Iterative Operator Sketching Framework for Large-Scale Imaging Inverse Problems

Junqi Tang, Subhadip Mukherjee, Carola-Bibiane Schönlieb

Iterative Operator Sketching Framework for Large-Scale Imaging Inverse Problems

1st Junqi Tang
School of Mathematics
University of Birmingham
Birmingham, UK
j.tang.2@bham.ac.uk

2nd Subhadip Mukherjee
ECE
IIT-Kharagpur
Kharagpur, India
subhadipju@gmail.com

3rd Carola-Bibiane Schönlieb
DAMTP
University of Cambridge
Cambridge, UK
cbs31@cam.ac.uk

***Abstract*—Although having demonstrated great numerical performance in various imaging applications, the iterative data-driven reconstruction (IDR) schemes, e.g., plug-and-play algorithms and deep unrolling networks, can have significant computational limitations in large-scale imaging inverse problems in practice. This is mostly due to the fact that they need to involve the high-dimensional forward/adjoint operators which are non-trivial to compute in each iteration. In this work, we propose a new operator sketching framework tailored for designing efficient IDR schemes, which are currently state-of-the-art solutions for imaging inverse problems. Our framework performs dimensionality reduction on both image and measurement data domain, leading to efficient computations. Using this framework, we derive several accelerated IDR schemes, such as the plug-and-play multi-stage sketched gradient (PnP-MS2G) and sketching-based primal-dual (LSPD and Sk-LSPD) deep unrolling networks. Our numerical experiments on X-ray CT image reconstruction demonstrate the remarkable effectiveness of our sketched IDR schemes.**

***Index Terms*—Plug-and-Play Priors, Deep Unrolling, Sketching, Stochastic Optimization.**

I. Introduction

Large-scale imaging inverse problems have become crucial in various medical and industrial applications, such as X-ray CT, MRI and PET. Such imaging systems can be generally expressed as:

$$b = Ax^{\dagger} + w, \tag{1}$$

where $x^{\dagger} \in \mathbb{R}^d$, $A \in \mathbb{R}^{n\times d}$, $b \in \mathbb{R}^n$, and $w \in \mathbb{R}^n$ denote the ground truth image, the forward operator which models the measurement physics, the measurement data, and the (possibly data-dependent) measurement noise, respectively. In this work, we propose a new paradigm for designing fast iterative data-driven reconstruction (IDR) schemes, such as the PnP algorithms [17], [21] and deep unrolling schemes [1], [10], [16], utilizing the principle of dimension reduction for solving high-dimensional imaging inverse problems.

Traditionally, to obtain an estimate of the ground truth $x^{\dagger}$, we typically solve an optimization problem of the form:

$$x^{\star} \in \arg\min_{x\in\mathbb{R}^d} f(b, Ax) + g(x), \tag{2}$$

where data fidelity term $f(b, Ax)$ is a convex function in x, and the most typical and widely used choice of the data fidelity is the squared-error loss $\|b - Ax\|_2^2$. $g(x)$ is a regularization function that is usually convex to ensure provable convergence and robustness of the estimation, such as the TV regularization applied on the image domain and ℓ_1 regularization in the wavelet/shearlet domain. The problem (2) can be efficiently solved using proximal splitting methods [5].

While these classical convex regularization approaches provide theoretically tractable solutions for inverse problems, they have been significantly outperformed by the PnP priors, constructed by advanced image denoisers or deep neural networks. For example, the PnP-ADMM of [21] and PnP-LBFGS of [18] extend the classical optimization methods ADMM and quasi-Newton, replacing the proximal operator with the denoiser and have been widely applied in solving inverse problems since then. On the other hand, a very similar approach named regularization-by-denoising (RED) has been proposed by [13], [15], which explicitly constructs the regularization term using the denoiser and provides improved convenience in parameter tuning. Since a strong link between PnP and RED is established in [4] under the RED-PRO unifying framework, we refer to both PnP and RED as "PnP" for simplicity. The deep unrolling schemes [1], [10], [16] have been developed in parallel to the PnP schemes. Although having similar algorithmic structure to PnP schemes, they are based on unfolding the proximal splitting methods while parameterizing the proximal operators with deep neural networks and then trained end-to-end. For example, the Learned Primal-Dual network of [1] is based on unrolling the iterations of the Chambolle-Pock algorithm [3] via parameterizing the proximal operators with CNNs, leading to state-of-the-art results in image reconstruction tasks.

This paper is a short conference version of the works recorded in our online technical reports [19], [20]. CBS acknowledges support from the Philip Leverhulme Prize, the Royal Society Wolfson Fellowship, the EPSRC advanced career fellowship EP/V029428/1, the EPSRC programme grant EP/V026259/1, and the EPSRC grants EP/S026045/1 and EP/T003553/1, EP/N014588/1, EP/T017961/1, the Wellcome Innovator Awards 215733/Z/19/Z and 221633/Z/20/Z, the European Union Horizon 2020 research and innovation programme under the Marie Skodowska-Curie grant agreement NoMADS and REMODEL, the Cantab Capital Institute for the Mathematics of Information and the Alan Turing Institute. This research was also supported by the NIHR Cambridge Biomedical Research Centre (NIHR203312). The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care.

The IDR schemes such as PnP and deep unrolling may require a long computational time to obtain a good estimate when the image and measurement dimensions are large. We propose a generic acceleration scheme for IDR using dimensionality-reduction/sketching both in the image space and measurement space.

II. Multi-Stage Sketched Iterative Reconstruction

Our framework performs sketching in both image domain (of dimension d) and data domain (of dimension n). For ease of illustration, we use the least-squares objective and linear forward operator here without loss of generality. For a given forward operator $A \in \mathbb{R}^{n\times d}$, we can often find a low dimensional proxy $A_s \in \mathbb{R}^{n\times m_0}$ discretized on a reduced image dimension $m_0 < d$ such that $Ax \approx A_s\mathcal{S}(x)$, where $\mathcal{S}(\cdot) : \mathbb{R}^d \to \mathbb{R}^{m_0}$ ($m_0 < d$) is a sketching/downsampling operator. Furthermore, we can also perform random sketching $M(\cdot) : \mathbb{R}^n \to \mathbb{R}^m$ ($m < n$) on the measurement/data domain, which corresponds to stochastic approximation [14]. One typical choice of this sketching operator M is the subsampling sketch – uniformly sampled minibatch of $I_{n\times n}$ [12], which is suitable for inverse problems. For the image domain sketching operator $\mathcal{S}$, we found that off-the-shelf down-sampling algorithms such as the bi-cubic interpolation suffice in our framework. Now we can summarize this double-sketching as follows:

$$\begin{aligned}\|b - Ax\|_2^2 &\approx \|b - A_s\mathcal{S}(x)\|_2^2 \\ &\propto \mathbf{E}_M\|Mb - MA_s\mathcal{S}(x)\|_2^2.\end{aligned} \tag{3}$$

Instead of using standalone data-domain sketches [12], our double-sketching framework is more effective in terms of dimensionality reduction and can be applied to generically accelerate PnP methods and also deep unrolling networks. When using the sketched loss in (3), we can have an approximate data fit that can be efficiently optimized by SGD [14] or its variance-reduced variants [7]. To recover the same reconstruction quality as the original program, we can adjust the image-domain sketch size m_0 stage-wise in a coarse-to-fine manner.

A. Doubly-Sketched PnP

We first apply our framework above to derive a plug-and-play multi-stage sketched gradient algorithm (PnP-MS2G). We present our PnP-MS2G in Algorithm 1, where we denote $\mathcal{D}$ as the denoiser, $\mathcal{S}$ as the image-domain sketching operator with varying sketch sizes, and $\mathcal{U}(\cdot) : \mathbb{R}^m \to \mathbb{R}^d$ as the upsampling operator which is an approximate inversion of the downsampler $\mathcal{S}$. To explain the motivation and derivation of Algorithm 1, we give here a concrete example with the squared error data fidelity. Noting that the PnP proximal gradient descent iteration can be written as:

$$x_{k+1} = \mathcal{D}[x_k - \eta \cdot (A^TAx_k - A^Tb)], \tag{4}$$

Algorithm 1 — Plug-and-Play with Multi-Stage Sketched Gradients (PnP-MS2G)

Initialization: $x_0 \in \mathbb{R}^d$, number of stages K, sketch-size $[m_1, ..., m_K]$, sketched forward operator $[A_{s_1}, ...A_{s_K}]$, sketching operators $[\mathcal{S}_1, ..., \mathcal{S}_K]$, up-sampling operators $[\mathcal{U}_1, ..., \mathcal{U}_K]$, number of inner-loops for each stage $[N_1, ..., N_K]$, step-size sequence $[\eta_1, ..., \eta_{\sum_{k=1}^K N_k}]$, $\alpha \in (0, 1]$, iteration counter $i = 0$
For $k = 1$ **to** K
  For $j = 1$ **to** N_k
    $i \leftarrow i + 1$
    Generate random subsampling mask M_i
    Compute the image-domain sketch: $v = \mathcal{S}_k(x_i)$
    Compute gradient estimate $G := \nabla_v f(M_i b, M_i A_{s_k} v)$
    $z_{i+1} = x_i - \eta_i \mathcal{U}_k G$,
    $x_{i+1} = (1 - \alpha)z_{i+1} + \alpha\mathcal{D}(z_{i+1})$,
  Endfor
Endfor
Output x_{i+1}

where $\mathcal{D}(\cdot)$ denotes the denoiser, which can be a denoising algorithm such as BM3D, or a classical proximal operator of some convex regularization (such as TV-prox), or a pre-trained denoising deep network such as (DnCNN [22]). Then we can perform one-side sketch first on the image domain and obtain a preliminary version of sketched gradient:

$$x_{k+1} = \mathcal{D}[x_k - \eta \cdot \mathcal{U}(A_s^TA_s\mathcal{S}(x_k) - A_s^Tb)], \tag{5}$$

where $\mathcal{U}(\cdot)$ denotes the upsamling operator. Numerically we found that off-the-shelf up/down-sampling algorithms such as the bi-cubic interpolation suffice to provide us with good estimates of the true gradients.

To further reduce the computational complexity, we can also utilize stochastic gradient estimate by applying the subsampling sketches on the measurement domain:

$$x_{k+1} = \mathcal{D}[x_k - \eta \cdot \mathcal{U}((M_kA_s)^TM_kA_s\mathcal{S}(x_k) - (M_kA_s)^TM_kb)] \tag{6}$$

where M_k is a random subsampling operator here for computing the stochastic gradient.

In Algorithm 1, we present our PnP-MS2G scheme, where we typically start by an aggressive sketch $\{A_{s_1}, \mathcal{S}_1\}$ with $m_1 \ll d$ for very fast initial convergence, and then for later stages we switch to medium size sketches $\{A_{s_k}, \mathcal{S}_k\}$ with $m_k < d$ which are increasingly more conservative, to reach a similar reconstruction accuracy as the unsketched counterpart. We also include a relaxation (optional) on the denoising step as suggested in [4].

B. Doubly-Sketched Deep Unrolling

Our multi-stage sketching scheme above can be easily incorporated within deep unrolling networks. Here we choose one of the state-of-the-art unrolling schemes, namely the learned primal-dual network [1]. It is based on unfolding the iteration of the Chambolle-Pock PDHG algorithm [3] by replacing the proximal operators $\text{prox}_{\sigma f^\star}(\cdot)$ and $\text{prox}_{\tau g}(\cdot)$ (here τ and σ are

TABLE I
LOW-DOSE CT TESTING RESULTS FOR LPD, LSPD, AND SKLSPD NETWORKS ON MAYO DATASET, WITH SUPERVISED TRAINING

METHOD	# CALLS ON A AND A^T	PSNR	SSIM	GPU TIME 1 PASS OF TEST SET
LPD	24	35.3177	0.9065	48.348
LSPD	6	35.0577	0.9014	31.196
SKLSPD	4	34.9749	0.9028	23.996

the primal and dual step-sizes) with multilayer convolutional neural networks $\mathcal{P}_{\theta_p}(\cdot)$ and $\mathcal{D}_{\theta_d}(\cdot)$, with sets of parameters θ_p and θ_d, applied on the both primal and dual spaces. The step sizes at each step are also set to be trainable. The learned primal-dual with K iterations can be written as the following, where the learnable parameters are $\{\theta_p^k, \theta_d^k, \tau_k, \sigma_k\}_{k=0}^{K-1}$:

Learned Primal-Dual (LPD) – Init. : $x_0 \in \mathbb{R}^d$ $y_0 \in \mathbb{R}^n$

For $k = 0, 1, 2, ..., K-1$

$$y_{k+1} = \mathcal{D}_{\theta_d^k}(y_k, \sigma_k, Ax_k, b);$$
$$x_{k+1} = \mathcal{P}_{\theta_p^k}(x_k, \tau_k, A^T y_{k+1});$$

By applying our sketching framework we can speedily approximate the products Ax_k, $A^T y_{k+1}$:

$$Ax_k \approx (M_i A_{s_k})\mathcal{S}(x_k), A^T y_{k+1} \approx \mathcal{U}((M_i A_{s_k})^T y_{k+1}). \quad (7)$$

We then apply dimensionality reduction using (7) to obtain the final network, referred to as Sketched LSPD (SKLSPD):

SkLSPD – Initialize $x_0 \in \mathbb{R}^d$ $y_0 \in \mathbb{R}^m$

For $k = 0, 1, 2, ..., K-1$

$$i = \mod(k, n/m);$$
(or pick i from $[0, n/m - 1]$ uniformly at random)
$$y_{k+1} = \mathcal{D}_{\theta_d^k}(y_k, \sigma_k, (M_i A_{s_k})\mathcal{S}(x_k), M_i b);$$
$$x_{k+1} = \mathcal{P}_{\theta_p^k}(x_k, \tau_k, \mathcal{U}((M_i A_{s_k})^T y_{k+1}));$$

Alternatively, we can also rearrange the last line as:

$$x_{k+1} = \mathcal{U}(\mathcal{P}_{\theta_p^k}(\mathcal{S}(x_k), \tau_k, (M_i A_{s_k})^T y_{k+1})) \quad (8)$$

for further improved numerical efficiency in practice.

In this work, for the sake of numerical comparison, we also implement a special case of SkLSPD which does not perform image-domain sketches (i.e. $\mathcal{U} = \mathcal{S} = I$), referred to as the LSPD (Learned Stochastic Primal-Dual) network.

III. NUMERICAL RESULTS

Sketched PnP. We start by presenting some numerical results for applying our sketching framework to accelerating PnP algorithms. We consider sparse-view CT reconstruction and compare our PnP-MS2G with the PnP stochastic gradient descent method [17], which has already been found to be superior over deterministic PnP schemes. For our PnP-MS2G we perform a $4\times$ downscale at the first 50 iterations, then a $2\times$ downscale afterward, leading to significant improvement in computational efficiency. We choose here the BM3D [6] as the denoiser. In the numerical results reported in Figure 1, we found that even in terms of the iteration counts, PnP-MS2G can achieve an improvement in the convergence speed, while additionally being more efficient per iteration compared to PnP-SGD. We believe that this is because PnP-MS2G prioritizes reconstructing the low-frequency component of the image in the early iterations since the gradient is taken on the low-dimensional image space, which leads to an algorithmic warm-start as a by-product.

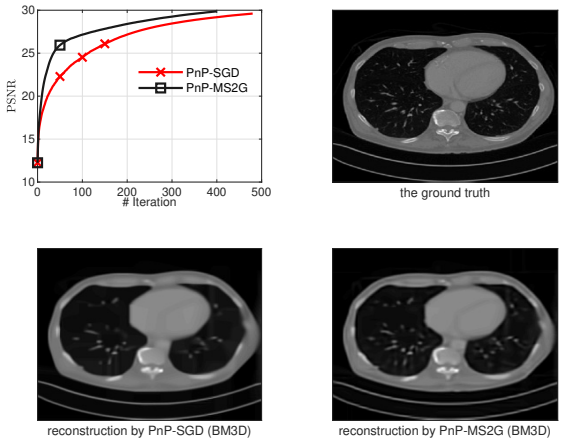

Fig. 1. Example for applying MS2G (minibatched) in X-ray CT reconstruction, comparing to the PnP stochastic gradient descent (PnP-SGD) proposed by [17]. Note that each iteration of PnP-MS2G is more computationally efficient than PnP-SGD due to the dimensionality reduction by our double operator sketching. Surprisingly, even in terms of iteration-number the PnP-MS2G can provide better convergence rate comparing to PnP-SGD.

Sketched deep unrolling. We then present the experimental performance of our proposed networks for low-dose X-ray CT. In real-world clinical practice, the low dosage CT is widely used and highly recommended, since the intense exposures to the X-ray could significantly increase the risk of inducing cancers. The low-dose CT takes a large number of low-energy X-ray views, leading to huge volumes of noisy measurements. This makes the reconstruction schemes struggle to achieve efficient and accurate estimations. In our X-ray CT experiments, we use the standard Mayo-Clinic dataset [11] which contains 10 patients' 3D CT scans. We use 2111 slices (from nine patients) of 2D images sized 512×512 for training and 118 slices of the remaining one patient for testing. We use the ODL toolbox [1] to simulate fan beam projection data with 800 equally-spaced angles of view (each view includes 800 rays). The fan-beam CT measurement is corrupted with Poisson noise: $b \sim \text{Poisson}(I_0 e^{-Ax^\dagger})$, where we make a low-dose choice of $I_0 = 7 \times 10^4$. We use the Beer-Lambert law to simulate the noisy projection data and to linearize the measurements, we consider the log data.

In our LSPD and SkLSPD networks, we interleave-partition (according to the angles) the forward/adjoint operators and

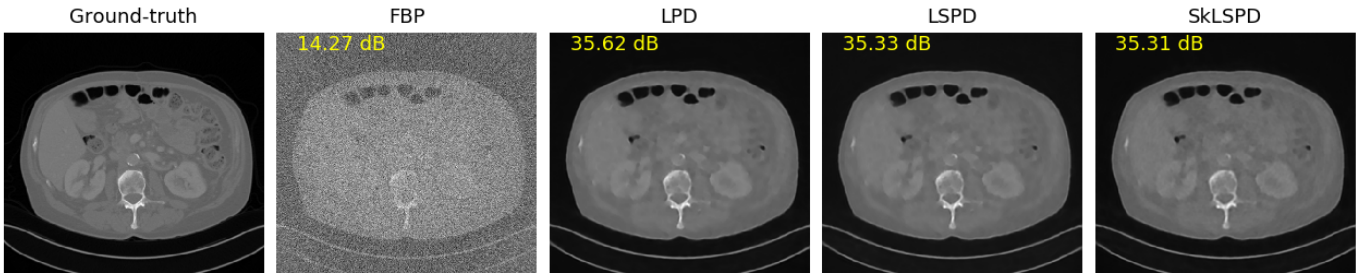

Fig. 2. Examples for Low-dose CT on the test set of Mayo dataset. We can observe that our LSPD and SkLSPD networks achieve similar reconstruction performance as the full-batch LPD.

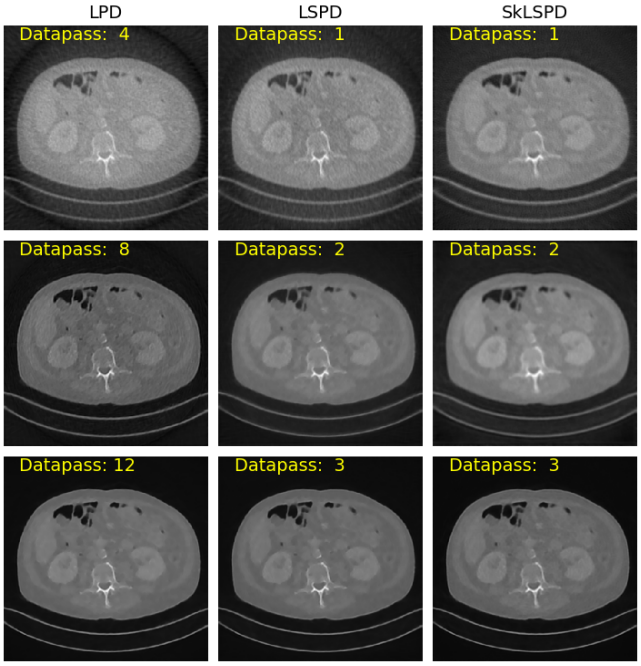

Fig. 3. Example for intermediate layer outputs for Low-dose CT on the test set of Mayo dataset. We can observe that LSPD/SkLSPD achieves competitive reconstruction quality with LPD across intermediate layers (the first row is for the 4th iteration/layer, the second row is for the 8th, while the third row is for the 12th).

data into $m = 4$ subsets. Our networks have $K = 12$ layers[1] hence correspond to three data-passes, which means it takes only three calls in total on the forward and adjoint operators. We compare it with the learned primal-dual (LPD) which has $K = 12$ layers, corresponding to 12 calls on the forward and adjoint operator. We train all the networks with 50 epochs of Adam optimizer [8] with batch size one, in the supervised manner.

For all these networks, we choose the subnetworks $\mathcal{P}_{\theta_k}$ and $\mathcal{D}_{\theta_k}$ to have three convolutional layers (with a skip connection between the first channel of input and the output) and 32 channels, with kernel size 5. The starting point x_0 is set to be the standard filtered-back projection for all the unrolling networks. We set all of them to have 12 algorithmic layers ($K = 12$). For the up/down-samplers in our Sketched LSPD, we choose the bilinear upsample and downsample functions in Pytorch. The up-sampler increases the input image 4 times (from 256×256 to 512×512), while the down-sampler makes the input image 4 times smaller (from 512×512 to 256×256). While the full forward operator A is defined on the grid of 512×512, the sketched operator A_s is defined on the grid of 256×256 and hence requires only half of the computation in this setting. We use the coarse-to-fine strategy for SkLSPD, where we double-sketch the first 8 layers, but leave the last four layers to have only measurement subsampling and no downsampling in the image and dual spaces.

We present the performance of the LPD, LSPD, and SkLSPD on the test set in Table 1, and some illustrative examples in Figure 2 and 3 for a visual comparison. From the numerical results, we found out that our LSPD and SkLSPD networks both achieve almost the same reconstruction accuracy compared to the LPD baseline in terms of PSNR (peak signal-to-noise ratio) and SSIM measures, requiring only a fraction of the computation of the forward and adjoint operators. In terms of run time on GPU, our methods achieve a reduction in reconstruction time of around 40% to 60% compared to the full batch LPD.

[1]each layer of LSPD includes a primal and a dual subnetwork with three convolutional layers with kernel size 5×5 and 32 channels, same for LPD.

IV. Conclusions

We proposed a generic acceleration framework for deriving efficient IDR schemes, such as PnP and deep unrolling networks. We presented several important instances such as PnP-MS2G, LSPD, and SkLSPD, and showed that the proposed methods achieved high-quality reconstruction under significant dimensionality reduction, demonstrating the promising potential of applicability in large-scale inverse problems. Meanwhile, we believe that our framework can also be effectively applied to accelerate MCMC schemes for uncertainty quantification in imaging, such as PnP-ULA [9], and NF-ULA [2] which leverages generative models.

REFERENCES

[1] J. Adler and O. Öktem, "Learned primal-dual reconstruction," *IEEE transactions on medical imaging*, vol. 37, no. 6, pp. 1322–1332, 2018.

[2] Z. Cai, J. Tang, S. Mukherjee, J. Li, C.-B. Schönlieb, and X. Zhang, "Nf-ula: Normalizing flow-based unadjusted langevin algorithm for imaging inverse problems," *SIAM Journal on Imaging Sciences*, vol. 17, no. 2, pp. 820–860, 2024.

[3] A. Chambolle and T. Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging," *Journal of mathematical imaging and vision*, vol. 40, no. 1, pp. 120–145, 2011.

[4] R. Cohen, M. Elad, and P. Milanfar, "Regularization by denoising via fixed-point projection (red-pro)," *SIAM Journal on Imaging Sciences*, vol. 14, no. 3, pp. 1374–1406, 2021.

[5] P. L. Combettes and J.-C. Pesquet, "Proximal splitting methods in signal processing," in *Fixed-point algorithms for inverse problems in science and engineering*. Springer, 2011, pp. 185–212.

[6] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3-d transform-domain collaborative filtering." *IEEE transactions on image processing: a publication of the IEEE Signal Processing Society*, vol. 16, no. 8, pp. 2080–2095, 2007.

[7] R. Johnson and T. Zhang, "Accelerating stochastic gradient descent using predictive variance reduction," in *Advances in neural information processing systems*, 2013, pp. 315–323.

[8] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Proceedings of 3rd International Conference on Learning Representations*, 2015.

[9] R. Laumont, V. D. Bortoli, A. Almansa, J. Delon, A. Durmus, and M. Pereyra, "Bayesian imaging using plug & play priors: when langevin meets tweedie," *SIAM Journal on Imaging Sciences*, vol. 15, no. 2, pp. 701–737, 2022.

[10] J. Liu, Y. Sun, W. Gan, X. Xu, B. Wohlberg, and U. S. Kamilov, "Sgd-net: Efficient model-based deep learning with theoretical guarantees," *IEEE Transactions on Computational Imaging*, vol. 7, pp. 598–610, 2021.

[11] C. McCollough, "Tu-fg-207a-04: Overview of the low dose ct grand challenge," *Medical physics*, vol. 43, no. 6Part35, pp. 3759–3760, 2016.

[12] M. Pilanci and M. J. Wainwright, "Randomized sketches of convex programs with sharp guarantees," *Information Theory, IEEE Transactions on*, vol. 61, no. 9, pp. 5096–5115, 2015.

[13] E. T. Reehorst and P. Schniter, "Regularization by denoising: Clarifications and new interpretations," *IEEE Transactions on Computational Imaging*, vol. 5, no. 1, pp. 52–67, 2018.

[14] H. Robbins and S. Monro, "A stochastic approximation method," *The Annals of Mathematical Statistics*, vol. 22, no. 3, pp. 400–407, 1951.

[15] Y. Romano, M. Elad, and P. Milanfar, "The little engine that could: Regularization by denoising (red)," *SIAM Journal on Imaging Sciences*, vol. 10, no. 4, pp. 1804–1844, 2017.

[16] J. Sun, H. Li, Z. Xu *et al.*, "Deep admm-net for compressive sensing mri," *Advances in neural information processing systems*, vol. 29, 2016.

[17] Y. Sun, B. Wohlberg, and U. S. Kamilov, "An online plug-and-play algorithm for regularized image reconstruction," *IEEE Transactions on Computational Imaging*, 2019.

[18] H. Y. Tan, S. Mukherjee, J. Tang, and C.-B. Schönlieb, "Provably convergent plug-and-play quasi-newton methods," *SIAM Journal on Imaging Sciences*, vol. 17, no. 2, pp. 785–819, 2024.

[19] J. Tang, "Accelerating plug-and-play image reconstruction via multi-stage sketched gradients," *arXiv preprint arXiv:2203.07308*, 2022.

[20] J. Tang, S. Mukherjee, and C.-B. Schönlieb, "Accelerating deep unrolling networks via dimensionality reduction," *arXiv preprint arXiv:2208.14784*, 2022.

[21] S. V. Venkatakrishnan, C. A. Bouman, and B. Wohlberg, "Plug-and-play priors for model based reconstruction," in *2013 IEEE Global Conference on Signal and Information Processing*. IEEE, 2013, pp. 945–948.

[22] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3142–3155, 2017.