Online Regularization by Denoising with Applications to Phase Retrieval

Zihui Wu Yu Sun Jiaming Liu Ulugbek S. Kamilov
Washington University in St. Louis
{ray.wu, sun.yu, jiaming.liu, kamilov}@wustl.edu
https://cigroup.wustl.edu

Abstract

Regularization by denoising (RED) is a powerful framework for solving imaging inverse problems. Most RED algorithms are iterative batch procedures, which limits their applicability to very large datasets. In this paper, we address this limitation by introducing a novel online RED (On-RED) algorithm, which processes a small subset of the data at a time. We establish the theoretical convergence of On-RED in convex settings and empirically discuss its effectiveness in non-convex ones by illustrating its applicability to phase retrieval. Our results suggest that On-RED is an effective alternative to the traditional RED algorithms when dealing with large datasets.[1]

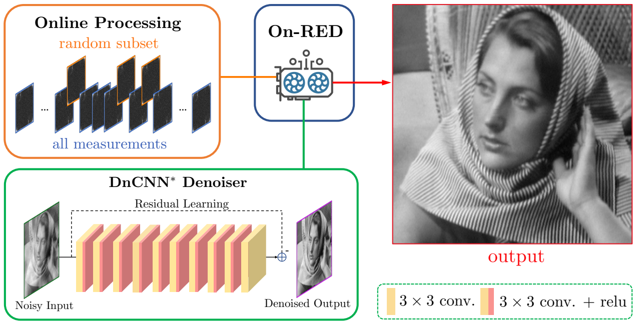

Figure 1. Conceptual illustration of *online regularization by denoising (On-RED)*. The proposed algorithm uses a *random subset of noisy measurements* at every iteration to reconstruct a high-quality image using a *convolutional neural network (CNN)* denoser.

1. Introduction

The recovery of an unknown image $\boldsymbol{x} \in \mathbb{R}^n$ from a set of noisy measurement is crucial in many applications, including computational microscopy [44], astronomical imaging [38], and phase retrieval [11]. The problem is usually formulated as a regularized optimization

$$\widehat{\boldsymbol{x}} = \mathop{\mathsf{arg\,min}}_{\boldsymbol{x} \in \mathbb{R}^N} \{f(\boldsymbol{x})\} \quad \text{with} \quad f(\boldsymbol{x}) = g(\boldsymbol{x}) + h(\boldsymbol{x}), \quad (1)$$

where g is the data-fidelity term that ensures the consistency with the measurements, and h is the regularizer that imposes the prior knowledge on the unknown image. Popular methods for solving such optimization problems include the family of proximal methods, such as proximal gradient method (PGM) [3, 4, 14, 19] and alternating direction method of multipliers (ADMM) [1, 7, 16, 30], due to their compatibility with non-differentiable regularizers [17, 18, 35].

Recent work has demonstrated the benefit of using denoisers as priors for solving imaging inverse problems [8, 12, 23, 26, 27, 37, 40, 41, 43, 49]. One popular framework, known as *plug-and-play priors (PnP)* [46], extends traditional proximal methods by replacing the proximal operator with a general denoising function. This grants PnP a remarkable flexibility in choosing image priors, but also complicates its analysis due to the lack of an explicit objective function.

An alternative strategy for leveraging denoisers is the *regularization by denoising (RED)* framework [34], which formulates an explicit regularizer h for certain classes of denoisers [33, 34]. Recent work has shown the effectiveness of RED under sophisticated denoisers for many different image reconstruction tasks [27, 33, 34, 39]. For example, Metzler *et al.* [27] demonstrated the state-of-the-art performance of RED for phase retrieval by using the DnCNN denoiser [48].

Typical PnP and RED algorithms are iterative *batch* procedures, which means that they processes the entire set of measurements at every iteration. This type of batch processing of data is known to be inefficient when dealing with large datasets [6, 24]. Recently, an online variant of PnP [40] has been proposed to address this problem, yet such an algorithm is still missing for the RED framework.

In order to address this gap, we propose an *online* extension of RED, called *online regularization by denoising (On-RED)*. Unlike its batch counterparts, On-RED adopts online processing of data by using only a random subset of measurements at a time (see Figure 1 for a conceptual

[1] This work was supported by National Science Foundation under Grant CCF-1813910.

illustration). This empowers the proposed method to effectively scale to datasets that are too large for batch processing. Moreover, On-RED can fully leverage the flexibility offered by deep learning by using *convolutional neural network (CNN)* denoisers.

The key contributions of this paper are as follows:

- We propose a novel On-RED algorithm for online processing of measurements. We provide the theoretical convergence analysis of the algorithm under several transparent assumptions. In particular, given a convex g and nonexpansive denoiser, which does not necessarily correspond to any explicit h, our analysis shows that On-RED converges to a fixed point at the worst-case rate of $O(1/\sqrt{t})$.
- We validate the effectiveness of On-RED for phase retrieval from *Coded Diffraction Patterns* (CDP) [11] under a CNN denoiser. Numerical results demonstrate the empirical fixed-point convergence of On-RED in this non-convex setting and show its potential for processing large datasets under nonconvex g.

2. Background

In this section, we first review the problem of regularized image reconstruction and then introduce some related work.

2.1. Inverse Problems in Imaging

Consider the inverse problem of recovering $\boldsymbol{x} \in \mathbb{R}^n$ from measurements $\boldsymbol{y} \in \mathbb{R}^m$ specified by the linear system

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{e}, \tag{2}$$

where the measurement matrix $\boldsymbol{H} \in \mathbb{R}^{m\times n}$ characterizes the response of the system, and $\boldsymbol{e}$ is usually assumed to be additive white Gaussian noise (AWGN). When the inverse problem is nonlinear, the measurement operator can be generalized to a mapping $\boldsymbol{H} : \mathbb{R}^n \rightarrow \mathbb{R}^m$. A common example is the problem of *phase retrieval (PR)*, which corresponds the following nonlinear system

$$\boldsymbol{y} = \boldsymbol{H}(\boldsymbol{x}) + \boldsymbol{e}, \quad \text{with} \quad \boldsymbol{H}(\boldsymbol{x}) = |\boldsymbol{A}\boldsymbol{x}| \tag{3}$$

where $|\cdot|$ denotes an element-wise absolute value, and $\boldsymbol{A} \in \mathbb{C}^{m\times n}$ is the measurement matrix.

Due to the ill-posedness, inverse problems are often formulated as (1). A widely-used data-fidelity term is the least-square loss

$$g(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{H}(\boldsymbol{x})\|_2^2, \tag{4}$$

which penalizes the mismatch to the measurements in terms of ℓ_2-norm. In particular, for the PR problem, the data-fidelity becomes $\frac{1}{2}\|\boldsymbol{y} - |\boldsymbol{A}\boldsymbol{x}|\|_2^2$, which is known to be *nonconvex*. Two common choices for the regularizer include the sparsity-enhancing ℓ_1 penalty $h(\boldsymbol{x}) = \tau\|\boldsymbol{x}\|_1$ and the total variation (TV) penalty $h(\boldsymbol{x}) = \tau\|\boldsymbol{D}\boldsymbol{x}\|_1$, where $\tau > 0$ controls the strength of regularization and $\boldsymbol{D}$ denotes the discrete gradient operator [10, 15, 22, 35, 45].

Two popular methods for solving (1) are PGM and ADMM. They circumvent the differentiation of non-smooth regularizers by using a mathematical concept called *proximal map* [29]

$$\mathsf{prox}_{\tau h}(\boldsymbol{z}) := \underset{\boldsymbol{x}\in\mathbb{R}^n}{\mathsf{arg\,min}} \left\{\frac{1}{2}\|\boldsymbol{x} - \boldsymbol{z}\|_2^2 + \tau h(\boldsymbol{x})\right\}. \tag{5}$$

A close inspection of (5) reveals that the proximal map actually corresponds to an image denoiser based on regularized optimization. This mathematical equivalence led to the development of PnP and RED.

2.2. Plug-and-play algorithms

Consider the ADMM iteration

$$\begin{aligned} \boldsymbol{z}^k &\leftarrow \mathsf{prox}_{\tau g}(\boldsymbol{x}^{k-1} - \boldsymbol{s}^{k-1}) \\ \boldsymbol{x}^k &\leftarrow \mathsf{prox}_{\tau h}(\boldsymbol{z}^k + \boldsymbol{s}^{k-1}) \\ \boldsymbol{s}^k &\leftarrow \boldsymbol{s}^{k-1} + (\boldsymbol{z}^k - \boldsymbol{x}^k), \end{aligned} \tag{6}$$

where $k \geq 1$ denotes the iteration number. In (6), the regularization is imposed by $\mathsf{prox}_{\tau h} : \mathbb{R}^n \rightarrow \mathbb{R}^n$, which denotes the proximal map of h.

Inspired by the equivalence that the proximal map is a denoiser, Venkatakrishnan *et al.* [46] introduced the PnP framework based on ADMM by replacing $\mathsf{prox}_{\tau h}$ in (6) with a general denoising function $\mathsf{D}_\sigma : \mathbb{R}^n \rightarrow \mathbb{R}^n$

$$\boldsymbol{x}^k \leftarrow \mathsf{D}_\sigma(\boldsymbol{z}^k + \boldsymbol{s}^{k-1})$$

where $\sigma > 0$ controls the strength of denoising. This simple replacement enables PnP to regularize the problem by using advanced denoisers, such as BM3D [13] and DnCNN. Numerical experiments show that PnP achieves the state-of-the-art performance in many applications. Similar PnP algorithms have been developed using PGM [23], primal-dual splitting [31], and approximate message passing (AMP) [20, 28].

Considerable effort has been made to understand the theoretical convergence of the PnP algorithms [9, 12, 26, 36, 37, 40, 42]. Recently, Sun *et al.* [40] proposed an online PnP algorithm based on PGM, named PnP-SPGM, and analyzed its fixed-point convergence using the monotone operator theory [2]. This paper extends their results to the RED framework by introducing a new algorithm and analyzing its theoretical convergence.

2.3. Regularization by Denoising

The RED framework, proposed by Romano *et al.* [34], is an alternative way to leverage image denoisers. RED has

Algorithm 1 GM-RED

```
input: $\boldsymbol{x}^0 \in \mathbb{R}^n$, $\tau > 0$, and $\sigma > 0$
for $k = 1, 2, \dots$ do
    $\nabla g(\boldsymbol{x}^{k-1}) \leftarrow \mathsf{fullGradient}(\boldsymbol{x}^{k-1})$
    $\mathsf{G}(\boldsymbol{x}^{k-1}) \leftarrow \nabla g(\boldsymbol{x}^{k-1}) + \tau(\boldsymbol{x}^{k-1} - \mathsf{D}_\sigma(\boldsymbol{x}^{k-1}))$
    $\boldsymbol{x}^k \leftarrow \boldsymbol{x}^{k-1} - \gamma \mathsf{G}(\boldsymbol{x}^{k-1})$
end for
```

Algorithm 2 On-RED

```
input: $\boldsymbol{x}^0 \in \mathbb{R}^n$, $\tau > 0$, $\sigma > 0$, and $B \geq 1$
for $k = 1, 2, \dots$ do
    $\widehat{\nabla} g(\boldsymbol{x}^{k-1}) \leftarrow \mathsf{minibatchGradient}(\boldsymbol{x}^{k-1}, B)$
    $\widehat{\mathsf{G}}(\boldsymbol{x}^{k-1}) \leftarrow \widehat{\nabla} g(\boldsymbol{x}^{k-1}) + \tau(\boldsymbol{x}^{k-1} - \mathsf{D}_\sigma(\boldsymbol{x}^{k-1}))$
    $\boldsymbol{x}^k \leftarrow \boldsymbol{x}^{k-1} - \gamma \widehat{\mathsf{G}}(\boldsymbol{x}^{k-1})$
end for
```

been shown successful in many regularized reconstruction tasks, including image deblurring [34], super-resolution [25], and phase retrieval [27]. The framework aims to find a fixed point $\boldsymbol{x}^*$ that satisfies

$$\mathsf{G}(\boldsymbol{x}^*) = \nabla g(\boldsymbol{x}^*) + \tau(\boldsymbol{x}^* - \mathsf{D}_\sigma(\boldsymbol{x}^*)) = 0, \tag{7}$$

where $\tau > 0$ and ∇g denotes the gradient of g. Equivalently, $\boldsymbol{x}^*$ lies in the zero set of $\mathsf{G} : \mathbb{R}^n \rightarrow \mathbb{R}^n$

$$\boldsymbol{x}^* \in \mathsf{zer}(\mathsf{G}) \;:=\; \{\boldsymbol{x} \in \mathbb{R}^n \mid \mathsf{G}(\boldsymbol{x}) = 0\}. \tag{8}$$

Romano *et al.* discussed several RED algorithms for finding such $\boldsymbol{x}^*$. One popular algorithm is the gradient descent (summarized in Algorithm 1)

$$\begin{aligned} &\boldsymbol{x}^k \leftarrow \boldsymbol{x}^{k-1} - \gamma(\nabla g(\boldsymbol{x}^{k-1}) + \mathsf{H}(\boldsymbol{x}^{k-1})) \\ &\quad \text{with} \quad \mathsf{H}(\boldsymbol{x}) \;:=\; \tau(\boldsymbol{x} - \mathsf{D}_\sigma(\boldsymbol{x})), \end{aligned} \tag{9}$$

where $\gamma > 0$ is the step-size. They have justified $\mathsf{H}(\cdot)$ as a gradient of some explicit function under some conditions. In particular, when denoiser D_σ is locally homogeneous and has a symmetric Jacobian [33, 34], H corresponds to the gradient of the following regularizer

$$h(\boldsymbol{x}) = \frac{\tau}{2}\boldsymbol{x}^{\mathsf{T}}(\boldsymbol{x} - \mathsf{D}_\sigma(\boldsymbol{x})). \tag{10}$$

By having a closed-form objective function, one can use the classical optimization theory to analyze the convergence of RED algorithms [34]. On the other hand, fixed-point convergence has also been established without having an explicit objective function [33,39]. Reehorst *et al.* [33] have shown that RED proximal gradient methods (RED-PG) converge to a fixed point by utilizing the montone operator theory. Sun *et al.* [39] have established the worst-case convergence for the block coordinate variant of RED algorithm (BC-RED) under a nonexpansive D_σ. In this paper, we extend the analysis of BC-RED in [39] to the randomized processing of measurements instead of image blocks, which opens up applications requiring the processing of a large number of measurements.

3. Online Regularization by Denoising

We now introduce the proposed online RED (On-RED), which processes the measurements in an online fashion. The online processing of measurements is especially beneficial for problems with the following data-fidelity

$$g(\boldsymbol{x}) = \mathbb{E}[g_i(\boldsymbol{x})] = \frac{1}{I}\sum_{i=1}^{I} g_i(\boldsymbol{x}), \tag{11}$$

which is composed of I component functions $g_i(\boldsymbol{x})$, each evaluated only on the subset $\boldsymbol{y}_i$ of the measurements $\boldsymbol{y}$. The computation of the gradient

$$\nabla g(\boldsymbol{x}) = \mathbb{E}[\nabla g_i(\boldsymbol{x})] = \frac{1}{I}\sum_{i=1}^{I} \nabla g_i(\boldsymbol{x}), \tag{12}$$

is proportional to the total number I. Note that the expectation in (11) and (12) is taken over a uniformly distributed random variable $i \in \{1, \dots, I\}$. Large I effectively precludes the usage of batch GM-RED algorithms because of large memory requirements or impractical computation times. The key idea of On-RED is to approximate the gradient at every iteration by averaging $B \ll I$ component gradients

$$\widehat{\nabla} g(\boldsymbol{x}) = \frac{1}{B}\sum_{b=1}^{B} \nabla g_{i_b}(\boldsymbol{x}), \tag{13}$$

where $i_1, \dots, i_B$ are independent random indices that are distributed uniformly over $\{1, \dots, I\}$. The *minibatch* size parameter $B \geq 1$ controls the number of gradient components used at every iteration.

Algorithm 2 summarizes the algorithmic details of On-RED, where the operation $\mathsf{minibatchGradient}$ computes the averaged gradients with respect to the selected minibatch components. Note that at each iteration, the minibatch is randomly sampled from the entire set of measurements. In the next section, we will present the theoretical convergence analysis of On-RED.

4. Convergence Analysis under Convexity

A fixed-point convergence of averaged operators is well known as Krasnosel'skii-Mann theorem [2], which was applied to the aforementioned analysis of PnP [40] and RED algorithms [33, 39]. Here, our analysis extends these results to the online processing of measurements and provides explicit worst-case convergence rates for On-RED. Note that our analysis does not assume that H corresponds to any explicit regularizer h. We first introduce the assumptions necessary for our analysis and then present the main results.

Assumption 1. *We make the following assumptions on the data-fidelity term g:*

(a) The component functions g_i are all convex and differentiable with the same Lipschitz constant $L > 0$.

(b) At every iteration, the gradient estimate is unbiased and has a bounded variance:

$$\mathbb{E}[\widehat{\nabla} g(\boldsymbol{x})] = \nabla g(\boldsymbol{x}), \;\; \mathbb{E}[\|\nabla g(\boldsymbol{x}) - \widehat{\nabla} g(\boldsymbol{x})\|_2^2] \leq \frac{\nu^2}{B},$$

for some constant $\nu > 0$.

Assumption 1(a) implies that the overall data-fidelity g is also convex and has Lipschitz continuous gradient with constant L. Assumption 1(b) assumes that the minibatch gradient is an unbiased estimate of the full gradient. The bounded variance assumption is a standard assumption used in the analysis of online and stochastic algorithms [5, 21, 40, 47]

Assumption 2. *Let operator G have a nonempty zero set $\mathsf{zer}(\mathsf{G}) \neq \varnothing$. The distance between the the farthest point in $\mathsf{zer}(\mathsf{G})$ and the sequence $\{\boldsymbol{x}^k\}_{k=0,1,\cdots}$ generated by On-RED is bounded by constant R_0*

$$\max_{\boldsymbol{x}^* \in \mathsf{zer}(\mathsf{G})} \|\boldsymbol{x}^k - \boldsymbol{x}^*\|_2 \leq R_0, \quad k \geq 0$$

This assumption indicates that the iterates of On-RED lie within a Euclidean ball of a bounded radius from $\mathsf{zer}(\mathsf{G})$.

Assumption 3. *Given $\sigma > 0$, the denoiser D_σ is a nonexpansive operator such that*

$$\|\mathsf{D}_\sigma(\boldsymbol{x}) - \mathsf{D}_\sigma(\boldsymbol{y})\|_2 \leq \|\boldsymbol{x} - \boldsymbol{y}\|_2 \quad \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n,$$

Since the proximal operator is nonexpansive [32], it automatically satisfies this assumption. Nonexpansive CNN denoisers can also be trained by using spectral normalization techniques [39]. Under the above assumptions, we now establish the convergence theorem for On-RED.

Theorem 1. *Run On-RED for $t \geq 1$ iterations under Assumptions 1-3 using a fixed step-size $\gamma \in (0, 1/(L + 2\tau)]$ and a fixed minibatch size $B \geq 1$. Then, we have*

$$\begin{aligned} &\mathbb{E}\left[\min_{k \in \{1,\ldots,t\}} \|\mathsf{G}(\boldsymbol{x}^{k-1})\|_2^2\right] \\ &\leq \mathbb{E}\left[\frac{1}{t}\sum_{k=1}^{t} \|\mathsf{G}(\boldsymbol{x}^{k-1})\|_2^2\right] \\ &\leq \frac{(L+2\tau)}{\gamma}\left[\frac{\nu^2\gamma^2}{B} + \frac{2\gamma\nu}{\sqrt{B}}R_0 + \frac{R_0^2}{t}\right]. \end{aligned}$$

Proof. See Section 7. □

When t goes to infinity, this theorem shows that the accuracy of the expected convergence of On-RED to an element of $\mathsf{zer}(\mathsf{G})$ improves with smaller γ and larger B. For example, we can have the convergence rate of $O(1/\sqrt{t})$ by setting $\gamma = 1/(L + 2\tau)$ and $B = t$

$$\mathbb{E}\left[\frac{1}{t}\sum_{k=1}^{t} \|\mathsf{G}(\boldsymbol{x}^{k-1})\|_2^2\right] \leq \frac{C}{\sqrt{t}},$$

where $C > 0$ is a constant and we use the bound $\frac{1}{t} \leq \frac{1}{\sqrt{t}}$ that is valid for $t \geq 1$.

5. Numerical Simulation for Phase Retrieval

In this section, we test the performance of On-RED on a nonconvex phase retrieval problem from *coded diffraction patterns (CDP)*. The state-of-the-art performance of RED for this problem was shown by Metzler *et al.* [27]. Here, we investigate the convergence of On-RED and show its effectiveness for reducing the per-iteration complexity of the traditional batch GM-RED. Our results show the potential of On-RED to scale to a large member of measurements under powerful denoisers that do not correspond to explicit regularizers.

5.1. Experiment Setup

In CDP, the object $\boldsymbol{x} \in \mathbb{R}^n$ is illuminated by a coherent light source. A random known phase mask modulates the light and the modulation code is denoted as $\boldsymbol{M}_i$ for the ith measurement. In this work, each entry of $\boldsymbol{M}_i$ is drawn uniformly from the unit circle in the complex plane. The light goes through the far-field Fraunhofer diffraction and a camera measures its intensity $\boldsymbol{y}_i \in \mathbb{R}_+$. Since Fraunhofer diffraction can be modeled by 2D Fourier Transform, the ith data-fidelity term of this phase reconstruction problem can be formulated as follows:

$$g_i(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{y}_i - |\boldsymbol{F}\boldsymbol{M}_i\boldsymbol{x}|\|_2^2$$

where $\boldsymbol{F}$ denotes 2D discrete Fast Fourier Transform (FFT). The total data-fidelity term for all the measurements then becomes

$$g(\boldsymbol{x}) = \mathbb{E}[g_i(\boldsymbol{x})] = \frac{1}{I}\sum_{i=1}^{I} g_i(\boldsymbol{x}).$$

Noticeably, this problem is well suited for On-RED because it has the same formulation as (11).

In the experiments, we reconstruct six 256×256 standard grayscale natural images, displayed in Figure 2. The simulated measurements are corrupted by AWGN corresponding

Figure 2. Test images used in the experiments. From left to right: *Barbara*, *Boat*, *Lenna*, *Monarch*, *Parrot*, *Pepper*.

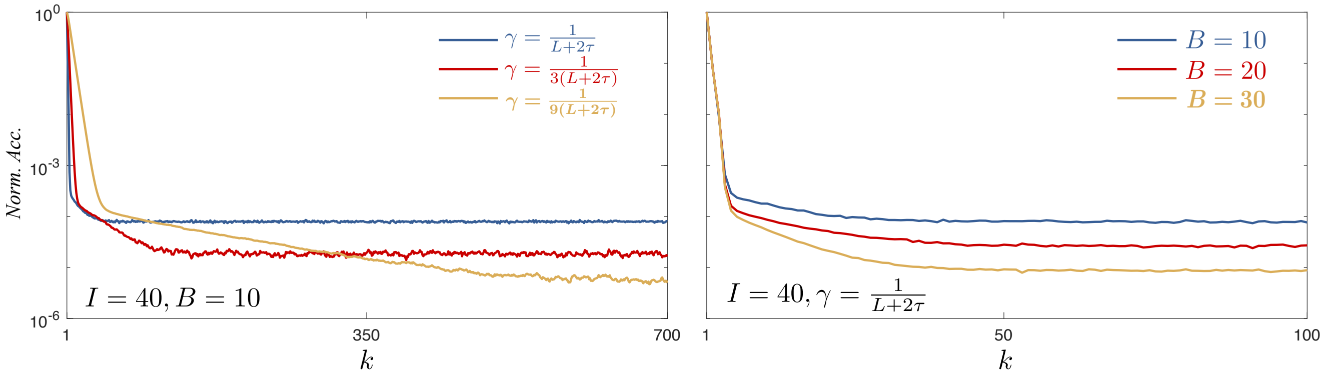

Figure 3. Illustration of the influence of γ and B on the convergence of On-RED for phase retrieval under DnCNN*. The left plot shows the convergence results of On-RED for three different step sizes with a fixed minibatch size $B = 10$ and the right plot shows the results of On-RED for three different minibatch sizes with a fixed step size $\gamma = \frac{1}{L+2\tau}$. Both experiments draw random samples from a total of $I = 40$ measurements. The plots validate that smaller γ and larger B improve the convergence accuracy in this nonconvex setting.

to 25 dB of input signal-to-noise ratio (SNR), defined as follows

$$\mathrm{SNR}(\hat{\boldsymbol{y}}, \boldsymbol{y}) = 20\log_{10} \frac{\|\boldsymbol{y}\|}{\|\boldsymbol{y} - \hat{\boldsymbol{y}}\|}$$

where $\hat{\boldsymbol{y}}$ represents the noisy vector and $\boldsymbol{y}$ denotes the ground truth. We also use SNR as a quantitative measure for the quality of reconstructions.

We used DnCNN* as our CNN denoiser for the experiments. The architecture of DnCNN* is illustrated in Figure 1 and was adopted from the popular DnCNN. We generated training examples by adding AWGN to images from BSD400 and applying standard data augmentation strategy including flipping, rotating, and rescaling. We used the residual learning technique where DnCNN* predicts the noise image from the input. The network was trained to minimize the following loss

$$\mathcal{L}_\theta = \frac{1}{n}\sum_{i=1}^{n}\left\{\|f_\theta(\boldsymbol{x}_i) - \boldsymbol{y}_i\|_2^2 + \|f_\theta(\boldsymbol{x}_i) - \boldsymbol{y}_i\|_1\right\}, \quad (14)$$

where $\boldsymbol{x}_i$ is the noisy input, $\boldsymbol{y}_i$ is the noise, and f_θ represents DnCNN*.

The hyperparameters for experiments in 5.2 and 5.3 are listed in Table 1. All algorithms start from $\boldsymbol{x}^0 = \mathbf{0}$, where $\mathbf{0} \in \mathbb{R}^n$ is all zeros. The value of τ for each image was optimized for the best SNR performance with respect to ground truth test images. In this paper, the values of B and I are set only to show the potential of On-RED dealing with large datasets.

Table 1. List of algorithmic hyperparameters

Hyperparameters		5.2	5.3
$\boldsymbol{x}^0$	initial point of reconstructions	$\mathbf{0}$	$\mathbf{0}$
σ	input noise level for denoisers	5	5
τ	level of regularization in RED	0.2	optimized
γ	step size	$\frac{1}{L+2\tau} \cdot \{1, \frac{1}{3}, \frac{1}{9}\}$	$\frac{1}{L+2\tau}$
B	minibatch size at every iteration	$\{10, 20, 30\}$	1
I	batch size	40	6

Table 2. Convergence accuracy averaged over the test images

Denoiser	Step-size (γ)			Mini-batch size (B)		
	$\frac{1}{L+2\tau}$	$\frac{1}{3(L+2\tau)}$	$\frac{1}{9(L+2\tau)}$	10	20	30
TV	8.65e-5	2.36e-5	9.43e-6	8.65e-5	2.81e-5	9.81e-6
BM3D	8.01e-5	1.59e-5	9.10e-6	8.01e-6	2.72e-5	8.93e-6
DnCNN*	7.63e-5	1.94e-6	5.03e-6	7.63e-5	2.72e-5	8.88e-6

5.2. Convergence of On-RED

Theorem 1 implies that the expected accuracy improves for a smaller step size γ and larger minibatch size B. In order to numerically evaluate the convergence, we define and consider the following normalized accuracy

$$\textit{Norm. Acc.} := \|\mathsf{G}(\boldsymbol{x}^k)\|_2^2 / \|\mathsf{G}(\boldsymbol{x}^0)\|_2^2$$

where G is defined in (7). As the sequence $\{\boldsymbol{x}^k\}_{k=0,1,...}$ converges to a fixed point in $\mathrm{zer}(\mathsf{G})$, the normalized accuracy

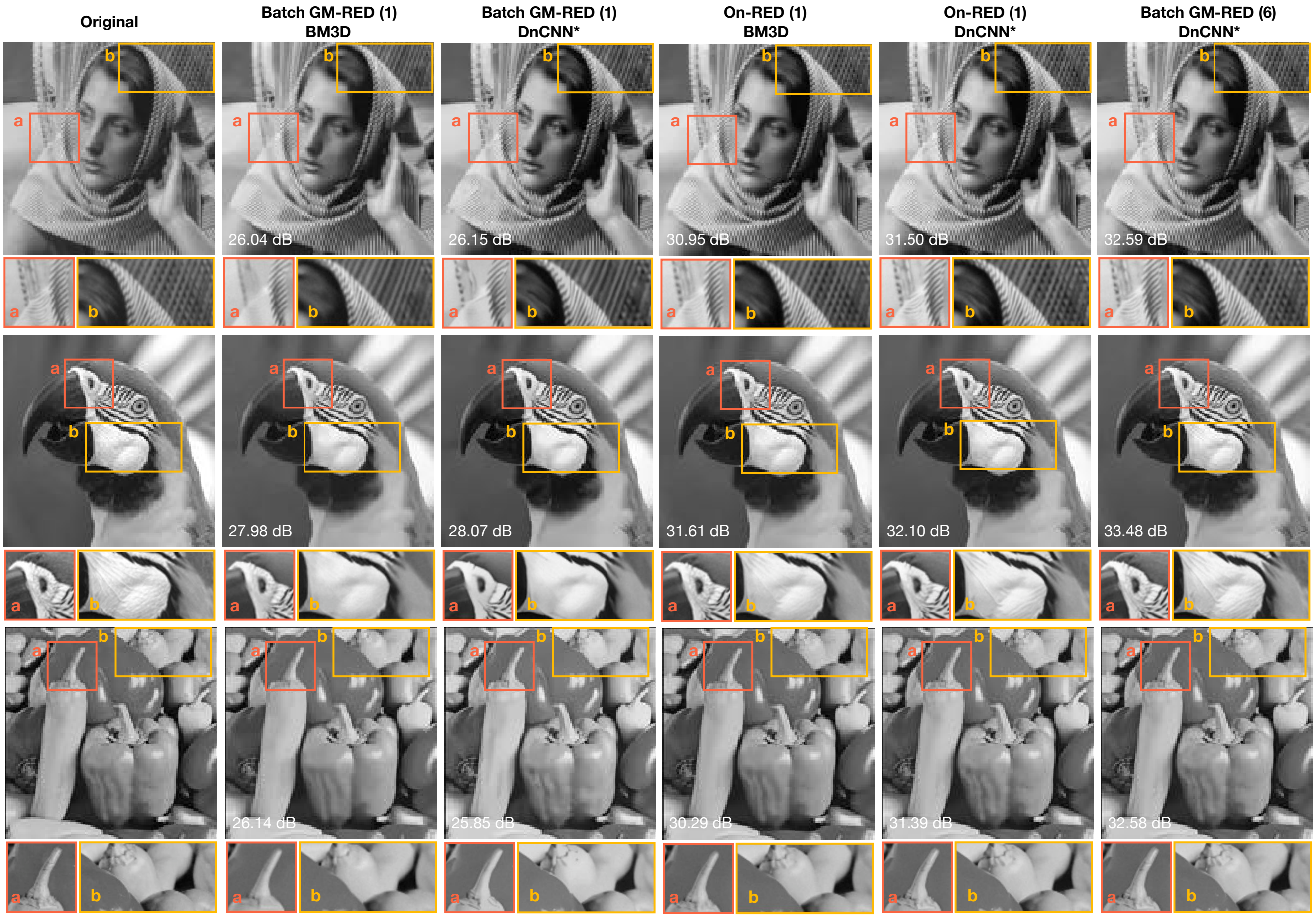

Figure 4. Visual examples of recontructed *Barbara*, *Parrot*, and *Pepper* images by GM-RED (1), On-RED (1), and GM-RED (6) with BM3D and DnCNN* denoisers. The original images are displayed in the first column. The second and the third columns show the results of batch GM-RED using 1 fixed measurement. The fourth and the fifth columns present the results of On-RED using a single randomly selected measurement per-iteration out of 6 total measurements. The results of the batch algorithm using all 6 measurements are given in the last column. Differences are zoomed in using boxes inside the images. Each reconstruction is labeled by its SNR (dB) with respect to the original image. Note that On-RED (1) recovers the details lost by GM-RED (1) by approaching the performance of GM-RED (6)

decreases to zero.

Figure 3 (left) shows the evolution of the convergence accuracy for $\gamma \in \{\frac{1}{L+2\tau}, \frac{1}{3(L+2\tau)}, \frac{1}{9(L+2\tau)}\}$ with DnCNN*. Here, L denotes the Lipschitz constant defined in Assumption 1 and τ represents the parameter of RED. We observe that the empirical performance of On-RED using DnCNN* is consistent with Theorem 1, as the accuracy improves with smaller step size. Moreover, Figure 3 (right) numerically evaluates the convergence accuracy of On-RED for minibatch size $B \in \{10, 20, 30\}$. This plot shows that the convergence accuracy improves when minibatch size B becomes larger. Therefore, the change of convergence accuracy with both step size γ and minibatch size B follows the same trend in Theorem 1 for this nonconvex problem.

We note that the similar trend generalizes to BM3D and TV denoisers as well. The summary in Table 2 gives the convergence results of all three denoisers.

5.3. Benefits of On-RED with a CNN Denoiser

In this subsection, we show the performance and efficiency of On-RED in solving CDP. To understand the potential of On-RED to scale to large datasets, we consider the scenario where the number of illuminations processed at every iteration is fixed to one.

Table 3 provides the SNR performance of different algorithms. GM-RED (fixed 1) uses 1 fixed measurement and On-RED $(B = 1)$ uses 1 random measurement out of 6 total measurements at every iteration, so they have the same per-iteration computation cost. On-RED outperforms GM-RED by 4.54 dB and 4.99 dB under BM3D and DnCNN*, respectively, by actually using all measurements. We also note that the average SNR of *stochastic gradient method (SGM)* $(B = 1)$ is higher than that of GM-RED (fixed 1) for both denoisers. This implies that the online processing in SGM

Table 3. Optimized SNR for each test image in dB

Algorithms $(I = 6)$	SGM $(B = 1)$	GM-RED (fixed 1)		On-RED $(B = 1)$		GM-RED (fixed 6)
Denoisers	—	BM3D	DnCNN*	BM3D	DnCNN*	DnCNN*
Barbara	27.37	26.04	26.15	30.95	31.50	32.59
Boat	27.68	26.90	27.53	31.65	32.61	33.17
Lenna	27.65	26.55	27.58	31.47	32.54	33.20
Monarch	27.51	24.76	26.34	29.66	31.31	32.63
Parrot	27.20	27.98	28.07	31.61	32.10	33.48
Pepper	27.08	26.14	25.85	30.29	31.39	32.58
Average	27.42	26.40	26.92	30.94	31.91	32.94

boosts the SNR more than the regularization of GM-RED. By combining online processing and advanced denoisers, On-RED largely improves the reconstruction performance, which is close to that of the batch algorithm GM-RED (6) using all 6 measurements.

Visual illustrations of *Barbara*, *Parrot*, and *Pepper* are given in Figure 4. It is clear that the images reconstructed by On-RED (1) preserve the features lost by GM-RED (1), such as the stripes in *Barbara*, the white feather in *Parrot*, and the stems in *Pepper*. Moreover, these features in the reconstructed images of On-RED (1) have no visual difference from the results of GM-RED (6), as illustrated by column 4, 5, and 6. This indicates that the online algorithm approaches the image quality of the batch algorithm with a lower per-iteration complexity.

6. Conclusion

In this paper, we proposed an online algorithm for the Regularization by Denoising framework. We provided the theoretical convergence proof under a few transparent assumptions and a detailed analysis in a convex problem setting. We then applied On-RED to a nonconvex phase retrieval problem from coded diffraction patterns to show its convergence. The performance of On-RED with our learning denoiser DnCNN* demonstrated that On-RED is well compatible with powerful denoisers that do not correspond to explicit regularizers. Our results showed that On-RED has the potential to solve data-intensive problems involving a large number of measurements by reducing per-iteration computation cost.

7. Proof of Theorem 1

We consider the following two operators

$$\mathsf{P} := \mathsf{I} - \gamma\mathsf{G} \quad \text{and} \quad \widehat{\mathsf{P}} := \mathsf{I} - \gamma\widehat{\mathsf{G}}$$

where $\widehat{\mathsf{P}}$ is the online variant of P. The iterates of On-RED can be expressed as

$$\boldsymbol{x}^k = \widehat{\mathsf{P}}(\boldsymbol{x}^{k-1}) = \boldsymbol{x}^{k-1} - \gamma\widehat{\mathsf{G}}(\boldsymbol{x}^{k-1}), \quad \text{with} \quad \widehat{\mathsf{G}} = \widehat{\nabla} g + \mathsf{H}.$$

Note also the following equivalence

$$\boldsymbol{x}^* \in \mathsf{zer}(\mathsf{G}) \quad \Leftrightarrow \quad \boldsymbol{x}^* \in \mathsf{fix}(\mathsf{P})$$

Proposition 1. *Consider an operater* P *and its online variant* $\widehat{\mathsf{P}}$. *If the data-fidelity* $g(\cdot)$ *satisfies Assumption 1, then we have*

$$\mathbb{E}[\widehat{\mathsf{P}}(\boldsymbol{x})] = \mathsf{P}(\boldsymbol{x}), \quad \mathbb{E}[\|\mathsf{P}(\boldsymbol{x}) - \widehat{\mathsf{P}}(\boldsymbol{x})\|_2^2] \leq \frac{\gamma^2\nu^2}{B}.$$

Proof. First, we can show

$$\mathbb{E}[\widehat{\mathsf{G}}(\boldsymbol{x})] = \mathbb{E}[\widehat{\nabla} g(\boldsymbol{x})] + \mathsf{H}(\boldsymbol{x}) = \mathsf{G}(\boldsymbol{x})$$

and

$$\mathbb{E}[\|\mathsf{G}(\boldsymbol{x}) - \widehat{\mathsf{G}}(\boldsymbol{x})\|_2^2] = \mathbb{E}[\|\nabla g(\boldsymbol{x}) - \widehat{\nabla} g(\boldsymbol{x})\|_2^2] \leq \frac{\nu^2}{B}$$

Then, we can prove the desired result

$$\mathbb{E}[\widehat{\mathsf{P}}(\boldsymbol{x})] = \mathsf{I} - \gamma\mathbb{E}[\widehat{\mathsf{G}}(\boldsymbol{x})] = \mathsf{P}(\boldsymbol{x})$$

and

$$\mathbb{E}[\|\mathsf{P}(\boldsymbol{x}) - \widehat{\mathsf{P}}(\boldsymbol{x})\|_2^2] = \gamma^2\,\mathbb{E}[\|\mathsf{G}(\boldsymbol{x}) - \widehat{\mathsf{G}}(\boldsymbol{x})\|_2^2] \leq \frac{\gamma^2\nu^2}{B}$$

□

Proposition 2. *Let the denoiser* D_σ *be such that it satisfies Assumption 3 and* ∇g *is L-Lipschitz continuous. For any* $\gamma \in (0, 1/(L + 2\tau)]$, *the operator* P *is nonexpansive*

$$\|\mathsf{P}(\boldsymbol{x}) - \mathsf{P}(\boldsymbol{y})\|_2 \leq \|\boldsymbol{x} - \boldsymbol{y}\|_2 \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$$

Proof. The proposition is a direct result of the part (c) of the proof of Theorem 1 (Section A) in the Supplementary Material of [39] by setting $\mathsf{U} = \mathsf{U}^\mathsf{T} = \mathsf{I}$ and $\mathsf{G}_i = \mathsf{G}$, which corresponds to the full-gradient RED algorithm of (9). □

Now we prove Theorem 1 in the paper. Consider a single iteration $\boldsymbol{x}^k = \widehat{\mathsf{P}}(\boldsymbol{x}^{k-1})$, then we can write for any $\boldsymbol{x}^* \in \mathsf{zer}(\mathsf{G})$ that

$$\begin{aligned}
&\|\boldsymbol{x}^k - \boldsymbol{x}^*\|_2^2 = \|\widehat{\mathsf{P}}(\boldsymbol{x}^{k-1}) - \mathsf{P}(\boldsymbol{x}^*)\|_2^2 \\
&= \|\widehat{\mathsf{P}}(\boldsymbol{x}^{k-1}) - \mathsf{P}(\boldsymbol{x}^{k-1}) + \mathsf{P}(\boldsymbol{x}^{k-1}) - \mathsf{P}(\boldsymbol{x}^*)\|_2^2 \\
&= \|\mathsf{P}(\boldsymbol{x}^{k-1}) - \mathsf{P}(\boldsymbol{x}^*)\|_2^2 + \|\widehat{\mathsf{P}}(\boldsymbol{x}^{k-1}) - \mathsf{P}(\boldsymbol{x}^{k-1})\|_2^2 \\
&\qquad + 2(\widehat{\mathsf{P}}(\boldsymbol{x}^{k-1}) - \mathsf{P}(\boldsymbol{x}^{k-1}))^\mathsf{T}(\mathsf{P}(\boldsymbol{x}^{k-1}) - \mathsf{P}(\boldsymbol{x}^*)) \\
&\leq \|\boldsymbol{x}^{k-1} - \boldsymbol{x}^*\|_2^2 - \left(\frac{\gamma}{L + 2\tau}\right)\|\mathsf{G}(\boldsymbol{x}^{k-1})\|_2^2 \qquad (15) \\
&\qquad + \|\widehat{\mathsf{P}}(\boldsymbol{x}^{k-1}) - \mathsf{P}(\boldsymbol{x}^{k-1})\|_2^2 \\
&\qquad + 2\|\widehat{\mathsf{P}}(\boldsymbol{x}^{k-1}) - \mathsf{P}(\boldsymbol{x}^{k-1})\|_2 \cdot \|\mathsf{P}(\boldsymbol{x}^{k-1}) - \mathsf{P}(\boldsymbol{x}^*)\|_2,
\end{aligned}$$

where we use the Cauchy-Schwarz inequality and adapt the bound (14) in the part (d) of the proof of Theorem 1 (Section A) in the Supplementary Material of [39] by setting

$\mathsf{U} = \mathsf{U}^\mathsf{T} = \mathsf{I}$ and $\mathsf{G}_i = \mathsf{G}$. According to Assumption 2 and Proposition 2, we have

$$\|\mathsf{P}(\boldsymbol{x}^{k-1}) - \mathsf{P}(\boldsymbol{x}^*)\|_2 \leq \|\boldsymbol{x}^{k-1} - \boldsymbol{x}^*\|_2 \leq R_0. \tag{16}$$

Additionally, by using Jensen's inequality, we can have for all $\boldsymbol{x} \in \mathbb{R}^n$ that

$$\begin{aligned} &\mathbb{E}\left[\|\mathsf{P}(\boldsymbol{x}) - \widehat{\mathsf{P}}(\boldsymbol{x})\|_2\right] = \mathbb{E}\left[\sqrt{\|\mathsf{P}(\boldsymbol{x}) - \widehat{\mathsf{P}}(\boldsymbol{x})\|_2^2}\right] \\ &\leq \sqrt{\mathbb{E}\left[\|\mathsf{P}(\boldsymbol{x}) - \widehat{\mathsf{P}}(\boldsymbol{x})\|_2^2\right]} \leq \frac{\gamma\nu}{\sqrt{B}}. \end{aligned} \tag{17}$$

By rearranging and taking a conditional expectation of (15) and using these bounds, we can obtain

$$\begin{aligned} &\mathbb{E}\left[\|\boldsymbol{x}^k - \boldsymbol{x}^*\|_2^2 - \|\boldsymbol{x}^{k-1} - \boldsymbol{x}^*\|_2^2 \mid \boldsymbol{x}^{k-1}\right] \\ &\leq \frac{2\gamma\nu}{\sqrt{B}}R_0 + \frac{\gamma^2\nu^2}{B} - \left(\frac{\gamma}{L + 2\tau}\right)\|\mathsf{G}(\boldsymbol{x}^{k-1})\|_2^2, \end{aligned}$$

which can be reorganized as

$$\begin{aligned} \|\mathsf{G}(\boldsymbol{x}^{k-1})\|_2^2 \leq \left(\frac{L + 2\tau}{\gamma}\right) \bigg[&\frac{\gamma^2\nu^2}{B} + \frac{2\gamma\nu}{\sqrt{B}}R_0 \\ &+ \mathbb{E}\left[\|\boldsymbol{x}^{k-1} - \boldsymbol{x}^*\|_2^2 - \|\boldsymbol{x}^k - \boldsymbol{x}^*\|_2^2 \mid \boldsymbol{x}^{k-1}\right]\bigg]. \end{aligned}$$

By averaging the inequality over $t \geq 1$ iterations, taking the total expectation, and dropping the last term, we obtain

$$\begin{aligned} &\mathbb{E}\left[\frac{1}{t}\sum_{k=1}^{t}\|\mathsf{G}(\boldsymbol{x}^{k-1})\|_2^2\right] \\ &\leq \frac{L + 2\tau}{\gamma}\left[\frac{\gamma^2\nu^2}{B} + \frac{2\gamma\nu}{\sqrt{B}}R_0 + \frac{R_0^2}{t}\right] \end{aligned}$$

where we apply the law of total expectation and Assumption 2. This establishes the Theorem 1.

References

[1] M. V. Afonso, J. M.Bioucas-Dias, and M. A. T. Figueiredo. Fast image recovery using variable splitting and constrained optimization. *IEEE Trans. Image Process.*, 19(9):2345–2356, September 2010. 1

[2] H. H. Bauschke and P. L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, 2 edition, 2017. 2, 3

[3] A. Beck and M. Teboulle. Fast gradient-based algorithm for constrained total variation image denoising and deblurring problems. *IEEE Trans. Image Process.*, 18(11):2419–2434, November 2009. 1

[4] J. Bect, L. Blanc-Feraud, G. Aubert, and A. Chambolle. A ℓ_1-unified variational framework for image restoration. In *Proc. Euro. Conf. Comp. Vis. (ECCV)*, volume 3024, pages 1–13, New York, 2004. 1

[5] J. Bernstein, Y.-X. Wang, K. Azizzadenesheli, and A. Anandkumar. signSGD: Compressed optimization for non-convex problems. In *Proc. 35th Int. Conf. Machine Learning (ICML)*, volume 80, pages 560–569, Stockholm, Sweden, July 2018. 4

[6] L. Bottou and O. Bousquet. The tradeoffs of large scale learning. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, pages 161–168, Vancouver, BC, Canada, December 3-6, 2007. 1

[7] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011. 1

[8] A. Brifman, Y. Romano, and M. Elad. Turning a denoiser into a super-resolver using plug and play priors. In *Proc. IEEE Int. Conf. Image Proc. (ICIP)*, pages 1404–1408, Phoenix, AZ, USA, September 25-28, 2016. 1

[9] G. T. Buzzard, S. H. Chan, S. Sreehari, and C. A. Bouman. Plug-and-play unplugged: Optimization free reconstruction using consensus equilibrium. *SIAM J. Imaging Sci.*, 11(3):2001–2020, September 2018. 2

[10] E. J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inf. Theory*, 52(2):489–509, February 2006. 2

[11] E. J. Candès, T. Strohmer, and V. Voroninski. PhaseLift: Exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics*, 66(8):1241–1274, 2013. 1, 2

[12] S. H. Chan, X. Wang, and O. A. Elgendy. Plug-and-play ADMM for image restoration: Fixed-point convergence and applications. *IEEE Trans. Comp. Imag.*, 3(1):84–98, March 2017. 1, 2

[13] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian. Image denoising by sparse 3-D transform-domain collaborative filtering. *IEEE Trans. Image Process.*, 16(16):2080–2095, August 2007. 2

[14] I. Daubechies, M. Defrise, and C. D. Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Commun. Pure Appl. Math.*, 57(11):1413–1457, November 2004. 1

[15] D. L. Donoho. Compressed sensing. *IEEE Trans. Inf. Theory*, 52(4):1289–1306, April 2006. 2

[16] J. Eckstein and D. P. Bertsekas. On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55:293–318, 1992. 1

[17] M. Elad and M. Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Trans. Image Process.*, 15(12):3736–3745, December 2006. 1

[18] M. A. T. Figueiredo and R. D. Nowak. Wavelet-based image estimation: An empirical Bayes approach using Jeffreys' noninformative prior. *IEEE Trans. Image Process.*, 10(9):1322–1331, September 2001. 1

[19] M. A. T. Figueiredo and R. D. Nowak. An EM algorithm for wavelet-based image restoration. *IEEE Trans. Image Process.*, 12(8):906–916, August 2003. 1

[20] A. K. Fletcher, P. Pandit, S. Rangan, S. Sarkar, and P. Schniter. Plug-in estimation in high-dimensional linear inverse problems: A rigorous analysis. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, pages 7451–7460, Montreal, QC, Canada, December 2-8, 2018. 2

[21] S. Ghadimi and G. Lan. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Math. Program. Ser. A*, 156(1):59–99, March 2016. 4

[22] U. S. Kamilov. A parallel proximal algorithm for anisotropic total variation minimization. *IEEE Trans. Image Process.*, 26(2):539–548, February 2017. 2

[23] U. S. Kamilov, H. Mansour, and B. Wohlberg. A plug-and-play priors approach for solving nonlinear imaging inverse problems. *IEEE Signal. Proc. Let.*, 24(12):1872–1876, December 2017. 1, 2

[24] D. Kim, D. Pal, J. Thibault, and J. A. Fessler. Accelerating ordered subsets image reconstruction for X-ray CT using spatially nonuniform optimization transfer. *IEEE Trans. Med. Imag.*, 32(11):1965–1978, Nov 2013. 1

[25] G. Mataev, M. Elad, and P. Milanfar. Deepred: Deep image prior powered by RED. 2019. arXiv:1903.10176 [cs.CV]. 3

[26] T. Meinhardt, M. Moeller, C. Hazirbas, and D. Cremers. Learning proximal operators: Using denoising networks for regularizing inverse imaging problems. In *Proc. IEEE Int. Conf. Comp. Vis. (ICCV)*, pages 1799–1808, Venice, Italy, October 22-29, 2017. 1, 2

[27] C. Metzler, P. Schniter, A. Veeraraghavan, and R. Baraniuk. prDeep: Robust phase retrieval with a flexible deep network. In *Proc. 35th Int. Conf. Machine Learning (ICML)*, pages 3501–3510, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. 1, 3, 4

[28] C. A. Metzler, A. Maleki, and R. G. Baraniuk. From denoising to compressed sensing. *IEEE Trans. Inf. Theory*, 62(9):5117–5144, September 2016. 2

[29] J. J. Moreau. Proximité et dualité dans un espace Hilbertien. *Bull. Soc. Math. France*, 93:273–299, 1965. 2

[30] M. K. Ng, P. Weiss, and X. Yuan. Solving constrained total-variation image restoration and reconstruction problems via alternating direction methods. *SIAM J. Sci. Comput.*, 32(5):2710–2736, August 2010. 1

[31] S. Ono. Primal-dual plug-and-play image restoration. *IEEE Signal. Proc. Let.*, 24(8):1108–1112, 2017. 2

[32] N. Parikh and S. Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3):123–231, 2014. 4

[33] E. T. Reehorst and P. Schniter. Regularization by denoising: Clarifications and new interpretations. *IEEE Trans. Comput. Imag.*, 5(1):52–67, Mar. 2019. 1, 3

[34] Y. Romano, M. Elad, and P. Milanfar. The little engine that could: Regularization by denoising (RED). *SIAM J. Imaging Sci.*, 10(4):1804–1844, 2017. 1, 2, 3

[35] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D*, 60(1–4):259–268, November 1992. 1, 2

[36] E. K. Ryu, J. Liu, S. Wang, X. Chen, Z. Wang, and W. Yin. Plug-and-play methods provably converge with properly trained denoisers. In *Proc. 36th Int. Conf. Machine Learning (ICML)*, pages 5546–5557, 2019. 2

[37] S. Sreehari, S. V. Venkatakrishnan, B. Wohlberg, G. T. Buzzard, L. F. Drummy, J. P. Simmons, and C. A. Bouman. Plug-and-play priors for bright field electron tomography and sparse interpolation. *IEEE Trans. Comput. Imaging*, 2(4):408–423, December 2016. 1, 2

[38] J. L. Starck, E. Pantin, and F. Murtagh. Deconvolution in astronomy: A review. *Pub. Astron. Soc. Pacific*, 114(800):1051–1069, October 2002. 1

[39] Y. Sun, J. Liu, and U. S. Kamilov. Block coordinate regularization by denoising. In *Proc. Advances in Neural Information Processing Systems 32*, Vancouver, BC, Canada, Dec. 2019. 1, 3, 4, 7

[40] Y. Sun, B. Wohlberg, and U. S. Kamilov. An online plug-and-play algorithm for regularized image reconstruction. *IEEE Trans. Comput. Imaging*, 2019. 1, 2, 3, 4

[41] Y. Sun, S. Xu, Y. Li, L. Tian, B. Wohlberg, and U. S. Kamilov. Regularized fourier ptychography using an online plug-and-play algorithm. In *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Process. (ICASSP)*, pages 7665–7669, Brighton, UK, May 12-17, 2019. 1

[42] A. Teodoro, J. M. Bioucas-Dias, and M. Figueiredo. Scene-adapted plug-and-play algorithm with convergence guarantees. In *Proc. IEEE Int. Workshop on Machine Learning for Signal Processing*, pages 1–6, Tokyo, Japan, September 25-28, 2017. 2

[43] A. M. Teodoro, J. M. Biocas-Dias, and M. A. T. Figueiredo. Image restoration and reconstruction using variable splitting and class-adapted image priors. In *Proc. IEEE Int. Conf. Image Proc. (ICIP)*, pages 3518–3522, Phoenix, AZ, USA, September 25-28, 2016. 1

[44] L. Tian and L. Waller. 3D intensity and phase imaging from light field measurements in an LED array microscope. *Optica*, 2:104–111, 2015. 1

[45] R. Tibshirani. Regression and selection via the lasso. *J. R. Stat. Soc. Series B (Methodological)*, 58(1):267–288, 1996. 2

[46] S. V. Venkatakrishnan, C. A. Bouman, and B. Wohlberg. Plug-and-play priors for model based reconstruction. In *Proc. IEEE Global Conf. Signal Process. and Inf. Process. (GlobalSIP)*, pages 945–948, Austin, TX, USA, December 3-5, 2013. 1, 2

[47] X. Xu and U. S. Kamilov. Signprox: One-bit proximal algorithm for nonconvex stochastic optimization. In *IEEE Int. Conf. Acoustics, Speech and Signal Process. (ICASSP)*, pages 7800–7804, Brighton, UK, May 2019. 4

[48] 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.*, 26(7):3142–3155, July 2017. 1

[49] K. Zhang, W. Zuo, S. Gu, and L. Zhang. Learning deep CNN denoiser prior for image restoration. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pages 3929–3938, Honolulu, USA, July 21-26, 2017. 1