Consider an affine approximation to $f(\mathbf{X}) = \left\| \mathbf{Y} - \mathbf{H}_c \mathbf{X} \mathbf{H}_r^\top \right\|_{\text{F}}^2$ in (2) at $\mathbf{X}^\ell$, and the following update rule:

$$\mathbf{X}^{\ell+1} = \arg\min_{\mathbf{X}} f\left(\mathbf{X}^\ell\right) + \text{Tr}\left(\mathbf{X} - \mathbf{X}^\ell\right)^\top \nabla f(\mathbf{X}^\ell) + \frac{1}{2\eta} \left\| \mathbf{X} - \mathbf{X}^\ell \right\|_{\text{F}}^2 + \lambda \mathcal{G}\left(\mathbf{X}\right), \tag{3}$$

where $\left\| \mathbf{X} - \mathbf{X}^\ell \right\|_{\text{F}}^2$ serves as the proximal term. For separable $\mathcal{G}(\mathbf{X})$, employing the definitions of proximal operators [7, 8], the update is expressed equivalently as

$$\mathbf{X}^{\ell+1} = \mathcal{P}_\nu \left(\mathbf{X}^\ell - \eta \nabla f\left(\mathbf{X}^\ell\right)\right), \tag{4}$$

where $\mathcal{P}_\nu$ is the proximal operator corresponding to $\mathcal{G}$ and $\nu = \lambda\eta$. For the case

$$\mathcal{G}(\mathbf{X}) = \|\mathbf{X}\|_1 \triangleq \sum_{i=1}^{M} \sum_{j=1}^{N} |X_{ij}|,$$

the proximal operator turns out to be the element-wise soft-thresholding (ST) function [9]:

$$\mathcal{P}_\nu\left(X_{ij}\right) = \text{sgn}(X_{ij}) \max\left\{|X_{ij}| - \nu, 0\right\},$$

where sgn$(\cdot)$ denotes the signum function. The resulting update corresponds to the standard iterative shrinkage and thresholding algorithm (ISTA) [10]. Using the gradient expression

$$\nabla f\left(\mathbf{X}\right) = \mathbf{H}_c{}^\top \left(\mathbf{H}_c \mathbf{X} \mathbf{H}_r - \mathbf{Y}\right) \mathbf{H}_r{}^\top,$$

(4) can be expressed in the following form

$$\mathbf{X}^{\ell+1} = \mathcal{P}_\nu \left(\mathbf{X}^\ell - \eta \mathbf{H}_c{}^\top \mathbf{H}_c \mathbf{X}^\ell \mathbf{H}_r \mathbf{H}_r{}^\top + \mathbf{C}\right), \tag{5}$$

where $\mathbf{C} = \eta \mathbf{H}_c{}^\top \mathbf{Y} \mathbf{H}_r{}^\top$. Gregor and LeCun encountered similar update equations in the context of sparse coding and interpreted the update steps as one layer of a feedforward DNN [11]. Inspired by a similar connection in the deconvolution problem at hand, we interpret (5) as the feedforward computation through a neural network (NN) with weight matrix determined by $\mathbf{H}_c$ and $\mathbf{H}_r$, and bias $\mathbf{C}$ shared across layers. The input to the NN is the initialization $\mathbf{X}^0$. The proximal operator $\mathcal{P}_\nu\left(\cdot\right)$ effectively plays the role of the nonlinear activation function ψ of the neurons in layer ℓ. An L-stage iterative algorithm now becomes equivalent to an L-layer DNN. This connection becomes particularly effective when the priors are not explicitly available but adequate training data is available from which the priors can be learnt. Effectively, this also corresponds to learning the appropriate regularizer, which renders the DNN capable of performing deconvolution. We refer to this network as the Bayesian Deep Deconvolutional Neural Network (BD^{2}N^2).

2.1 Learning Priors

We parameterize the activation function ψ as a linear combination of K derivatives of a Gaussian (DoG) and the linear function [12] as

$$\psi(u) = \sum_{k=1}^{K} c_k \phi_k(u), u \in \mathbb{R},$$

where

$$\phi_k(u) = u \, \exp\left(-\frac{(k-1)u^2}{2\tau^2}\right).$$

The activation function, although nonlinear in u, is linear in the coefficients $\{c_k\}$, which can be learnt using a training dataset containing blurred images and their ground-truth target pairs. The primary motivation for using a DoG-based parameterization is its high success over the soft-threshold in several 2-D and 3-D deconvolution [12] and denoising problems [13]. This parameterization actually leads to a wide variety of sparsity inducing regularizers, beyond the standard ℓ_1 norm — this will be reported separately in a journal version of this paper. ISTA has a convergence rate of $\mathcal{O}\left(\frac{1}{\ell}\right)$, which can be improved to $\mathcal{O}\left(\frac{1}{\ell^2}\right)$ by incorporating a momentum factor [14, 15], leading to the fast ISTA (FISTA) algorithm. It can be shown that FISTA has a deep residual network type architecture [16].

2.2 Learning Optimal Activation of BD^{2}N^2

The training dataset $\mathcal{D}$ consists of N examples $\{(\mathbf{Y}_q, \mathbf{X}_q)\}_{q=1}^N$, where $\mathbf{Y}_q = \mathbf{H}_c\mathbf{X}_q\mathbf{H}_r^\top + \boldsymbol{\xi}_q$. The random noise vectors $\boldsymbol{\xi}_q$ are assumed to be independent and identically distributed. Let $\mathbf{c}^\ell \in \mathbb{R}^K$, $\ell = 1 : L$, be the coefficients of the LET activation in layer ℓ. For the q^{th} example in the dataset, the prediction $\mathbf{X}_q^L$ of the L^{th} layer is a function of the corresponding measurement vector $\mathbf{Y}_q$ and the LET coefficient super-vector $\mathbf{c} = \left[\left[\mathbf{c}^1\right]^\top \mid \left[\mathbf{c}^2\right]^\top \mid \left[\mathbf{c}^3\right]^\top \cdots \left[\mathbf{c}^L\right]^\top\right]^\top$. The optimal set of activation parameters $\mathbf{c}^*$ is obtained by minimizing the squared estimation error over all training examples:

$$J(\mathbf{c}) = \frac{1}{2}\sum_{q=1}^{N} \|\mathbf{X}_q^L\left(\mathbf{Y}_q, \mathbf{c}\right) - \mathbf{X}_q\|_2^2.$$

The optimization requires knowledge of the gradient of $J(\mathbf{c})$ with respect to $\mathbf{c}$. In general, the optimization of $J(\mathbf{c})$ using vanilla gradient-descent (GD) tends to diverge, unless a very small step size is chosen. We overcome this hurdle by noting that the Hessian need not be computed explicitly. All that is needed is the Hessian-vector product to train the parameters of the network. This is precisely what the *Hessian-free optimization* (HFO) technique [17] guarantees, which is therefore employed in our training procedure. In the i^{th} epoch of HFO [17], the search direction $\boldsymbol{\delta}_{\mathbf{c}}^*$ is obtained by minimizing a second-order Taylor-series approximation $\tilde{J}\left(\mathbf{c}\right)$ to the actual cost $J(\mathbf{c})$ at the i^{th} iterate $\mathbf{c}_i$:

$$\tilde{J}\left(\mathbf{c}_i + \boldsymbol{\delta}_{\mathbf{c}}\right) = J\left(\mathbf{c}_i\right) + \boldsymbol{\delta}_{\mathbf{c}}^\top \mathbf{g}_i + \frac{1}{2}\boldsymbol{\delta}_{\mathbf{c}}^\top \mathbf{H}_i \boldsymbol{\delta}_{\mathbf{c}},$$

where $\mathbf{g}_i = \nabla J(\mathbf{c})\big|_{\mathbf{c}=\mathbf{c}_i}$, $\mathbf{H}_i = \nabla^2 J(\mathbf{c})\big|_{\mathbf{c}=\mathbf{c}_i}$, and $\boldsymbol{\delta}_{\mathbf{c}}$ is the search direction to be chosen optimally at every iteration by minimizing a regularized quadratic approximation:

$$\boldsymbol{\delta}_{\mathbf{c}}^* = \arg\min_{\boldsymbol{\delta}_{\mathbf{c}}} \tilde{J}\left(\mathbf{c}_i + \boldsymbol{\delta}_{\mathbf{c}}\right) + \gamma\|\boldsymbol{\delta}_{\mathbf{c}}\|_2^2.$$

2.3 Experimental Validation on Synthetic Data

In order to evaluate the performance of our network architecture, we tested our BD^{2}N^2 technique on synthetic datasets. We created three sets of 50 synthetic images of size 128×128. The three sets were divided into images containing 15, 25, and 35 randomly generated points, respectively. Each image is then convolved with a Gaussian PSF of size 3×3. Further, to each set of the 50 images, we added AWGN with standard deviation of 3, 5, and 7. In each set, we took 10 images for training and retained 40 for testing. Finally, the reconstruction peak signal-to-noise ratio (PSNR) was calculated by averaging over the PSNR outputs of the test images. We observe from Fig. 1, the BD^{2}N^2 approach is able to reconstruct a deblurred image from the noisy and blurred one. The reconstructed image is of high quality and resembles the ground-truth image very closely. The PSNR gain over FISTA was computed to be about 4 dB, which is significant.

3 Application to Super-Resolution Localization Microscopy

We now present experimental results related to localization of point-sources in stochastic optical reconstruction microscopic (STORM) [18] imaging of *Fibronectin* samples. STORM is a super-resolution imaging technique [19], which achieves a resolution of the order of 10 nm, which is below the diffraction limit. In STORM imaging, one acquires a sequence of images where each image can be modeled as in (1) and then localizes the point-sources [20, 21] using a Gaussian peak-fitting

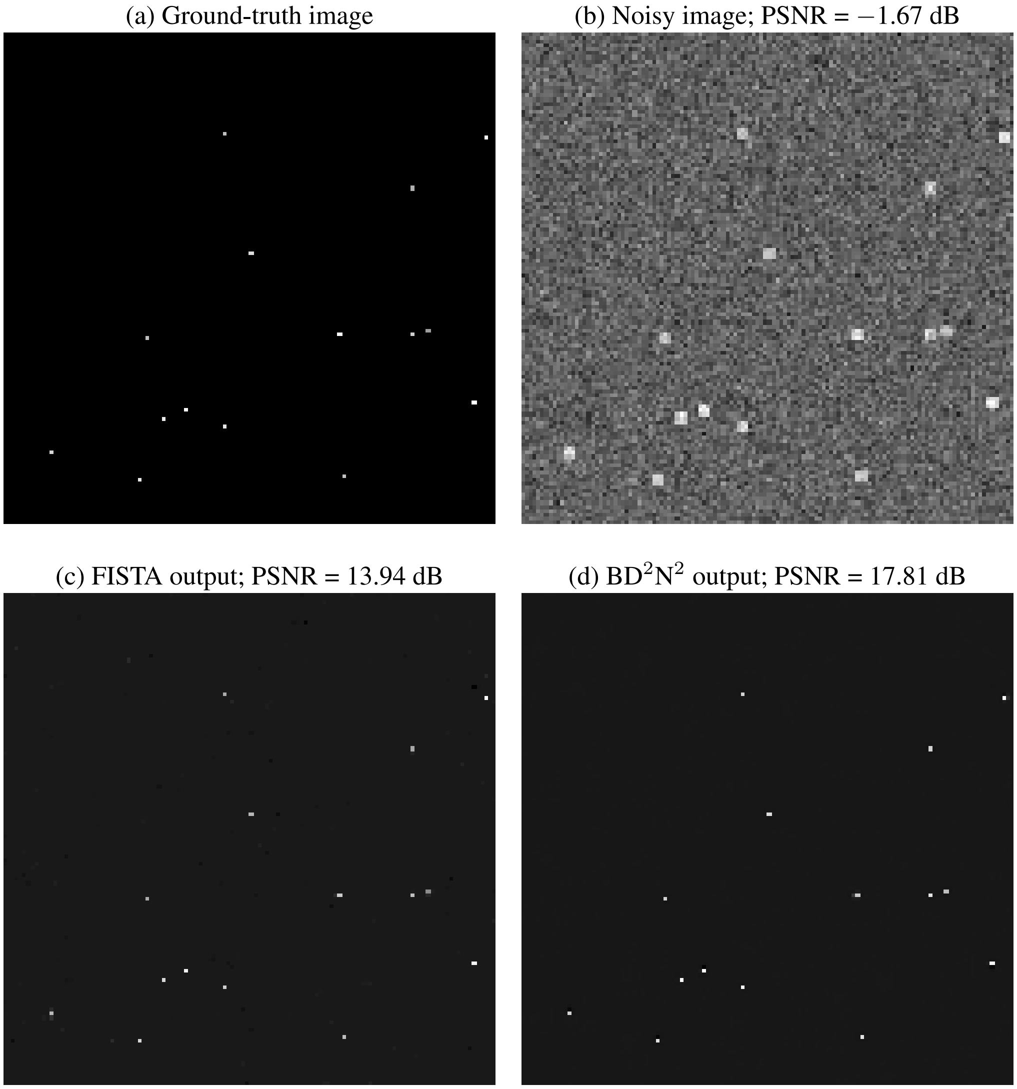
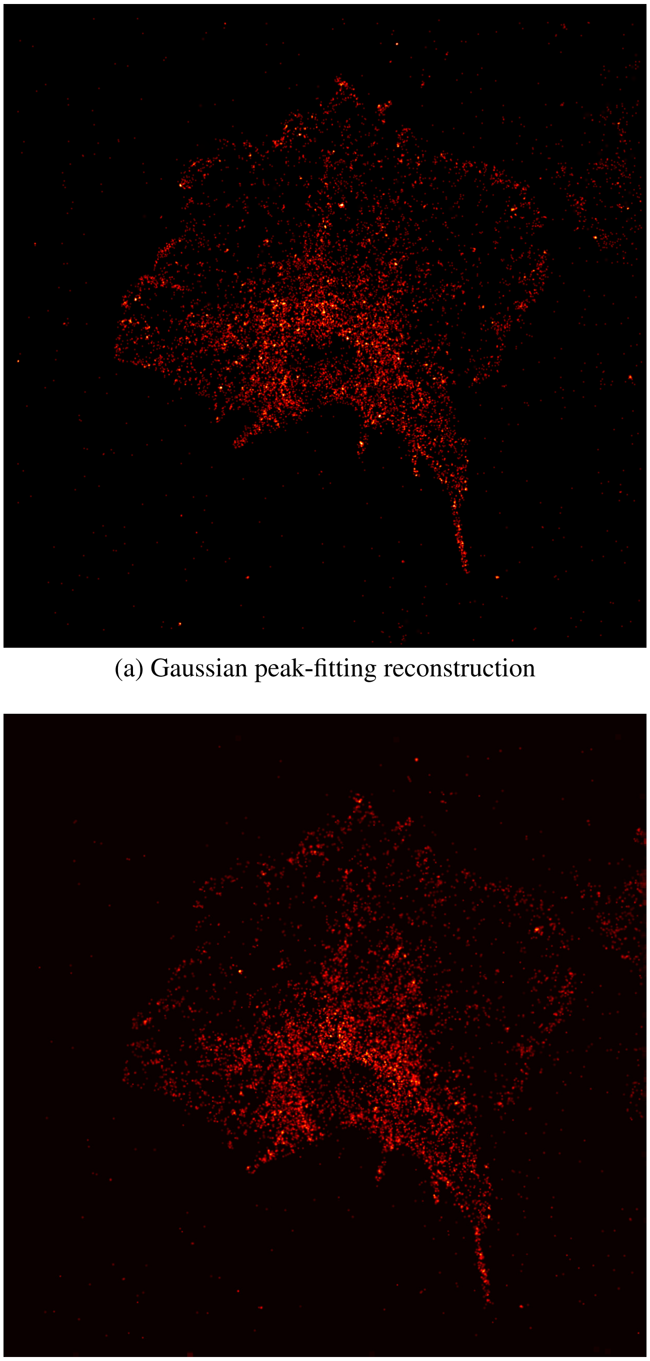


Figure 1: Deconvolution results on a synthetic image comprising 15 point sources: **(a)** Ground-truth image; **(b)** Noisy image; **(c)** FISTA output; and **(d)** BD^{2}N^2 output.

technique. We employ the proposed BD^{2}N^2 based deconvolution technique to perform localization. The dataset consists of 9994 low-resolution frames of size 129×129. The PSF is a Gaussian kernel with $\sigma_{\text{PSF}} = 150$ nm. For training the BD^{2}N^2, 50 frames were randomly chosen from the dataset, and the remaining were used for testing. Each localized point denoted as $(\hat{x}_\ell, \hat{y}_\ell)$ is further associated with a Gaussian uncertainty blob as determined by Thompson's rule [22]. The final image is rendered by creating a super-resolved grid of size 1290×1290 and placing a Gaussian uncertainty blob at $[\text{round}(10\hat{x}_\ell), \text{round}(10\hat{y}_\ell)]$. The reconstructed images are shown in Fig. 2. We observe that the image reconstructed by the BD^{2}N^2 is on par with the benchmark Gaussian peak detection and fitting technique.

4 Conclusions

We considered the problem of sparse spike deconvolution in the presence of noise within a Bayesian formulation. We considered an iterative algorithm to solve the deconvolution problem and established a one-to-one equivalence with a deep neural network architecture. Upon training the network to learn the activation function, it becomes capable of learning priors and consequently the optimal regularizers. Deconvolution of synthetic datasets showed that the BD^{2}N^2 is capable of giving high



(a) Gaussian peak-fitting reconstruction

(b) BD^2N^2 reconstruction

Figure 2: (Color online) A comparison of super-resolved *Fibronectin* images reconstructed by the BD^2N^2 technique vis-à-vis the *de facto* standard Gaussian peak-fitting algorithm [20, 21].

accuracy reconstruction, about 4 dB higher in PSNR compared with FISTA. We also demonstrated successful application of the BD^2N^2 approach for performing deconvolution of image stacks acquired in a super-resolution localization microscopy setup. The BD^2N^2 approach resulted in a high quality of image reconstruction competitive with the *de facto* standard.

References

[1] E. Betzig, G. H. Patterson, R. Sougrat, O. W. Lindwasser, S. Olenych, J. S. Bonifacino, M. W. Davidson, J. Lippincott-Schwartz, and H. F. Hess, "Imaging intracellular fluorescent proteins at nanometer resolution," *Science*, vol. 313, no. 5793, pp. 1642–1645, 2006.

[2] J. Fölling, M. Bossi, H. Bock, R. Medda, C. A. Wurm, B. Hein, S. Jakobs, C. Eggeling, and S. W. Hell, "Fluorescence nanoscopy by ground-state depletion and single-molecule return," *Nature Methods*, vol. 5, no. 11, pp. 943–945, 2008.

[3] V. Studer, J. Bobin, M. Chahid, H. S. Mousavi, E. Candès, and M. Dahan, "Compressive fluorescence microscopy for biological and hyperspectral imaging," in *Proceedings of National Academy of Sciences*, vol. 109, no. 26, 2012, pp. E1679–E1687.

[4] E. Pantin, J. L. Starck, and F. Murtagh, "Deconvolution and blind deconvolution in astronomy," in *Blind Image Deconvolution: Theory and Applications*. CRC press, 2007, pp. 100–138.

[5] P. Sudhakar, L. Jacques, X. Dubois, P. Antoine, and L. Joannes, "Compressive schlieren deflectometry," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 5999–6003.

[6] P. C. Hansen, J. G. Nagy, and D. P. O'leary, *Deblurring Images: Matrices, Spectra, and Filtering*. SIAM Press, 2006.

[7] J.-J. Moreau, "Fonctions convexes duales et points proximaux dans un espace Hilbertien," *Reports of the Paris Academy of Sciences, Series A*, vol. 255, pp. 2897–2899, 1962.

[8] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski, "Optimization with sparsity-inducing penalties," *Foundations and Trends in Machine Learning*, vol. 4, no. 1, pp. 1–106, 2012.

[9] D. L. Donoho and I. M. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage," *Journal of the American Statistical Association*, vol. 90, no. 432, pp. 1200–1224, 1995.

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

[11] K. Gregor and Y. LeCun, "Learning fast approximations of sparse coding," in *Proceedings of the 27th International Conference on Machine Learning*, 2010, pp. 399–406.

[12] H. Pan and T. Blu, "An iterative linear expansion of thresholds for-based image restoration," *IEEE Transactions on Image Processing*, vol. 22, no. 9, pp. 3715–3728, 2013.

[13] T. Blu and F. Luisier, "The SURE-LET approach to image denoising," *IEEE Transactions on Image Processing*, vol. 16, no. 11, pp. 2778–2786, 2007.

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

[15] Y. Nesterov, "A method of solving a convex programming problem with convergence rate $\mathcal{O}(1/k^2)$," in *Soviet Mathematics Doklady*, vol. 27, no. 2, 1983, pp. 372–376.

[16] D. Wipf and S. Nagarajan, "Iterative reweighted ℓ_1 and ℓ_2 methods for finding sparse solutions," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 2, pp. 317–329, Apr. 2010.

[17] J. Martens, "Deep learning via Hessian-free optimization," in *Proceedings of the 27th International Conference on Machine Learning*, 2010, pp. 735–742.

[18] M. J. Rust, M. Bates, and X. Zhuang, "Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM)," *Nature Methods*, vol. 3, no. 10, pp. 793–796, 2006.

[19] L. Schermelleh, R. Heintzmann, and H. Leonhardt, "A guide to super-resolution fluorescence microscopy," *The Journal of Cell Biology*, vol. 190, no. 2, pp. 165–175, 2010. [Online]. Available: http://jcb.rupress.org/content/190/2/165

[20] M. K. Cheezum, W. F. Walker, and W. H. Guilford, "Quantitative comparison of algorithms for tracking single fluorescent particles," *Biophysical Journal*, vol. 81, no. 4, pp. 2378–2388, 2001.

[21] E. J. Rees, M. Erdelyi, G. S. K. Schierle, A. Knight, and C. F. Kaminski, "Elements of image processing in localization microscopy," *Journal of Optics*, vol. 15, no. 9, p. 094012, 2013.

[22] K. I. Mortensen, L. S. Churchman, J. A. Spudich, and H. Flyvbjerg, "Optimized localization analysis for single-molecule tracking and super-resolution microscopy," *Nature Methods*, vol. 7, no. 5, pp. 377–381, 2010.