Edinburgh Research Explorer

A Fast Stochastic Plug-and-Play ADMM for Imaging Inverse Problems

Link:
Link to publication record in Edinburgh Research Explorer

OPEN ACCESS

A Fast Stochastic Plug-and-Play ADMM for Imaging Inverse Problems

Junqi Tang, Mike Davies, *Fellow, IEEE*

Abstract—In this work we propose an efficient stochastic plug-and-play (PnP) algorithm for imaging inverse problems. The PnP stochastic gradient descent methods have been recently proposed and shown improved performance in some imaging applications over standard deterministic PnP methods. However, current stochastic PnP methods need to frequently compute the image denoisers which can be computationally expensive. To overcome this limitation, we propose a new stochastic PnP-ADMM method which is based on introducing stochastic gradient descent inner-loops within an inexact ADMM framework. We provide the theoretical guarantee on the fixed-point convergence for our algorithm under standard assumptions. Our numerical results demonstrate the effectiveness of our approach compared with state-of-the-art PnP methods.

Index Terms—Stochastic ADMM, Plug-and-Play Priors.

I. INTRODUCTION

Recent trends in the research of computational imaging have been focusing on developing algorithms which are able to jointly utilize the power of classical physical models and advanced image priors [1]–[5]. These methods typically take the form of well-known optimization algorithms, and plug in a pretrained deep neural network [6] or a patch-based denoiser with non-local denoising properties [7]–[10]. In this work we propose a novel stochastic plug-and-play method for imaging inverse problems. Consider the following observation model for a linear inverse problem:

$$b = Ax^\dagger + w, \quad A \in \mathbb{R}^{n \times d} \quad (1)$$

where $x^\dagger$ denotes the vectorized (raster) ground truth image, $A \in \mathbb{R}^{n \times d}$ represents the forward measurement model, $b \in \mathbb{R}^n$ denotes the observation, while $w \in \mathbb{R}^n$ represents the random additive noise. Traditionally, in order to get a good estimate of the ground truth $x^\dagger$, we typically seek to find the minimizer of a composite objective function:

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

where $f(x) = \frac{1}{n}\sum_{i=1}^n f_i(x) = \frac{1}{n}\sum_{i=1}^n f(a_i, b_i, x)$ is the data fidelity term which is assumed to be convex and smooth, such as the least-square loss $f(x) = \frac{1}{2n}\|Ax - b\|_2^2$, and here we denote a_i as the i-th row of A and b_i as the i-th element of b. Meanwhile $g(x)$ in (2) denotes a regularization term which encodes image priors, with classical examples including the sparsity-inducing regularization in wavelet domain and the total-variation regularization [11], etc. The composite loss

J.Tang and M.Davies are with the School of Engineering, University of Edinburgh. Correspondence to J.Tang@ed.ac.uk

function (2) can be effectively minimized via a class of iterative algorithms which are known as the proximal splitting methods [12], including the forward-backward splitting [13]–[15], primal-dual splitting [16]–[18] and the Douglas-Rachfort splitting/alternating direction method of multipliers (ADMM) [19]–[21], etc.

Considering the link between proximal operators and denoising, researchers [1]–[3] have discovered that if they simply replace the proximal operator on g with a direct call to an off-the-shelf denoising algorithm, such as NLM [7], TNRD [9], BM3D [8], or the DnCNN [6], excellent image recovery/reconstruction results can often be attained. Although the imaging community has very limited theoretical understanding and convergence analysis for such algorithms so far, these ad-hoc approaches have shown state-of-the-art performances in various imaging applications.

Recently, inspired by the success of stochastic gradient descent (SGD) methods in solving large-scale optimization tasks in machine learning [22]–[24] and some imaging applications [25]–[27], Sun et al [28] have extended the deterministic plug-and-play ISTA/FISTA method [3] and proposed PnP-SGD in order to improve the computational efficiency. In each iteration of PnP-SGD, a minibatch stochastic gradient is computed as an unbiased estimator of the full gradient, which yields a computational benefit in each iteration. However, as discussed in [29], current stochastic gradient methods in general need to compute the proximal operator/denoisers more frequently than the deterministic gradient methods within the same amount of gradient evaluations. When the denoiser is computationally expensive, the actual performance benefit of using stochastic gradient techniques may be compromised due to this computational overhead. In this work, we seek to address this issue for stochastic PnP methods and propose a more practical approach for utilizing the power of stochastic gradient techniques to accelerate the deterministic plug-and-play algorithms.

A. Main contributions

This paper's contribution is two-fold:

- We propose an efficient stochastic PnP method which empirically improves upon previous stochastic approach in [28] by reducing the number of calls on the modern denoisers which are usually a bottleneck for computation. We demonstrate the effectiveness of our approach in X-ray computed tomography (CT) imaging problems.

- We provide a theoretical fixed-point convergence analysis for our stochastic PnP algorithm under standard assumptions.

II. STOCHASTIC PNP-ADMM

In a recent work of Sun et al [28], a stochastic PnP algorithm is proposed for image restoration and reconstruction, which can be written as the following:

PnP-SGD – Initialize $x^0, z^0 = x^0$

For $k = 1, 2, ..., t$
$$
\begin{aligned}
x^k &= \mathcal{D}[z^{k-1} - \eta \cdot \nabla f_{S_k}(z^{k-1})] \\
z^k &= x^k + \alpha^k(x^k - x^{k-1})
\end{aligned}
$$

where $\nabla f_{S_k}(x^{k-1})$ denotes a minibatch stochastic gradient with a randomly subsampled index S_k, chosen uniformly at random from a partitioned index $\hat{\mathcal{I}} = \{\mathcal{I}_1, \mathcal{I}_2, ..., \mathcal{I}_K\}$, where $\mathcal{I}_1 \cup \mathcal{I}_2 \cup ... \cup \mathcal{I}_K = [n]$ and $\mathcal{I}_i \cap \mathcal{I}_j = \emptyset, \forall i \neq j \in [K]$. The PnP-SGD algorithm is essentially a plug-and-play variant of the stochastic proximal gradient descent [30] which is based on the forward-backward splitting [13]. In each iteration of PnP-SGD, a minibatch stochastic gradient estimate is computed:

$$
\nabla f_{S_k}(z^k) = \frac{1}{m} \sum_{i \in S_k} \nabla f_i(z^k), \tag{3}
$$

where $m = \frac{n}{K}$. It first performs a stochastic gradient descent step with a step-size η, then a denoising step is computed using an off-the-shelf denoiser [6]–[9] denoted as $\mathcal{D}(\cdot)$. Finally a momentum step is performed for empirical convergence acceleration with a momentum parameter α_k as in the FISTA algorithm [14]. The computational benefit of PnP-SGD over its deterministic counterparts (PnP-ISTA/PnP-FISTA [3]) comes from using an approximation of the full gradient by the minibatch gradient (3) which can be efficiently computed. However, the PnP-SGD does not have the capability to reduce the cost of computing the denoising step – the denoiser has to be called at each iteration. To overcome this computational bottleneck, one plausible approach is to decouple the gradient step and the denoising step via Douglas-Rachford splitting/ADMM instead of the forward-backward splitting. In this work we study and propose a stochastic gradient extension of the PnP-ADMM algorithm[1] [1]:

PnP-ADMM – Initialize $x^0 = z^0 \in \mathbb{R}^d$;

For $k = 0, 2, ..., t$
$$
\begin{aligned}
y^k &= \text{prox}_{\tau f}[z^k] \\
x^{k+1} &= \mathcal{D}[2y^k - z^k] \\
z^{k+1} &= z^k + x^{k+1} - y^k.
\end{aligned}
$$

where in each iteration an exact proximal step on the data-fidelity term $f(x)$ is computed with a constant step-size τ:

$$
\text{prox}_{\tau f}(\cdot) = \arg\min_x \frac{1}{2}\|x - \cdot\|_2^2 + \tau f(x). \tag{4}
$$

[1]We write the update rule of PnP-ADMM in this paper using the equivalent Douglas-Rachford splitting reformulation [31, Section 9.1] for the simplicity of notation in analysis.

Algorithm 1 — Stochastic PnP-ADMM

Initialization: number of inner-iterations: $[N_1, N_2, ..., N_K]$, momentum parameter sequence: $[\alpha_1, \alpha_2, ..., \alpha_{\max_{j \in [K]} N_j}]$, partition index $\hat{\mathcal{I}} = \{\mathcal{I}_1, \mathcal{I}_2, ..., \mathcal{I}_K\}$, $z^0 \in \mathbb{R}^d$, $y_0^0 \in \mathbb{R}^d$.

for $k = 1$ to K **do**
    for $j = 1$ to N_k **do**
        Randomly sample $S_j \in \hat{\mathcal{I}}$ with replacement.
        Compute a stochastic gradient estimator $\nabla f_{S_j}(y_{j-1}^k)$
        $v_j^k = y_{j-1}^k - \eta_k \cdot [\tau \nabla f_{S_j}(y_{j-1}^k) + y_{j-1}^k - z^k]$
        Momentum: $y_j^k = v_j^k + \alpha_j(v_j^k - v_{j-1}^k)$
    end for
    $x^{k+1} = \mathcal{D}(2y_{N_k}^k - z^k)$
    $z^k = z^{k-1} + x^{k+1} - y_{N_k}^k$
    $y_0^{k+1} = x^{k+1}$
end for
Output x^K

In classical ADMM the step-size τ can be any positive constant to ensure convergence. The update rule of the PnP-ADMM can be written as $z^{k+1} = T(z^k)$, where $T(\cdot)$ is an operator defined as [31]:

$$
T = \frac{1}{2}I + \frac{1}{2}(2\mathcal{D} - I)(2\text{prox}_{\tau f} - I). \tag{5}
$$

Our proposed solution, presented in algorithm 1, is to use SGD with momentum to *approximately* solve the prox step (4) within PnP-ADMM framework. We denote the number of inner-iterations at the k-th outerloop as N_k. In each inner-iteration a stochastic gradient descent step is performed with a step-size η_k, and then followed by a momentum step for empirical acceleration. Unlike the PnP-SGD which needs to call the denoiser in every stochastic gradient descent iteration, the proposed method only needs to compute the denoiser once every N_k iterations. Our theoretical analysis is restricted to the case where we set $\alpha_j = 0$ and the parameters N_k and η_k are chosen adaptively in each outer-iteration. However, in practice, a constant step size η which is inversely proportional to the Lipschitz constant $\eta_k = O(\frac{1}{L})$, $\tau = O(1)$, a constant number of inner-iterations $N_k = O(K)$, and a FISTA-like momentum parameter $\alpha_j = \frac{j-1}{j+3}$ [17] are suggested for good empirical performance.

III. CONVERGENCE ANALYSIS

In this section we provide theoretical analysis for our stochastic PnP-ADMM. When we run a stochastic gradient-based innerloop, we are effectively making an approximation of the proximal step $y^k = \text{prox}_{\tau f}[z^k]$, hence we can write our stochastic PnP-ADMM algorithm as the inexact recursion:

$$
z^{k+1} = T(z^k) + \varepsilon^k. \tag{6}
$$

where ε^k denotes the approximation error. Now the desired fixed-point convergence can be established for (6), if we make the following standard assumptions as in [31] on the denoiser and the data-fidelity term.

A. Generic Assumptions

A. 1: The denoiser satisfies:

$$\|(\mathcal{D} - I)(x) - (\mathcal{D} - I)(y)\|_2 \leq \beta \|x - y\|_2, \ \forall x, y \in \mathbb{R}^d, \quad (7)$$

with $\beta > 0$.

It is easy to show that A.1 implies a relaxed non-expansiveness condition on $\mathcal{D}(\cdot)$ which reads $\|\mathcal{D}(x) - \mathcal{D}(y)\|_2 \leq (1 + \beta)\|x - y\|_2$ and is satisfied for a wide class of modern denoisers such as NLM and properly trained DnCNNs [31].

A. 2: $f(\cdot)$ is μ-strongly-convex:

$$f(x) - f(y) - \langle \nabla f(y), x - y \rangle \geq \mu \|x - y\|_2^2, \ \forall x, y \in \mathbb{R}^d, \quad (8)$$

with $\mu > 0$. Meanwhile for a given minibatch partition index $\hat{\mathcal{I}} = \{\mathcal{I}_1, \mathcal{I}_2, ..., \mathcal{I}_K\}$ such that $f(\cdot) = \frac{1}{K} \sum_{k=1}^{K} f_{\mathcal{I}_k}(\cdot)$, each $f_{\mathcal{I}_k}$ is L-smooth, such that $\forall x, y \in \mathbb{R}^d$:

$$f_{\mathcal{I}_k}(x) - f_{\mathcal{I}_k}(y) - \langle \nabla f_{\mathcal{I}_k}(y), x - y \rangle \leq L\|x - y\|_2^2. \quad (9)$$

The strong-convexity assumption is necessary for our analysis. It seems pessimistic since a number of imaging inverse problems do not have strong-convexity. We believe that the assumption on strong-convexity could be relaxed, e.g. following the ideas from [32]–[36]. On the other hand, one can instead run Algorithm 1 on a regularized objective $\hat{f}(x) = f(x) + \frac{\epsilon}{2}\|x\|_2^2$ to manually enforce strong-convexity, which is a classical trick in convex optimization [37]. Nevertheless, we believe that relaxing this assumption is an important future direction for the analysis.

B. Analysis

We first apply an existing convergence result for SGD for establishing the approximation accuracy of the inner-loop:

Lemma III.1: Under Assumption A.2, denote that for k-th outer-loop of Stochastic PnP-ADMM $y_\star^k = \text{prox}_{\tau f}[z^k]$, and define the following quantities for each outer-iteration k:

$$\sigma_k^2 := \mathbb{E}_q[\|\tau \nabla f_{\mathcal{I}_q}(y_\star^k) + y_\star^k - z^k\|_2^2],$$
$$\xi_k := \|y_\star^k - x^k\|_2^2, \quad (10)$$

then if the step size $\eta_k = \frac{\mu_0 \varepsilon}{2\varepsilon \mu_0 L_0 + 2\sigma_k^2}$, $\alpha_j = 0$ for all j, $N_k = 2\log(\frac{\xi_k}{\varepsilon})(\frac{L_0}{\mu_0} + \frac{\sigma_k^2}{\mu_0^2 \varepsilon})$ with $\mu_0 = \tau\mu + 1$, $L_0 = \tau L + 1$, then we have the approximation error of the proximal step bounded as:

$$\mathbb{E}\|y_{N_k}^k - \text{prox}_{\tau f}[z^k]\|_2^2 \leq \varepsilon, \quad (11)$$

where the expectation is taken over the random sampling of the indices within the inner-loop.

Proof. We first observe that the proximal step $y_\star^k = \text{prox}_{\tau f}[z^k]$ can be written precisely as a finite-sum optimization problem of the follow form:

$$\text{prox}_{\tau f}(z^k) = \arg\min_x \frac{1}{2}\|x - z^k\|_2^2 + \tau f(x)$$
$$= \arg\min_x \frac{1}{K}\sum_{q=1}^{K}[\tau f_{\mathcal{I}_q}(x) + \frac{1}{2}\|x - z^k\|_2^2], \quad (12)$$

which is a $(\tau\mu + 1)$-strongly-convex objective and each of the element in the sum is $(\tau L + 1)$-smooth.

According to [38, Theorem 2.1], if we run SGD (starting at $x^k \in \mathbb{R}^d$) with uniform random sampling and a step size $\frac{\mu_0 \varepsilon}{2\varepsilon \mu_0 L_0 + 2\sigma_k^2}$, then after $N = 2\log(\frac{\|y_\star^k - x^k\|_2}{\varepsilon})(\frac{L_0}{\mu_0} + \frac{\sigma_k^2}{\mu_0^2 \varepsilon})$ iterations, we have $\mathbb{E}\|y_{N_k}^k - y_\star^k\|_2^2 \leq \varepsilon$. $\qquad\square$

Now, we are able to prove the fixed-point convergence for the inexact recursion (6), and hence for our proposed method.

Theorem III.2: Assume A.1 and A.2 with $\beta < 1$, denote positive values σ and ξ such that $\forall k$, $\sigma_k \leq \sigma$ and $\xi_k \leq \xi$, and the quantities $\mu_0, L_0, \sigma_k^2, \xi_k$ are defined as in Lemma III.1. If we choose the step-size parameters as $\tau > 1/(1 + \beta - 2\beta^2)$, $\alpha_j = 0$, $\eta_k = \frac{\mu_0}{2\mu_0 L_0 + 2k\sigma^2}$, $N_k = 2\log(k\xi)(\frac{L_0}{\mu_0} + \frac{k\sigma^2}{\mu_0^2})$, we have the following fixed-point convergence for Algorithm 1:

$$\mathbb{E}\|z^{k+1} - z^k\|_2 \to 0, \quad (13)$$

when $k \to +\infty$.

Our main theorem suggests that for the basic form of Algorithm 1 where we choose the momentum $\alpha_j = 0$, the outerloop step-size $\tau = O(1)$, the inner-loop step-size decreasing in k and the number of inner-iterations increasing in k, Algorithm 1 is guaranteed to converge to a fix point. However our numerical results in section IV suggest that we may set the number of inner-loop N_k and step size η_k to be constant and use FISTA-type of momentum [14], [15], [17] for good empirical performance in practice.

C. Proof for Theorem III.2

Firstly, due to assumption A.1 we have:

$$\beta\|x - y\|_2 \geq \|(\mathcal{D} - I)(x) - (\mathcal{D} - I)(y)\|_2$$
$$\geq \|\mathcal{D}(x) - \mathcal{D}(y)\|_2 - \|x - y\|_2, \ \forall x, y \in \mathbb{R}^d, \quad (14)$$

hence $\|\mathcal{D}(x) - \mathcal{D}(y)\|_2 \leq (1 + \beta)\|x - y\|_2, \ \forall x, y \in \mathbb{R}^d$. Denote $u_k = \text{prox}_{\tau f}(z^k) - y_{N_k}^k = y_\star^k - y_{N_k}^k$, we have:

$$\|\varepsilon^k\|_2 = \|z^{k+1} - T(z^k)\|_2$$
$$\leq \|u_k\|_2 + \|\mathcal{D}(2y_{N_k}^k - z^k) - \mathcal{D}(2y_{N_k}^k - z^k + 2u_k)\|_2$$
$$\leq (3 + 2\beta)\|u_k\|_2. \quad (15)$$

Applying Lemma III.1 gives $\mathbb{E}\|\varepsilon^k\|_2 \leq \frac{3 + 2\beta}{k}$. Now according to [31, Theorem 2], under assumption A.1 and A.2, we can ensure that:

$$\|T(x) - T(y)\|_2 \leq \delta\|x - y\|_2, \ \forall x, y \in \mathbb{R}^d, \quad (16)$$

where $\delta = \frac{1 + \beta + \beta\tau\mu + 2\beta^2\tau\mu}{1 + \tau\mu + 2\beta\tau\mu}$. Moreover, if $\tau > 1/(1 + \beta - 2\beta^2)$ and $\beta < 1$, then $\delta < 1$. Hence we have:

$$\mathbb{E}\|z^{k+1} - z^k\|_2 = \mathbb{E}\|T(z^k) - T(z^{k-1}) + \varepsilon^k - \varepsilon^{k-1}\|_2$$
$$\leq \mathbb{E}\|T(z^k) - T(z^{k-1})\|_2$$
$$+ \mathbb{E}\|\varepsilon^k\|_2 + \mathbb{E}\|\varepsilon^{k-1}\|_2$$
$$\leq \delta\mathbb{E}\|z^k - z^{k-1}\|_2 + \frac{3 + 2\beta}{k} + \frac{3 + 2\beta}{k - 1}$$
$$\leq \delta\mathbb{E}\|z^k - z^{k-1}\|_2 + \frac{15}{k} \quad (17)$$

If we recursively apply the same argument we will get:

$$\mathbb{E}\|z^{k+1} - z^k\|_2 \leq \delta^k \mathbb{E}\|z^1 - z^0\|_2 + \frac{15}{k}\sum_{i=0}^{k-1}\frac{k\delta^i}{k - i}. \quad (18)$$
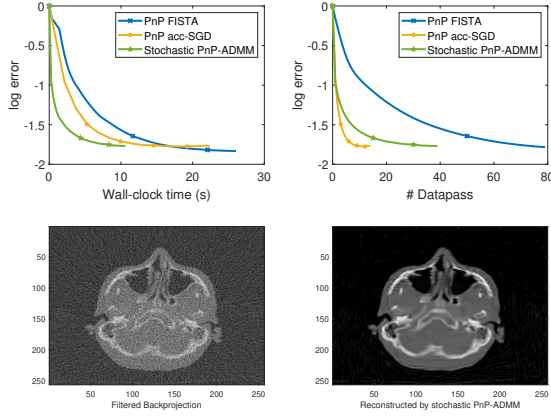
Fig. 1: The estimation error plot of the compared algorithms on low-dose CT example

Fig. 2: The estimation error plot of the compared algorithms on sparse-view CT example

Then we use a classic criterion to show the boundedness of series $\sum_{i=0}^{+\infty} v_i$ where $v_i = \frac{k\delta^i}{k-i}$. For any finite $p > 1$, we have:

$$\lim_{i \to +\infty} i^p v_i = \lim_{i \to +\infty} \frac{\delta^i i^p k}{k-i} = 0, \tag{19}$$

where we take $k \to +\infty$ and $i \le k - 1$, and then:

$$\sum_{i=0}^{k-1} \frac{k\delta^i}{k-i} < +\infty, \qquad \frac{15}{k} \sum_{i=0}^{k-1} \frac{k\delta^i}{k-i} \to 0. \tag{20}$$

Hence by taking $k \to +\infty$, we have $\mathbb{E}\|z^{k+1} - z^k\|_2 \to 0$. Thus finishes the proof for Theorem III.2.

IV. NUMERICAL EXPERIMENTS

For our numerical experiments, we choose the X-ray CT imaging as an example since it is know to favor stochastic gradient methods [29]. We compare our algorithm with the state-of-the-art stochastic PnP method with momentum acceleration proposed by Sun et al [28], as well as the PnP FISTA algorithm [3]. We use MATLAB R2018a in a machine with 1.6 GB RAM, 1.80 GHz Intel Core i7-8550 CPU.

We first test the compared methods on low-dose CT imaging problems, where low-energy noisy X-ray measurements with $I_0 = 10^3$ are used, which demands strong image priors are used in order to achieve good-quality reconstructions. Meanwhile we also compare these algorithms in sparse-view CT imaging with $I_0 = 10^4$, where fewer X-ray measurements are taken compared to the number of pixels to be inferenced[2]. For low-dose CT example, we choose the penalized weighted least-squares objective as the data-fidelity term, which is tailored for low-dose CT [39]. For sparse-view CT example, we choose the standard least-squares loss as the data fidelity term. For the randomized methods we partition the data into 10 minibatches. The noisy CT observations are obtained via $y \sim \text{Poisson}(I_0 e^{-Ax})$ where the forward operator A is implemented using the *AIRtools* package [40]. For our algorithm,

we set $N_j = 10$ for all j such that in each inner-loop we make exactly one pass of the data, outer-loop step-size $\tau = 1$, inner-loop step-size $\eta_k = \frac{1}{L}$, and the momentum parameter $\alpha_j = \frac{j-1}{j+3}$ as suggested in [17].

We choose the BM3D [8] with the denoiser-scaling [41] as the denoiser:

$$\mathcal{D}_\gamma(x) = \frac{1}{\gamma} \text{BM3D}(\gamma x), \tag{21}$$

and maximize the reconstruction performance for each of the compared algorithms via grid-searching the parameter γ.

We present the numerical results of the algorithms in Figure 1 for low-dose CT inverse problem of size $A \in \mathbb{R}^{88256 \times 65536}$, and in Figure 2 for sparse-view CT imaging task of size $A \in \mathbb{R}^{92160 \times 262144}$. We plot the estimation error $\log_{10} \|x - x^\dagger\|_2$ to the ground-truth image $x^\dagger$, against the actual run time as well as the number of datapasses. We can observe that both PnP-SGD and our method are much faster than PnP-FISTA in terms of datapass. The PnP-SGD appears to be faster than our method in terms of number of datapasses. However, in terms of actual run time, the PnP-SGD is slower than our method, due to the need to compute the costly BM3D at each stochastic gradient iteration.

V. CONCLUSION

In this work we propose a stochastic PnP-ADMM algorithm which is able to provide practical acceleration with stochastic gradient techniques, for efficiently solving imaging inverse problems. This is an effective approach to make the stochastic PnP schemes truly practical, by reducing the computational overhead of the modern denoisers. We provide a fixed-point convergence analysis, and demonstrate the effectiveness of our method in numerical experiments.

---

[2]The numerical result in this example suggests that empirically the strong-convexity is not needed for the stochastic PnP-ADMM to converge.
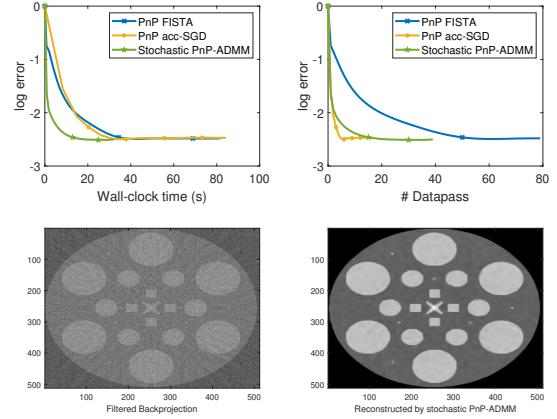
References

[1] 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.

[2] K. Egiazarian, A. Foi, and V. Katkovnik, "Compressed sensing image reconstruction via recursive spatially adaptive filtering," in *2007 IEEE International Conference on Image Processing*, vol. 1. IEEE, 2007, pp. I–549.

[3] U. S. Kamilov, H. Mansour, and B. Wohlberg, "A plug-and-play priors approach for solving nonlinear imaging inverse problems," *IEEE Signal Processing Letters*, vol. 24, no. 12, pp. 1872–1876, 2017.

[4] 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.

[5] 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.

[6] 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.

[7] A. Buades, B. Coll, and J.-M. Morel, "A non-local algorithm for image denoising," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 2. IEEE, 2005, pp. 60–65.

[8] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image restoration by sparse 3d transform-domain collaborative filtering," in *Image Processing: Algorithms and Systems VI*, vol. 6812. International Society for Optics and Photonics, 2008, p. 681207.

[9] Y. Chen and T. Pock, "Trainable nonlinear reaction diffusion: A flexible framework for fast and effective image restoration," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 6, pp. 1256–1272, 2017.

[10] H. Talebi and P. Milanfar, "Global image denoising," *IEEE Transactions on Image Processing*, vol. 23, no. 2, pp. 755–768, 2013.

[11] A. Chambolle and T. Pock, "An introduction to continuous optimization for imaging," *Acta Numerica*, vol. 25, pp. 161–319, 2016.

[12] 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.

[13] P.-L. Lions and B. Mercier, "Splitting algorithms for the sum of two nonlinear operators," *SIAM Journal on Numerical Analysis*, vol. 16, no. 6, pp. 964–979, 1979.

[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] ——, "Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems," *IEEE Transactions on Image Processing*, vol. 18, no. 11, pp. 2419–2434, 2009.

[16] 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.

[17] A. Chambolle and C. Dossal, "On the convergence of the iterates of the fast iterative shrinkage/thresholding algorithm," *Journal of Optimization theory and Applications*, vol. 166, no. 3, pp. 968–982, 2015.

[18] J.-C. Pesquet and A. Repetti, "A class of randomized primal-dual algorithms for distributed optimization," *arXiv preprint arXiv:1406.6404*, 2014.

[19] J. Douglas and H. H. Rachford, "On the numerical solution of heat conduction problems in two and three space variables," *Transactions of the American mathematical Society*, vol. 82, no. 2, pp. 421–439, 1956.

[20] R. I. Boţ, E. R. Csetnek, and C. Hendrich, "Inertial douglas–rachford splitting for monotone inclusion problems," *Applied Mathematics and Computation*, vol. 256, pp. 472–487, 2015.

[21] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein *et al.*, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine learning*, vol. 3, no. 1, pp. 1–122, 2011.

[22] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT'2010*. Springer, 2010, pp. 177–186.

[23] S. Shalev-Shwartz, Y. Singer, N. Srebro, and A. Cotter, "Pegasos: Primal estimated sub-gradient solver for svm," *Mathematical programming*, vol. 127, no. 1, pp. 3–30, 2011.

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

[25] A. Chambolle, M. J. Ehrhardt, P. Richtarik, and C.-B. Schonlieb, "Stochastic primal-dual hybrid gradient algorithm with arbitrary sampling and imaging applications," *SIAM Journal on Optimization*, vol. 28, no. 4, pp. 2783–2808, 2018.

[26] E. Chouzenoux and J.-C. Pesquet, "A stochastic majorize-minimize subspace algorithm for online penalized least squares estimation," *IEEE Transactions on Signal Processing*, vol. 65, no. 18, pp. 4770–4783, 2017.

[27] M. J. Ehrhardt, P. Markiewicz, A. Chambolle, P. Richtárik, J. Schott, and C.-B. Schönlieb, "Faster pet reconstruction with a stochastic primal-dual hybrid gradient method," in *Wavelets and Sparsity XVII*, vol. 10394. International Society for Optics and Photonics, 2017, p. 103941O.

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

[29] J. Tang, K. Egiazarian, M. Golbabaee, and M. Davies, "The practicality of stochastic optimization in imaging inverse problems," *arXiv preprint arXiv:1910.10100*, 2019.

[30] L. Rosasco, S. Villa, and B. C. Vũ, "Convergence of stochastic proximal gradient algorithm," *arXiv preprint arXiv:1403.5074*, 2014.

[31] E. Ryu, J. Liu, S. Wang, X. Chen, Z. Wang, and W. Yin, "Plug-and-play methods provably converge with properly trained denoisers," in *International Conference on Machine Learning*, 2019, pp. 5546–5557.

[32] S. Oymak, B. Recht, and M. Soltanolkotabi, "Sharp time–data tradeoffs for linear inverse problems," *IEEE Transactions on Information Theory*, vol. 64, no. 6, pp. 4129–4158, 2017.

[33] J. Bolte, A. Daniilidis, and A. Lewis, "The łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems," *SIAM Journal on Optimization*, vol. 17, no. 4, pp. 1205–1223, 2007.

[34] J. Liang, J. Fadili, and G. Peyré, "Local convergence properties of douglas–rachford and alternating direction method of multipliers," *Journal of Optimization Theory and Applications*, vol. 172, no. 3, pp. 874–913, 2017.

[35] J. Tang, M. Golbabaee, and M. E. Davies, "Gradient projection iterative sketch for large-scale constrained least-squares," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 70. PMLR, 2017, pp. 3377–3386.

[36] J. Tang, M. Golbabaee, F. Bach, and M. E. davies, "Rest-katyusha: Exploiting the solution's structure via scheduled restart schemes," in *Advances in Neural Information Processing Systems 31*. Curran Associates, Inc., 2018, pp. 427–438.

[37] Y. Nesterov, *Introductory lectures on convex optimization: A basic course*. Springer Science & Business Media, 2013, vol. 87.

[38] D. Needell, R. Ward, and N. Srebro, "Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm," in *Advances in neural information processing systems*, 2014, pp. 1017–1025.

[39] J. Wang, T. Li, H. Lu, and Z. Liang, "Penalized weighted least-squares approach to sinogram noise reduction and image reconstruction for low-dose x-ray computed tomography," *IEEE transactions on medical imaging*, vol. 25, no. 10, pp. 1272–1283, 2006.

[40] P. C. Hansen and M. Saxild-Hansen, "AIR toolsa MATLAB package of algebraic iterative reconstruction methods," *Journal of Computational and Applied Mathematics*, vol. 236, no. 8, pp. 2167–2178, 2012.

[41] X. Xu, J. Liu, Y. Sun, B. Wohlberg, and U. S. Kamilov, "Boosting the performance of plug-and-play priors via denoiser scaling," *arXiv preprint arXiv:2002.11546*, 2020.