Rethinking Gradient Step Denoiser: Towards Truly Pseudo-Contractive Operator

Shuchang Zhang
College of Science
National University of Defense Technology
Changsha, 410073
zhangshuchang19@nudt.edu.cn

Yaoyun Zeng*
College of Science
National University of Defense Technology
Changsha, 410073
yaoyun_zeng@nudt.edu.cn

Kangkang Deng
College of Science
National University of Defense Technology
Changsha, 410073
freedeng1208@gmail.com

Hongxia Wang †
College of Science
National University of Defense Technology
Changsha, 410073
wanghongxia@nudt.edu.cn

Abstract

Learning pseudo-contractive denoisers is a fundamental challenge in the theoretical analysis of Plug-and-Play (PnP) methods and the Regularization by Denoising (RED) framework. While spectral methods attempt to address this challenge using the power iteration method, they fail to guarantee the truly pseudo-contractive property and suffer from high computational complexity. In this work, we rethink gradient step (GS) denoisers and establish a theoretical connection between GS denoisers and pseudo-contractive operators. We show that GS denoisers, with the gradients of convex potential functions parameterized by input convex neural networks (ICNNs), can achieve truly pseudo-contractive properties. Furthermore, we integrate the learned truly pseudo-contractive denoiser into the RED-PRO (RED via fixed-point projection) model, definitely ensuring convergence in terms of both iterative sequences and objective functions. Extensive numerical experiments confirm that the learned GS denoiser satisfies the truly pseudo-contractive property and, when integrated into RED-PRO, provides a favorable trade-off between interpretability and empirical performance on inverse problems.

1 Introduction

The pseudo-contractive operators constitute a wide class of operators that arise in iterative methods for solving fixed point problems [17]. Here, we recall that an operator $T : \mathcal{H} \to \mathcal{H}$ is d-pseudo-contractive (d-PC) if there exists a constant $d \in (-\infty, 1]$ such that for any $\mathbf{x}, \mathbf{y} \in \mathcal{H}$, it holds that [12, 31]

$$\|T(\mathbf{x}) - T(\mathbf{y})\|^2 \leq \|\mathbf{x} - \mathbf{y}\|^2 + d\|(\mathbf{x} - T(\mathbf{x})) - (\mathbf{y} - T(\mathbf{y}))\|^2, \tag{1}$$

where $\mathcal{H}$ is a Hilbert space. When $d < 1$, the operator T is called d-strictly PC (d-SPC) operator [17]. The operator T is nonexpansive and firmly nonexpansive (FNE) when $d = 0$ and $d = -1$ [6]. The pseudo-contractive assumption has played an important role in the convergence analysis of PnP methods [61, 54, 14, 52, 57, 60, 40, 48, 58, 49, 30, 4, 11, 62, 47, 53, 10] and RED

*Equal contribution
†Corresponding author

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

framework [51, 50, 21, 42]. Firmly nonexpansive (FNE) denoisers [60, 49, 58, 11, 10], averaged nonexpansive denoisers [57, 48, 30, 47], nonexpansive denoisers [54, 51, 14, 50, 40, 53], and contractive residuals or operators [52, 42, 35, 4] all belong to the class of pseudo-contractive operators [17]. The demicontractive assumption (see Definition 3.1) further generalizes the SPC concept, enabling the inclusion of a broader range of denoisers. Based on the fixed-point projection of demicontractive denoisers, Cohen et al. proposed the RED-PRO model [21] to theoretically bridge between PnP and RED priors. The indicator of the fixed-point set can serve as a regularizer and plays an increasingly important role in solving inverse problems. It is critically important to develop an efficient technique for testing the demicontractivity of a given mapping [21]. Of course, obtaining a pseudo-contractive denoiser also ensures the demicontractive property, as pseudo-contractive denoisers are a subclass of demicontractive operators shown in Figure 1. However, ensuring nonexpansivity, or more generally, a Lipschitz constraint on artificial neural networks is not easy in practice [49, 11]. Therefore, how to design an efficient training framework that enables neural networks to learn theoretically guaranteed denoising mappings under the weak assumption, such as (1), is a highly challenging problem. In this paper, we will try to answer the open question by establishing a theoretical connection between the GS denoiser and the pseudo-contractive denoiser. **The GS denoiser is all you need—simply parameterizing convex potential functions with ICNNs is sufficient without complex training techniques.**

Recently, spectral methods have been proposed to address the aforementioned problem [52, 49, 35, 62]. The SPC-DRUNet trained by spectral methods has achieved the state-of-the-art performance in image restoration problems [62]. Spectral methods typically use the power iteration (PI) method to compute the spectral norm of the Jacobian matrix, which is then either normalized by dividing the convolutional kernels by the spectral norm [52] or incorporated into the loss function as a penalty during training [49, 35, 62]. Ryu et al. [52] utilize a PI method, treating the convolution as a linear operator that performs a matrix-vector product. To balance speed and precision, PI only runs one iteration. They enforced the contractiveness of the residual $I - T_\theta$ of the denoiser T_θ by real spectral normalization (RealSN), which normalized the spectral norm of each layer. To obtain a FNE denoiser T_θ, which is equivalent to the nonexpansiveness of $Q_\theta = 2T_\theta - I$ (see [16, Theorem 2.2.10] and [7, Proposition 4.4]), Pesquet et al. [49] added the penalty term of the spectral norm to the loss function, i.e.,

$$\mathcal{L}(\theta) = \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x}),\mathbf{n}\sim\mathcal{N}(\mathbf{0},\sigma^2\mathbf{I})} \left\|T_\theta(\mathbf{x}+\mathbf{n}) - \mathbf{x}\right\|^2 + \lambda \max\{\left\|J_{Q_\theta}(\tilde{\mathbf{x}})\right\|_*, 1-\varepsilon\},$$

where $p(\mathbf{x})$ is the data distribution, $J_{Q_\theta}(\tilde{\mathbf{x}})$ is the Jacobian of Q_θ at point $\tilde{\mathbf{x}} = \varrho\mathbf{x} + (1-\varrho)T_\theta(\mathbf{x}+\mathbf{n})(\varrho \in [0,1])$, the noise $\mathbf{n}$ follows a Gaussian distribution with mean $\mathbf{0}$ and standard deviation $\sigma\mathbf{I}$, $\lambda > 0$ is a penalization parameter, and the parameter $\varepsilon \in (0,1)$ controls the penalty term. Then the denoiser T_θ is a resolvent of a maximally monotone operator (MMO). Later, the GS denoiser $T_\theta = I - \nabla\psi_\theta$ was proposed [20, 34], where $\psi_\theta : \mathbb{R}^n \to \mathbb{R}\cup\{\infty\}$ is parameterized by the differentiable neural network such as ICNN [2] and DRUNet [67]. Based on the GS denoiser, Hurault et al. proposed the proximal DRUNet (Prox-DRUNet) [35], which requires that $\nabla\psi_\theta$ is L-contractive with $L < 1$. They fine-tuned the previously trained GS denoiser by the spectral method with the following loss:

$$\mathcal{L}(\theta) = \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x}),\mathbf{n}\sim\mathcal{N}(\mathbf{0},\sigma^2\mathbf{I})} \left\|T_\theta(\mathbf{x}+\mathbf{n}) - \mathbf{x}\right\|^2 + \lambda \max\{\left\|J_{\nabla\psi_\theta}(\mathbf{x}+\mathbf{n})\right\|_*, 1-\varepsilon\}.$$

During training, the spectral norm $\left\|J_{\nabla\psi_\theta}(\mathbf{x}+\mathbf{n})\right\|_*$ is estimated with 50 iterations. Wei et al. trained the d-SPC ($d < 1$) denoiser by the spectral method with the following loss [62]

$$\mathcal{L}(\theta) = \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x}),\mathbf{n}\sim\mathcal{N}(\mathbf{0},\sigma^2\mathbf{I})} \left\|T_\theta(\mathbf{x}+\mathbf{n}) - \mathbf{x}\right\|^2 + \lambda \max\{\left\|dI + (1-d)J_{T_\theta}(\mathbf{x}+\mathbf{n})\right\|_*, 1-\varepsilon\}.$$

However, these spectral methods constrain the spectral norm using a finite number of training samples $\{\mathbf{x}_i, \mathbf{x}_i^*\}_{i=1}^{N_0}$ (noisy and clean image pairs), rather than all samples from the entire space. It is unable to accurately constrain the global spectral norm such as $\max_{\mathbf{x}\in\mathbb{R}^n} \left\|2J_{T_\theta}(\mathbf{x}) - I\right\|_*$ in [49], which may violate the SPC property. Moreover, PI methods have notable drawbacks: they can be computationally expensive [62], are not fully deterministic due to random initialization, and intermediate iterations do not guarantee a strict upper bound on the spectral norm [25].

Amos et al. [2] introduced ICNNs, which guarantee the convexity by imposing non-negative constraints on network weights and employing convex, non-decreasing activation functions. The gradients of ICNNs possess universal approximation capabilities [33], making them particularly valuable in data-driven optimal transport [44, 33]. The GS denoiser leverages ICNNs to learn gradients of convex implicit regularizers [20]. Fang et al. [28] further proposed learned proximal networks (LPN), which

are parameterized by the gradient of ICNN. In this work, we rethink the GS denoiser $T_\theta = I - \nabla\psi_\theta$, where ψ_θ is an ICNN, that can serve as a truly SPC denoiser. We are the first to establish the essential connection between SPC denoisers and GS denoisers. The main contributions of this work are summarized as follows:

1. **Theoretical contributions.** We rethink that the GS denoiser corresponds to a truly SPC denoiser, as demonstrated in Proposition 4.1. We also theoretically propose another novel construction method for truly pseudo-contractive denoisers, as presented in Proposition 4.3. This work is the first to theoretically confirm the feasibility of training truly SPC denoisers and to practically address the unresolved challenge of training demicontractive denoisers posed by the RED-PRO model [21].
2. **Algorithms and applications.** We integrate the learned truly SPC denoiser into the RED-PRO model and propose Algorithm 1 for solving imaging inverse problems. Theorem 4.6 further establishes the convergence of objective functions in RED-PRO, complementing the prior work [21] that focused solely on sequence convergence.
3. **Experimental validation.** Through extensive numerical experiments, we validate the SPC property of the GS denoiser compared to spectral methods [49, 62]. Our results highlight the ability of the method to balance interpretability and performance in addressing inverse problems.

2 Related works

Regularization models are important in image restoration problems, which can be formulated as follows:

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) + \lambda g(\mathbf{x}), \tag{2}$$

where f is the data fidelity, g is the regularizer, and $\lambda > 0$ is a regularization parameter. For example, the data fidelity $f(\mathbf{x}) = \frac{1}{2\sigma^2}\left\|\mathbf{A}\mathbf{x} - \mathbf{y}\right\|^2$ corresponds to $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{n}$ with given linear operator $\mathbf{A}$ and Gaussian noise $\mathbf{n}$.

2.1 PnP methods

PnP methods that combine splitting algorithms with denoiser priors have been widely applied in practical problems [61, 54, 1, 37, 63, 28, 64, 38] and have achieved state-of-the-art performance in inverse imaging tasks [67, 26, 35, 59]. Venkatakrishnan et al. [61] first proposed the PnP method using the alternating direction method of multipliers (ADMM). PnP methods solve the problem (2) by replacing the proximal operator $\mathrm{prox}_g(\mathbf{x})$ with denoisers, such as non-local means (NLM) [13] and block-matching 3D filtering (BM3D) [24], within ADMM or forward-backward splitting (FBS), also known as the proximal gradient method [9]. Zhang et al. [67] extended PnP methods using trained DNNs to achieve state-of-the-art performance in image restoration [67]. The convergence of PnP methods has been extensively studied. Sreehari et al. established theoretical conditions for PnP-ADMM, requiring that the Jacobian ∇D_σ be a doubly stochastic and symmetric matrix with all real eigenvalues in the range $(0, 1]$ [54]. Buzzard et al. provided a Consensus Equilibrium interpretation on denoiser priors [14]. Chen et al. [18] analyzed fixed-point convergence under bounded denoisers, while Ryu et al. [52] proved the fixed-point convergence of PnP-FBS and PnP-ADMM using the Banach contraction principle, assuming strongly convex data fidelity and nonexpansiveness of the residual of DnCNN [68]. Diffusion models (DMs) [32] can also act as efficient PnP priors, which have been widely used in physical sciences such as black hole imaging problems [64, 69], and image restoration [70].

2.2 RED framework

Romano et al. [51] introduced the well-known RED framework, which constructs an explicit objective function and flexibly incorporates various denoisers, such as NLM [13], BM3D [24], or trainable nonlinear reaction diffusion (TNRD) [19]. The RED framework has been widely used in computational imaging [55, 56, 41, 42]. Reehorst and Schniter et al. [50] highlighted that some existing denoisers do not satisfy the assumptions of RED, and provided the SMD (score-matching by denoising) interpretation. To explore the relationship between PnP priors and RED, the RED-PRO

framework was proposed from the perspective of fixed-point projection. Since the fixed-point set $\mathrm{Fix}(T) = \{\mathbf{x} \in \mathcal{H} : T(\mathbf{x}) = \mathbf{x}\}$ is closed and convex [21, Theorem 3.8], RED-PRO reformulates RED as a convex optimization problem for image restoration via fixed-point projection, the regularizer g in (2) is the indicator of the fixed-point set $\mathrm{Fix}(T)$, i.e.,

$$g(\mathbf{x}) = \begin{cases} 0, & \text{if } \mathbf{x} \in \mathrm{Fix}(T), \\ \infty, & \text{otherwise.} \end{cases}$$

Building on the idea of RED, the GS denoiser $T_\theta = I - \nabla\psi_\theta$ was proposed [20, 34], where ψ_θ is a differentiable function. Based on the GS denoiser, Hurault et al. [35] established the convergence theory of PnP methods. He et al. proposed simultaneous local and nonlocal RED (SLN-RED) for image restoration [29]. To avoid tuning the regularization parameter, Cascarano et al. proposed the constrained RED called CRED [15] based on the discrepancy principle [27].

3 Preliminaries

Cohen et al. first introduced the following demicontractive assumption on denoisers [21]. Here is the definition of demicontractive operators.

Definition 3.1. The mapping $T : \mathcal{H} \to \mathcal{H}$ is d-demicontractive with $d < 1$, if for any $\mathbf{x} \in \mathcal{H}$ and $\mathbf{z} \in \mathrm{Fix}(T)$ it holds that

$$\|T(\mathbf{x}) - \mathbf{z}\|^2 \leq \|\mathbf{x} - \mathbf{z}\|^2 + d\,\|T(\mathbf{x}) - \mathbf{x}\|^2 .$$

The d-SPC operator is a d-demicontractive operator.

Let $\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3, \mathcal{C}_4$ denote the classes of all operators $T : \mathcal{H} \to \mathcal{H}$ satisfying the assumptions of demicontractive, SPC, NE, and FNE, respectively. For example, consider the inclusion defined as follows. Let $T \in \mathcal{C}_4$ be arbitrary. Then, for all $\mathbf{x}, \mathbf{y} \in \mathcal{H}$, it holds that

$$\|T(\mathbf{x}) - T(\mathbf{y})\|^2 \leq \|\mathbf{x} - \mathbf{y}\|^2 - \|(\mathbf{x} - T(\mathbf{x})) - (\mathbf{y} - T(\mathbf{y}))\|^2 \implies \|T(\mathbf{x}) - T(\mathbf{y})\| \leq \|\mathbf{x} - \mathbf{y}\|,$$

which means $T \in \mathcal{C}_3$. Therefore, $\mathcal{C}_4 \subset \mathcal{C}_3$. The relationship between different classes of operators is shown in Figure 1.

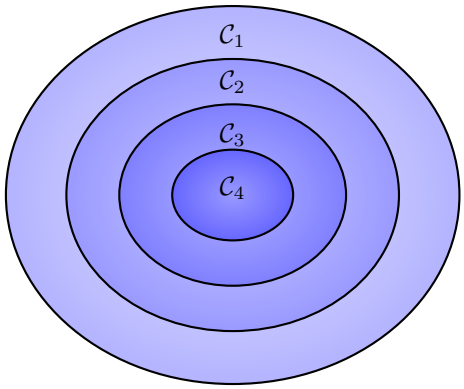


Figure 1: Relationship between different classes of operators.

The operator $T : \mathcal{H} \to \mathcal{H}$ is called conically λ-averaged for $\lambda > 0$ [5] if there exists a nonexpansive operator U such that $T = (1-\lambda)I + \lambda U$, where I denotes the identity. In particular, when $\lambda \in (0, 1)$ the operator is λ-averaged, a class that plays an important role in fixed-point algorithms [22, 65, 66, 8, 23]. If U is FNE, then $T = (1-\lambda)I + \lambda U$ is λ-relaxed FNE (λ-RFNE).

Next, we give some equivalent relationships between d-SPC operators, conically averaged operators, and RFNE operators demonstrated in Proposition 3.2. We give the proof in Appendix A.

Proposition 3.2 ([17])**.** *Let $\lambda = \frac{1}{1-d}$. Then the following statements are equivalent:*

(i) Let T be d-SPC ($d < 1$).

(ii) T *is a conically* λ*-averaged operator.*

(iii) T *is a* 2λ*-RFNE operator.*

The following Proposition 3.3 serves as a key connection for exploring how to learn a truly pseudo-contractive denoiser. Please see the proof in Appendix B.

Proposition 3.3 ([16]). *Let* $R = I - T$ *and* $\mu > 0$*, then* $T : \mathcal{H} \to \mathcal{H}$ *is* μ*-RFNE if and only if for all* $\mathbf{x}, \mathbf{y} \in \mathcal{H}$*,*

$$\langle \mathbf{x} - \mathbf{y}, R(\mathbf{x}) - R(\mathbf{y}) \rangle \geq \frac{1}{\mu} \|R(\mathbf{x}) - R(\mathbf{y})\|^2. \tag{3}$$

The residual $R = I - T$ is also called $\frac{1}{\mu}$-cocoercive in (3). Figure 2 shows all equivalence relationships between $1 - \frac{1}{\lambda}$-SPC operator, conically λ-averaged operator, 2λ-RFNE operator, and $\frac{1}{2\lambda}$-cocoercive residual, where $\lambda = \frac{1}{1-d}$.

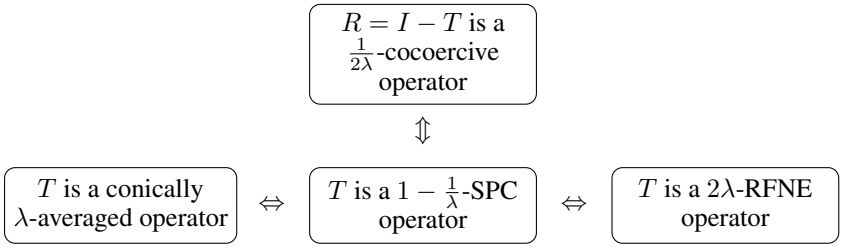


Figure 2: Equivalent relationships.

In the following Proposition 3.4, we introduce an important property of α-averaged operators for the convergence rate of objective functions about the RED-PRO model. The proof is given in Appendix C.

Proposition 3.4 ([66]). *Let* T *be a* α*-averaged operator with* $\alpha \in (0, 1)$*, then* T *is a* $\frac{1-\alpha}{\alpha}$*-strongly quasi-nonexpansive operator, i.e., for any* $\mathbf{z} \in \mathrm{Fix}(T)$*, it holds that*

$$\|T(\mathbf{x}) - \mathbf{z}\|^2 \leq \|\mathbf{x} - \mathbf{z}\|^2 - \frac{1-\alpha}{\alpha} \|\mathbf{x} - T(\mathbf{x})\|^2. \tag{4}$$

Finally, we give several equivalent characterizations of the L-smoothness property over the entire space $\mathcal{H}$. Here is the definition of L-smoothness.

Definition 3.5 ([9]). Let $L > 0$. The function $h : \mathcal{H} \to \mathbb{R} \cup \{\infty\}$ is L-smooth if it is differentiable over $\mathcal{H}$ and satisfies

$$\|\nabla h(\mathbf{x}) - \nabla h(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|, \quad \forall \mathbf{x}, \mathbf{y} \in \mathcal{H}. \tag{5}$$

Theorem 3.6 ([9]). *Let* $h : \mathcal{H} \to \mathbb{R} \cup \{\infty\}$ *be a convex function, differentiable over* $\mathcal{H}$*, and let* $L > 0$*. Then the following claims are equivalent:*

(i) h *is* L*-smooth.*

(ii) For all $\mathbf{x}, \mathbf{y} \in \mathcal{H}$*,*

$$\langle \nabla h(\mathbf{x}) - \nabla h(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq \frac{1}{L} \|\nabla h(\mathbf{x}) - \nabla h(\mathbf{y})\|^2. \tag{6}$$

4 Learned truly SPC denoiser

4.1 The GS denoiser is all you need

In this section, we will prove that the GS denoiser $T_\theta = I - \nabla \psi_\theta$ in which ψ_θ is an ICNN, can precisely correspond to a truly SPC denoiser. In [20], although Cohen et al. proposed the GS denoiser early, they failed to explore its connection with the SPC denoiser. The pursuit of learning SPC denoisers, once perceived as distant, is now within reach. Two inequalities (3) and (6) form a crucial bridge, enabling us to establish the essential connection between the GS denoiser and the SPC denoiser. We give the following result to theoretically guarantee the SPC property of the GS denoiser.

Proposition 4.1. *Consider a scalar-valued $(K+1)$-layered neural network $\psi_\theta : \mathbb{R}^n \to \mathbb{R}$ defined by $\psi_\theta(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{z}_K + b$ and the recursion*

$$\mathbf{z}_1 = \phi(\mathbf{H}_1\mathbf{x} + \mathbf{b}_1), \quad \mathbf{z}_k = \phi(\mathbf{W}_k\mathbf{z}_{k-1} + \mathbf{H}_k\mathbf{x} + \mathbf{b}_k), \quad k = 2, 3, \ldots, K,$$

where $\Theta = \{\mathbf{w}, b, \{\mathbf{W}_k\}_{k=2}^K, \{\mathbf{H}_k\}_{k=1}^K, \{\mathbf{b}_k\}_{k=1}^K\}$ are learnable parameters, $\phi : \mathbb{R} \to \mathbb{R}$ is a convex, non-decreasing and continuously differentiable scalar function, which operates pointwise. Assume that all entries of $\mathbf{W}_k$ and $\mathbf{w}$ are non-negative, and let ψ_θ be L_θ-smooth, then the GS denoiser $T_\theta = I - \nabla\psi_\theta$ is $\frac{L_\theta - 2}{L_\theta}$-SPC operator.

Proof. Since $\mathbf{W}_k$ and $\mathbf{w}$ are non-negative, it follows that ψ_θ is convex from [2, Proposition 1]. Since ψ_θ is L_θ-smooth, by (6), we have

$$\langle \nabla\psi_\theta(\mathbf{x}) - \nabla\psi_\theta(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq \frac{1}{L_\theta} \|\nabla\psi_\theta(\mathbf{x}) - \nabla\psi_\theta(\mathbf{y})\|^2.$$

Let $T_\theta = I - \nabla\psi_\theta$, by Proposition 3.2 and recall (3) in Proposition 3.3, we directly derive $\frac{2}{1-d} = L_\theta$, i.e., $d = \frac{L_\theta - 2}{L_\theta}$. Therefore, the denoiser T_θ is a $\frac{L_\theta - 2}{L_\theta}$-SPC operator, which completes the proof. □

Given the ICNN ψ_θ, we train the GS denoiser $T_\theta = I - \nabla\psi_\theta$ with the following loss function

$$\mathcal{L}(\theta) = \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x}), \mathbf{n}\sim\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})} \left\| T_\theta(\mathbf{x} + \mathbf{n}) - \mathbf{x} \right\|^2.$$

As previously proposed in [20], the training process is straightforward and does not require any additional spectral norm penalty.

Our Proposition 4.1 is the first to realize a truly $\frac{L_\theta - 2}{L_\theta}$-SPC operator via the ICNN GS denoiser, whose assumption is weaker than FNE and nonexpansive, and thus easier to satisfy in practice. Once the GS denoiser meets the $\frac{L_\theta - 2}{L_\theta}$-SPC condition, the RED-PRO framework automatically guarantees sequence convergence and objective convergence rate, as shown in Theorems 4.5 and 4.6, without requiring stronger FNE or nonexpansive assumptions. The result of Proposition 4.1 can further benefit existing PnP/RED theoretical works by enabling the ICNN GS denoiser to satisfy stronger FNE or averaged nonexpansive assumptions in two ways:

- Controlling $L_\theta \leq 1$, e.g., by normalizing convolution kernels via spectral methods or penalizing the network's Lipschitz constant in the loss function, so that the FNE and nonexpansive assumptions required in [58, Assumption 2], [54, Theorem III.1], and [50, Lemma 5] are met;
- Estimating L_θ via the power method and tuning the weight $w < \frac{2}{L_\theta}$ so that $T_w = wT_\theta + (1-w)I$ is a $\frac{wL_\theta}{2}$-averaged operator, thus satisfying the averaged operator assumptions in [57, Assumption 2(b)] and [48, Theorem 3.5, Theorem 3.6]

Therefore, Proposition 4.1 provides valuable practical guidance for existing PnP/RED theoretical works that require FNE, and averaged nonexpansive assumption.

Remark 4.2. Although we use the same GS denoiser as in [20], the difference is that we are the first to theoretically address the relationship between d-SPC denoisers and GS denoisers, and also practically demonstrate the feasibility of training SPC denoisers that fully satisfy the demicontractive condition required by RED-PRO [21]. We believe that this theoretical finding is valuable.

In contrast to spectral methods [52, 49, 35, 62], our approach, like [20], directly embeds the underlying mathematical structure, i.e., (3) and (6), into the denoiser, thereby naturally satisfying the SPC property. This eliminates the need for adding extra penalty terms in the loss function and overcomes the limitation that spectral methods cannot constrain the global spectral norm.

However, as pointed out in [20], one limitation is that the non-negative weights may constrain the expressivity of ICNNs. We are now theoretically give another alternative construction method for any truly pseudo-contractive neural networks.

Proposition 4.3. *Let $R_\theta : \mathbb{R}^n \to \mathbb{R}^n$ be a L_θ-Lipschitz continuous convolutional neural networks. Denote $\tilde{R}_\theta = R_\theta + L_\theta I$, then the denoiser $T_\theta = I - \tilde{R}_\theta$ is a pseudo-contractive operator, i.e.,*

$$\|T_\theta(\mathbf{x}) - T_\theta(\mathbf{y})\|^2 \leq \|\mathbf{x} - \mathbf{y}\|^2 + \|(\mathbf{x} - T_\theta(\mathbf{x})) - (\mathbf{y} - T_\theta(\mathbf{y}))\|^2.$$

Proof. Since R_θ is L_θ-Lipschitz continuous, for any $\mathbf{x}, \mathbf{y} \in \mathcal{H}$ we have

$$\|R_\theta(\mathbf{x}) - R_\theta(\mathbf{y})\| \leq L_\theta \|\mathbf{x} - \mathbf{y}\|, \tag{7}$$

by (7) and Cauthy-Schwarz inequality, we have

$$\|\langle R_\theta(\mathbf{x}) - R_\theta(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle\| \leq L_\theta \|\mathbf{x} - \mathbf{y}\|^2.$$

We can derive that

$$\langle R_\theta(\mathbf{x}) - R_\theta(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle + L_\theta \|\mathbf{x} - \mathbf{y}\|^2 \geq 0.$$

Thus, we have

$$\langle \tilde{R}_\theta(\mathbf{x}) - \tilde{R}_\theta(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle = \langle R_\theta(\mathbf{x}) - R_\theta(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle + L_\theta \|\mathbf{x} - \mathbf{y}\|^2 \geq 0.$$

By [7, Example 20.8], it follows that $T_\theta = I - \tilde{R}_\theta$ is a pseudo-contractive operator. □

The core problem in Proposition 4.3 is how to training Lipschitz-constrained neural networks. Ryu et al. normalized each convolutional kernel $\mathcal{K}$ with estimated spectral norm $\|\mathcal{K}\|_*$, i.e., $\frac{\mathcal{K}}{\|\mathcal{K}\|_*}$, which can obtain 1-Lipschitz CNN. Anil et al. [3] proposed that by combining GroupSort activation functions with orthonormal weight matrices, one can construct networks that are provably 1-Lipschitz and capable of approximating any 1-Lipschitz function arbitrarily well. These methods can be used to train the 1-Lipschitz-constrained neural networks R_θ in Proposition 4.3. In this case, the Lipschitz constant L_θ is equal to 1, then $\tilde{R}_\theta = R_\theta + I$, and $\tilde{R}_\theta = R_\theta + I$ can be viewed as a residual connection, which is used to fit the noise distribution $\mathbf{n}$. That is, $\tilde{R}_\theta$ is obtained by minimizing the following loss function:

$$\min_\theta \left\{ \mathcal{L}(\theta) = \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x}), \mathbf{n}\sim\mathcal{N}(\mathbf{0},\sigma^2\mathbf{I})} \left\| \tilde{R}_\theta(\mathbf{x} + \mathbf{n}) - \mathbf{n} \right\|^2 \right\},$$

and the pseudo-contractive denoiser $T_\theta = I - \tilde{R}_\theta$ is constructed. Moreover, Delattre et al. [25] controlled a L-Lipschitz convolutional kernel $\mathcal{K}_j (1 \leq j \leq l)$, the training loss becomes: $\mathcal{L}(\theta) + \mu_{\text{reg}} \sum_{j=1}^{l} \mathcal{L}_{\text{reg}}(\mathcal{K}_j)$ with

$$\mathcal{L}_{\text{reg}}(\mathcal{K}_j) = \sigma_{\text{GI}}(\mathcal{K}_j) \mathbf{1}_{\sigma_{\text{GI}}(\mathcal{K}_j) > L},$$

where σ_{GI} denotes the spectral norm computed by Gram iteration (GI), which is more efficient and accurate than the power method, and $x \to \mathbf{1}_{x>L}$ indicates 1 if $x > L$, and 0 otherwise.

Although the above methods yields pseudo-contractive denoisers by training Lipschitz-constrained neural networks, such networks may practically suffer from limited expressive capacity [3, 36].

Remark 4.4. According to Proposition 4.3, as long as we accurately compute the Lipschitz constant of the neural network, we can construct the pseudo-contractive neural network. The recent work [25] has shown that it is feasible to accurately obtain the Lipschitz constant of neural networks. We believe that more expressive truly pseudo-contractive neural networks with inherent interpretability shown in Proposition 4.3 will be developed in the future.

4.2 RED-PRO with the learned SPC denoiser

Based on the fixed-point projection of the demicontractive denoiser, Cohen et al. proposed the following RED-PRO model

$$\min_{\mathbf{x}\in \text{Fix}(T_\theta)} f(\mathbf{x}), \tag{8}$$

where the denoiser T_θ is assumed to be d-demicontractive ($d < 1$) and $f(\mathbf{x}) = \frac{1}{2\sigma^2} \|\mathbf{A}\mathbf{x} - \mathbf{y}\|^2$. The hybrid steepest descent algorithms [65, 66] is used to solve (8), i.e.,

$$\mathbf{z}^k = T_w(\mathbf{z}^{k-1} - \mu_k \nabla f(\mathbf{z}^{k-1})), \tag{9}$$

where $T_w = wT_\theta + (1 - w)I$, and $\{\mu_k\}_{k\in\mathbb{N}}$ is diminishing, i.e., $\sum_{k\in\mathbb{N}} \mu_k = +\infty, \lim_{k\to\infty} \mu_k = 0$, and $w \in (0, \frac{1-d}{2})$ for d-demicontractive denoiser T. Proposition 3.2 shows that T_w is $w\theta$-averaged and thus always nonexpansive for $0 < w < 1 - d$. Algorithm 1 shows the detailed steps of RED-PRO with the learned truly SPC denoiser T_θ, which is written into the following compact form,

$$\mathbf{x}^k = T_w(\mathbf{x}^{k-1}) - \mu_k \nabla f(T_w(\mathbf{x}^{k-1})), \tag{10}$$

where $T_w = wT_\theta + (1-w)I$. In fact, the above iteration (10) is equivalent to the iteration (9).

Algorithm 1 RED-PRO with the learned truly SPC denoiser

Require: initialization $\mathbf{x}^0 \in \mathbb{R}^n$, $\mu_k = \frac{c}{(1+k)^\alpha}$, $w \in (0, \frac{2}{L_\theta})$, and the GS denoiser $T_\theta = I - \nabla\psi_\theta$.

1: **for** $k = 1, 2, \cdots, K$ **do**
2: $\quad \mathbf{y}^k = (1-w)\mathbf{x}^{k-1} + wT_\theta(\mathbf{x}^{k-1})$
3: $\quad \mathbf{x}^k = \mathbf{y}^k - \mu_k \nabla f(\mathbf{y}^k)$
4: **end for**

Ensure: $\mathbf{x}^K$.

The known Theorem 4.5 provides the convergence guarantee of Algorithm 1 to an optimal solution of (2) with the learned truly d-SPC denoiser T_θ. Compared to RED-PRO [21, Theorem 4.3], we can extend the interval of w from $(0, \frac{1-d}{2})$ to $(0, 1-d)$, and explore that w depends on the L_θ-smooth property of the ICNN. We further complement the convergence rate of the objective function for RED-PRO in Theorem 4.6. The proof is given in the Appendix D.

Theorem 4.5 ([66, 21])**.** *Let $T_\theta = I - \nabla\psi_\theta$ be a continuous d-SPC denoiser, and $f(\mathbf{x})$ be a proper convex l.s.c. differentiable function with L-Lipschitz gradient. Then the sequence $\{\mathbf{x}^k\}_{k\in\mathbb{N}}$ generated by Algorithm 1 converges to a solution of* (8)*.*

Theorem 4.6. *Let $\{\mathbf{x}^k\}_{k\in\mathbb{N}}$ and $\{\mathbf{y}^k\}_{k\in\mathbb{N}}$ be sequences generated by Algorithm 1. Assume that $\mathcal{S} = \arg\min_{\mathbf{x}\in \mathrm{Fix}(T_\theta)} f(\mathbf{x})$ is the solution set of the RED-PRO model and the sequence $\{\mathbf{x}^k\}_{k\in\mathbb{N}}$ is bounded, then*

(i) For any $\mathbf{x}' \in \mathcal{S}$ and $k \geq 1$, there exist $D_1, D_2 > 0$ such that

$$u_k \leq \frac{D_1^2}{ck^{1-\alpha}} + \frac{cD_2^2}{k^\alpha},$$

where $u_k = \min\{\langle \nabla f(\mathbf{y}^j), \mathbf{y}^j - \mathbf{x}'\rangle : k \leq j \leq 2k\}$.

(ii) For any $\mathbf{x}' \in \mathcal{S}$ and $k \geq 1$, we have

$$f(\mathbf{y}^k_{best}) - f(\mathbf{x}') \leq \frac{D_1^2}{ck^{1-\alpha}} + \frac{cD_2^2}{k^\alpha},$$

where $k_{best} = \arg\min_{k\leq j\leq 2k} f(\mathbf{y}^j)$.

Remark 4.7. The boundedness of $\{\mathbf{x}^k\}_{k\in\mathbb{N}}$ in Theorem 4.6 is straightforward to verify. Specifically, we can replace T_w with $P_{[0,1]^n} \circ T_w$ in Algorithm 1, where $P_{[0,1]^n}$ denotes the metric projection onto the unit hypercube $[0,1]^n$. Thus the sequence $\{\mathbf{y}^k\}_{k\in\mathbb{N}}$ is bounded such that $\mathbf{y}^k \in [0,1]^n$. Since $\mathbf{x}^k = \mathbf{y}^k - \mu_k \nabla f(\mathbf{y}^k)$, for any $\mathbf{z} \in \mathrm{Fix}(T)$, we have $\left\|\mathbf{x}^k - \mathbf{z}\right\| \leq \left\|\mathbf{y}^k - \mathbf{z}\right\| + c\left\|\nabla f(\mathbf{y}^k)\right\|$. Therefore, the sequence $\{\mathbf{x}^k\}_{k\in\mathbb{N}}$ is also bounded.

Remark 4.8. Compared to the convergence results of RED-PRO in [21, Theorem 4.3 and Theorem 4.4], we provide a new complementary analysis by establishing the outer convergence rate of the objective function in Theorem 4.6 (ii).

5 Experiments

In this section, we present some experiments to evaluate the performance of the learned truly SPC denoiser. We benchmark it against state-of-the-art spectral methods, including MMO [49] and SPCNet [62]. Specifically, we validate the SPC property of the learned denoiser. RED-PRO with the learned truly SPC denoiser has the theoretical guarantee, which can be applied to complicated imaging inverse problems. Our primary focus is on achieving theoretical interpretability rather than pursuing state-of-the-art performance.

5.1 Implementation details

We compare spectral methods with the DnCNN architecture. SPCNet [62] adopts $d = 0.5$ for MNIST and CelebA datasets [39, 43], and $d = 0.8$ for BSD400 [45], while the MMO method [49] uses

$d = -1$ in (1). Spectral methods are configured with $\lambda = 10^{-3}$, $\varepsilon = 0.1$, and 20 PI iterations for CelebA and BSD400, or 30 iterations for MNIST. The ICNN models start with an initial learning rate of 10^{-3}, decaying to 10^{-4} after half the epochs. The DnCNN models begin with 10^{-4}, decaying to 5×10^{-5} mid-training.

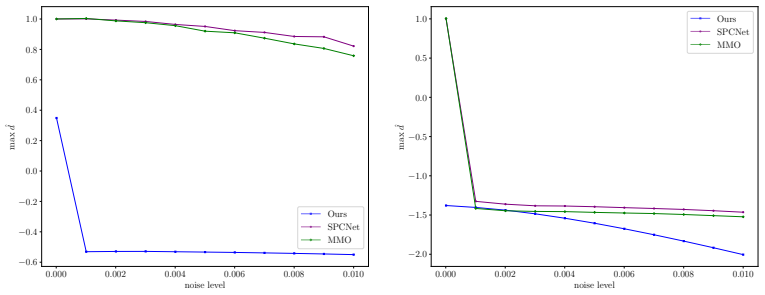
For the MNIST dataset, ICNN is implemented with four convolutional layers, each containing 64 hidden neurons and softplus activation function $\phi(x) = \frac{1}{\beta}\log(1 + \mathrm{e}^{\beta x})$ with $\beta = 10$. For BSD400 and CelebA, ICNN uses 256 hidden neurons with $\beta = 100$. All models are trained for Gaussian denoising with a noise level of $\sigma = 5/255$ and a batch size of 128. Training spans 50 epochs for MNIST and BSD400, and 30 epochs for CelebA. All experiments are conducted on one NVIDIA A800 GPU using the PyTorch framework.

5.2 Validation of SPC property

We test the SPC property of the learned GS denoiser $T_\theta = I - \nabla\psi_\theta$ with ICNN ψ_θ, MMO [49], and SPCNet [62] on two MNIST and test12 datasets. We calculate the maximum $\hat{d}$ defined by

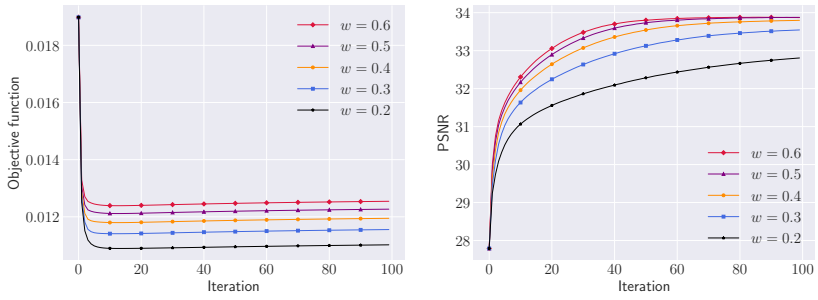
$$\hat{d} = \frac{\|T(\mathbf{x}) - T(\mathbf{y})\|^2 - \|\mathbf{x} - \mathbf{y}\|^2}{\|(\mathbf{x} - T(\mathbf{x})) - (\mathbf{y} - T(\mathbf{y}))\|^2},$$

where $\mathbf{y} = \mathbf{x} + \mathbf{n}, \mathbf{n} \sim \mathcal{N}(0, \sigma^2\mathbf{I})$. Noise levels are uniformly sampled at 11 points in the interval $[10^{-5}, 10^{-2}]$. As shown in Figure 3, the SPCNet [62] obtains a maximum $\hat{d}$ that exceeds 0.8 and 0.5 at a noise level of 10^{-5}, whereas the MMO method [49] yields a $\hat{d}$ value that exceeds -1. Therefore, denoisers trained by spectral methods violate the SPC property.



(a) MNIST (b) test12

Figure 3: Validation of the truly SPC property on two different datasets.



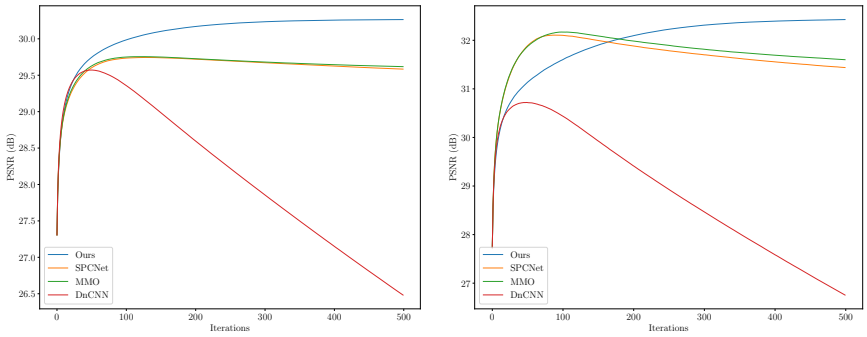
(a) (b)

Figure 4: Convergence of Algorithm 1 on one image in the CelebA dataset. (a) Objective function. (b) PSNR.

5.3 Competitive results with theoretical guarantees

In Figure 4, we show the trend of the fidelity term $\frac{1}{2\sigma^2}\|\mathbf{y}-\mathbf{A}\mathbf{x}\|^2$ and PSNR (dB) throughout the iterations with different w. We provide additional results in Appendix E.

In the following, we demonstrate the effectiveness of RED-PRO with the SPC denoiser in inverse problems, highlighting the ability to achieve competitive results while strictly satisfying theoretical constraints. Compared to non-SPC methods such as DPIR [67], which achieve high PSNR in only 8 iterations but do not converge with more iterations (see Appendix G.5 in [35]). As shown in Table 1. RED-PRO can offer both competitive performance and guaranteed convergence. We also provide visual PSNR curves shown in Figure 5 to clearly demonstrate that non-SPC methods may not converge.



(a) Man (b) House

Figure 5: PSNR curves of RED-PRO with the learned truly SPC denoiser and non-SPC methods on two images in the Gaussian deblurring task.

Table 1: Comparison of RED-PRO with various non-SPC denoisers on the Gaussian deblurring task. All non-SPC and the ICNN GS denoisers use Algorithm 1 with the same hyperparameters.

Methods	Parrot	House	Boat	Couple	Man
SPC-DnCNN [62]	25.73	31.43	28.53	28.17	29.58
MMO [49]	25.87	31.58	28.57	28.27	29.60
DnCNN [68]	25.30	26.74	26.23	25.98	26.47
DRUNet [67]	27.00	30.44	29.15	28.64	29.85
GS denoiser [35]	**27.38**	**32.58**	29.41	29.08	29.91
Ours	27.17	32.40	**29.54**	**29.17**	**30.25**

6 Conclusion

In this paper, we proposed a novel perspective to construct a truly SPC denoiser by directly embedding the underlying mathematical structure into the neural network architecture. Our theoretical analysis shown in Proposition 4.1 rethink that the known GS denoiser [20], built upon an ICNN, definitely satisfies the d-SPC property, which plays a crucial role in convergence in PnP methods and RED framework. Unlike spectral methods that require additional penalty terms and suffer from high computational cost due to costly PI iterations, our method naturally guarantees interpretability while offering the advantage in terms of time complexity. Furthermore, our theoretical insight shown in Proposition 4.3 can pave the way for future research in developing interpretative neural networks for imaging inverse problems.

Acknowledgments

We sincerely thank the Associate Professor Hui Zhang for valuable discussions. We sincerely thank four anonymous reviewers for their valuable and constructive feedback, which has greatly improved our work. This work was supported by the following grants: the National Key Research and Development Program of China (No. 2020YFA0713504), the National Natural Science Foundation of China (Grant No. 12471401), and the National Natural Science Foundation of China — Young Scientists Fund (Grant No. 12401419).

References

[1] Rizwan Ahmad, Charles A. Bouman, Gregery T. Buzzard, Stanley Chan, Sizhuo Liu, Edward T. Reehorst, and Philip Schniter. Plug-and-Play Methods for Magnetic Resonance Imaging: Using Denoisers for Image Recovery. *IEEE Signal Processing Magazine*, 37(1):105–116, 2020.

[2] Brandon Amos, Lei Xu, and J. Zico Kolter. Input convex neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 146–155. PMLR, 06–11 Aug 2017.

[3] Cem Anil, James Lucas, and Roger Grosse. Sorting out lipschitz function approximation. In *International conference on machine learning*, pages 291–301. PMLR, 2019.

[4] Chirayu D. Athalye, Kunal N. Chaudhury, and Bhartendu Kumar. On the contractivity of plug-and-play operators. *IEEE Signal Processing Letters*, 30:1447–1451, 2023.

[5] Sedi Bartz, Minh N Dao, and Hung M Phan. Conical averagedness and convergence analysis of fixed point algorithms. *Journal of Global Optimization*, 82(2):351–373, 2022.

[6] Heinz H Bauschke and Jonathan M Borwein. On projection algorithms for solving convex feasibility problems. *SIAM review*, 38(3):367–426, 1996.

[7] Heinz H. Bauschke and Patrick L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer Cham, second edition, 2017.

[8] Heinz H Bauschke, Dominikus Noll, and Hung M Phan. Linear and strong convergence of algorithms involving averaged nonexpansive operators. *Journal of Mathematical Analysis and Applications*, 421(1):1–20, 2015.

[9] Amir Beck. *First-order methods in optimization*. SIAM, 2017.

[10] Younes Belkouchi, Jean-Christophe Pesquet, Audrey Repetti, and Hugues Talbot. Learning truly monotone operators with applications to nonlinear inverse problems. *SIAM Journal on Imaging Sciences*, 18(1):735–764, 2025.

[11] Kristian Bredies, Jonathan Chirinos-Rodriguez, and Emanuele Naldi. Learning firmly nonexpansive operators, 2024.

[12] F.E Browder and W.V Petryshyn. Construction of fixed points of nonlinear mappings in hilbert space. *Journal of Mathematical Analysis and Applications*, 20(2):197–228, 1967.

[13] 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)*, volume 2, pages 60–65 vol. 2, 2005.

[14] Gregery T. Buzzard, Stanley H. Chan, Suhas Sreehari, and Charles A. Bouman. Plug-and-play unplugged: Optimization-free reconstruction using consensus equilibrium. *SIAM Journal on Imaging Sciences*, 11(3):2001–2020, 2018.

[15] Pasquale Cascarano, Alessandro Benfenati, Ulugbek S. Kamilov, and Xiaojian Xu. Constrained regularization by denoising with automatic parameter selection. *IEEE Signal Processing Letters*, 31:556–560, 2024.

[16] Andrzej Cegielski. *Iterative methods for fixed point problems in Hilbert spaces*, volume 2057. Springer, 2012.

[17] Andrzej Cegielski. Strict pseudocontractions and demicontractions, their properties, and applications. *Numerical Algorithms*, 95(4):1611–1642, Apr 2024.

[18] Stanley H. Chan, Xiran Wang, and Omar A. Elgendy. Plug-and-play admm for image restoration: Fixed-point convergence and applications. *IEEE Transactions on Computational Imaging*, 3(1):84–98, 2017.

[19] Yunjin Chen and Thomas Pock. Trainable nonlinear reaction diffusion: A flexible framework for fast and effective image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1256–1272, 2017.

[20] Regev Cohen, Yochai Blau, Daniel Freedman, and Ehud Rivlin. It has potential: Gradient-driven denoisers for convergent solutions to inverse problems. In *Advances in Neural Information Processing Systems*, volume 34, pages 18152–18164. Curran Associates, Inc., 2021.

[21] Regev Cohen, Michael Elad, and Peyman Milanfar. Regularization by denoising via fixed-point projection (red-pro). *SIAM Journal on Imaging Sciences*, 14(3):1374–1406, 2021.

[22] Patrick L Combettes*. Solving monotone inclusions via compositions of nonexpansive averaged operators. *Optimization*, 53(5-6):475–504, 2004.

[23] Patrick L Combettes and Isao Yamada. Compositions and convex combinations of averaged nonexpansive operators. *Journal of Mathematical Analysis and Applications*, 425(1):55–70, 2015.

[24] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE Transactions on Image Processing*, 16(8):2080–2095, 2007.

[25] Blaise Delattre, Quentin Barthélemy, Alexandre Araujo, and Alexandre Allauzen. Efficient Bound of Lipschitz Constant for Convolutional Layers by Gram Iteration. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 7513–7532. PMLR, 23–29 Jul 2023.

[26] Weisheng Dong, Peiyao Wang, Wotao Yin, Guangming Shi, Fangfang Wu, and Xiaotong Lu. Denoising prior driven deep neural network for image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(10):2305–2318, 2019.

[27] H. W. Engl. Discrepancy principles for tikhonov regularization of ill-posed problems leading to optimal convergence rates. *Journal of Optimization Theory and Applications*, 52(2):209–215, 1987.

[28] Zhenghan Fang, Sam Buchanan, and Jeremias Sulam. What's in a Prior? Learned Proximal Networks for Inverse Problems. In *The Twelfth International Conference on Learning Representations*, 2024.

[29] Liangtian He, Qinghua Zhang, Xuesong Yang, Yilun Wang, and Chao Wang. Sln-red: Regularization by simultaneous local and nonlocal denoising for image restoration. *IEEE Signal Processing Letters*, 30:578–582, 2023.

[30] Johannes Hertrich, Sebastian Neumayer, and Gabriele Steidl. Convolutional proximal neural networks and plug-and-play algorithms. *Linear Algebra and its Applications*, 631:203–234, 2021.

[31] Troy L Hicks and John D Kubicek. On the mann iteration process in a hilbert space. *Journal of Mathematical Analysis and Applications*, 59(3):498–504, 1977.

[32] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020.

[33] Chin-Wei Huang, Ricky T. Q. Chen, Christos Tsirigotis, and Aaron Courville. Convex potential flows: Universal probability distributions with optimal transport and convex optimization. In *International Conference on Learning Representations*, 2021.

[34] Samuel Hurault, Arthur Leclaire, and Nicolas Papadakis. Gradient step denoiser for convergent plug-and-play. In *International Conference on Learning Representations*, 2022.

[35] Samuel Hurault, Arthur Leclaire, and Nicolas Papadakis. Proximal denoiser for convergent plug-and-play optimization with nonconvex regularization. In *International Conference on Machine Learning*, pages 9483–9505. PMLR, 2022.

[36] Todd Huster, Cho-Yu Jason Chiang, and Ritu Chadha. Limitations of the lipschitz constant as a defense against adversarial examples. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 16–29. Springer, 2018.

[37] Ulugbek S. Kamilov, Charles A. Bouman, Gregery T. Buzzard, and Brendt Wohlberg. Plug-and-Play Methods for Integrating Physical and Learned Models in Computational Imaging: Theory, algorithms, and applications. *IEEE Signal Processing Magazine*, 40(1):85–97, 2023.

[38] Maximilian B. Kiss, Ander Biguri, Zakhar Shumaylov, Ferdia Sherry, K. Joost Batenburg, Carola-Bibiane Schönlieb, and Felix Lucka. Benchmarking learned algorithms for computed tomography image reconstruction tasks. *Applied Mathematics for Modern Challenges*, 3(0):1–43, 2025.

[39] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 2002.

[40] Jiaming Liu, Salman Asif, Brendt Wohlberg, and Ulugbek Kamilov. Recovery analysis for plug-and-play priors using the restricted eigenvalue condition. In *Advances in Neural Information Processing Systems*, volume 34, pages 5921–5933. Curran Associates, Inc., 2021.

[41] Jiaming Liu, Yu Sun, Weijie Gan, Xiaojian Xu, Brendt Wohlberg, and Ulugbek S. Kamilov. SGD-Net: Efficient Model-Based Deep Learning With Theoretical Guarantees. *IEEE Transactions on Computational Imaging*, 7:598–610, 2021.

[42] Jiaming Liu, Xiaojian Xu, Weijie Gan, shirin shoushtari, and Ulugbek Kamilov. Online deep equilibrium learning for regularization by denoising. In *Advances in Neural Information Processing Systems*, volume 35, pages 25363–25376. Curran Associates, Inc., 2022.

[43] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.

[44] Ashok Makkuva, Amirhossein Taghvaei, Sewoong Oh, and Jason Lee. Optimal transport mapping via input convex neural networks. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6672–6681. PMLR, 13–18 Jul 2020.

[45] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings eighth IEEE international conference on computer vision. ICCV 2001*, volume 2, pages 416–423. IEEE, 2001.

[46] Cynthia McCollough. Tu-fg-207a-04: overview of the low dose ct grand challenge. *Medical physics*, 43(6Part35):3759–3760, 2016.

[47] Pravin Nair and Kunal N. Chaudhury. Averaged deep denoisers for image regularization. *Journal of Mathematical Imaging and Vision*, 66(3):362–379, June 2024.

[48] Pravin Nair, Ruturaj G. Gavaskar, and Kunal Narayan Chaudhury. Fixed-point and objective convergence of plug-and-play algorithms. *IEEE Transactions on Computational Imaging*, 7:337–348, 2021.

[49] Jean-Christophe Pesquet, Audrey Repetti, Matthieu Terris, and Yves Wiaux. Learning maximally monotone operators for image recovery. *SIAM Journal on Imaging Sciences*, 14(3):1206–1237, 2021.

[50] Edward T. Reehorst and Philip Schniter. Regularization by denoising: Clarifications and new interpretations. *IEEE Transactions on Computational Imaging*, 5(1):52–67, 2019.

[51] Yaniv Romano, Michael Elad, and Peyman Milanfar. The little engine that could: Regularization by denoising (red). *SIAM Journal on Imaging Sciences*, 10(4):1804–1844, 2017.

[52] Ernest Ryu, Jialin Liu, Sicheng Wang, Xiaohan Chen, Zhangyang Wang, and Wotao Yin. Plug-and-play methods provably converge with properly trained denoisers. In *International Conference on Machine Learning*, pages 5546–5557. PMLR, 2019.

[53] Ferdia Sherry, Elena Celledoni, Matthias J. Ehrhardt, Davide Murari, Brynjulf Owren, and Carola-Bibiane Schönlieb. Designing stable neural networks using convex analysis and odes. *Physica D: Nonlinear Phenomena*, 463:134159, 2024.

[54] Suhas Sreehari, S. V. Venkatakrishnan, Brendt Wohlberg, Gregery T. Buzzard, Lawrence F. Drummy, Jeffrey P. Simmons, and Charles A. Bouman. Plug-and-play priors for bright field electron tomography and sparse interpolation. *IEEE Transactions on Computational Imaging*, 2(4):408–423, 2016.

[55] Yu Sun, Jiaming Liu, and Ulugbek S. Kamilov. Block coordinate regularization by denoising. *IEEE Transactions on Computational Imaging*, 6:908–921, 2020.

[56] Yu Sun, Jiaming Liu, Yiran Sun, Brendt Wohlberg, and Ulugbek Kamilov. Async-RED: A Provably Convergent Asynchronous Block Parallel Stochastic Method using Deep Denoising Priors. In *International Conference on Learning Representations*, 2021.

[57] Yu Sun, Brendt Wohlberg, and Ulugbek S. Kamilov. An online plug-and-play algorithm for regularized image reconstruction. *IEEE Transactions on Computational Imaging*, 5(3):395–408, 2019.

[58] Yu Sun, Zihui Wu, Xiaojian Xu, Brendt Wohlberg, and Ulugbek S. Kamilov. Scalable plug-and-play admm with convergence guarantees. *IEEE Transactions on Computational Imaging*, 7:849–863, 2021.

[59] Hong Ye Tan, Subhadip Mukherjee, Junqi Tang, and Carola-Bibiane Schönlieb. Provably Convergent Plug-and-Play Quasi-Newton Methods. *SIAM Journal on Imaging Sciences*, 17(2):785–819, 2024.

[60] Matthieu Terris, Audrey Repetti, Jean-Christophe Pesquet, and Yves Wiaux. Building firmly nonexpansive convolutional neural networks. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8658–8662, 2020.

[61] Singanallur V. Venkatakrishnan, Charles A. Bouman, and Brendt Wohlberg. Plug-and-play priors for model based reconstruction. In *2013 IEEE Global Conference on Signal and Information Processing*, pages 945–948, 2013.

[62] Deliang Wei, Peng Chen, and Fang Li. Learning pseudo-contractive denoisers for inverse problems. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 52500–52524. PMLR, 21–27 Jul 2024.

[63] Zhongming Wu, Chaoyan Huang, and Tieyong Zeng. Extrapolated Plug-and-Play Three-Operator Splitting Methods for Nonconvex Optimization with Applications to Image Restoration. *SIAM Journal on Imaging Sciences*, 17(2):1145–1181, 2024.

[64] Zihui Wu, Yu Sun, Yifan Chen, Bingliang Zhang, Yisong Yue, and Katherine Bouman. Principled Probabilistic Imaging using Diffusion Models as Plug-and-Play Priors. In *Advances in Neural Information Processing Systems*, volume 37, pages 118389–118427. Curran Associates, Inc., 2024.

[65] Isao Yamada. The hybrid steepest descent method for the variational inequality problem over the intersection of fixed point sets of nonexpansive mappings. *Inherently parallel algorithms in feasibility and optimization and their applications*, 8:473–504, 2001.

[66] Isao Yamada and Nobuhiko Ogura. Hybrid steepest descent method for variational inequality problem over the fixed point set of certain quasi-nonexpansive mappings. *Numerical Functional Analysis and Optimization*, 25(7-8):619–655, 2005.

[67] Kai Zhang, Yawei Li, Wangmeng Zuo, Lei Zhang, Luc Van Gool, and Radu Timofte. Plug-and-play image restoration with deep denoiser prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[68] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Transactions on Image Processing*, 26(7):3142–3155, 2017.

[69] Hongkai Zheng, Wenda Chu, Bingliang Zhang, Zihui Wu, Austin Wang, Berthy Feng, Caifeng Zou, Yu Sun, Nikola Borislavov Kovachki, Zachary E Ross, Katherine Bouman, and Yisong Yue. Inversebench: Benchmarking plug-and-play diffusion priors for inverse problems in physical sciences. In *The Thirteenth International Conference on Learning Representations*, 2025.

[70] Yuanzhi Zhu, Kai Zhang, Jingyun Liang, Jiezhang Cao, Bihan Wen, Radu Timofte, and Luc Van Gool. Denoising diffusion models for plug-and-play image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1219–1229, June 2023.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: we rethink that the GS denoiser corresponds to a truly SPC denoiser, as demonstrated in Proposition 4.1. This work is the first to theoretically confirm the feasibility of training truly SPC denoisers and to practically address the unresolved challenge of training demicontractive denoisers posed by the RED-PRO model [21].

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: one limitation is that the non-negative weights may constrain the expressivity of ICNNs.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Please see section 2 and Appendix. We provide the full set of assumptions and a complete (and correct) proof.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: we provide Implementation details and the hyperparameter setting.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: we provide the training and test details in numerical experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: error bars are not reported because it is useless for our experiments, we mainly verify the theory.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: we just evaluate the performance of the proposed method with the state-of-the art methods.

   Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: we confirm that the research conducted in the paper conforms, in every respect, with the NeurIPS Code of Ethics

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: This is a theoretical paper, so we think there is no societal impact of the work performed.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: we do not release any data or models that have a high risk for misuse.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: we use the open source code and datasets.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: we do not release any new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: we do not conduct any crowdsourcing experiments and research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: we do not conduct any crowdsourcing experiments and research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: we do not use LLMs in this research.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A The proof of Proposition 3.2

Let $\lambda = \frac{1}{1-d}$. We first show the equivalence between conically λ-averaged and d-SPC operators. Let $S = (1-d)T + dI$, then

$$\begin{aligned}\|S(\mathbf{x}) - S(\mathbf{y})\|^2 &= \|(1-d)(T(\mathbf{x}) - T(\mathbf{y})) + d(\mathbf{x}-\mathbf{y})\|^2 \\ &= (1-d)\|T(\mathbf{x}) - T(\mathbf{y})\|^2 + d\|\mathbf{x}-\mathbf{y}\|^2 \\ &\quad - d(1-d)\|(\mathbf{x} - T(\mathbf{x})) - (\mathbf{y} - T(\mathbf{y}))\|^2\end{aligned}$$

$$\begin{aligned}S \text{ is nonexpansive } &\Leftrightarrow \|S(\mathbf{x}) - S(\mathbf{y})\|^2 \le \|\mathbf{x}-\mathbf{y}\|^2 \\ &\Leftrightarrow (1-d)\|T(\mathbf{x}) - T(\mathbf{y})\|^2 + d\|\mathbf{x}-\mathbf{y}\|^2 \\ &\quad - d(1-d)\|(\mathbf{x} - T(\mathbf{x})) - (\mathbf{y} - T(\mathbf{y}))\|^2 \le \|\mathbf{x}-\mathbf{y}\|^2 \\ &\Leftrightarrow \|T(\mathbf{x}) - T(\mathbf{y})\|^2 \le \|\mathbf{x}-\mathbf{y}\|^2 + d\|(\mathbf{x} - T(\mathbf{x})) - (\mathbf{y} - T(\mathbf{y}))\|^2 \\ &\Leftrightarrow T \text{ is } d\text{-SPC}.\end{aligned}$$

By the above equivalence, we obtain $T = (1-\lambda)I + \lambda S$ is conically λ-averaged, where S is nonexpansive.

(i) $\Leftrightarrow$ (ii): Take $d = -1$, then the FNE operator V is equivalent to $\frac{1}{2}$-averaged operator. There exists a NE operator S such that $V = \frac{S+I}{2}$, then $S = 2V - I$. Substituting $S = 2V - I$ into $T = (1-\lambda)I + \lambda S$ yields

$$T = (1-2\lambda)I + 2\lambda V,$$

hence T is (2λ)-relaxed FNE (i.e. 2λ-RFNE). The conically λ-averaged operator is equivalent to the 2λ-FNE operator.

B The proof of Proposition 3.3

Let $T = \mu S + (1-\mu)I$ for a FNE operator S. Since S is FNE, then

$$\|S(\mathbf{x}) - S(\mathbf{y})\|^2 \le \|\mathbf{x}-\mathbf{y}\|^2 - \|(\mathbf{x} - S(\mathbf{x})) - (\mathbf{y} - S(\mathbf{y}))\|^2.$$

Since

$$\begin{aligned}\|S(\mathbf{x}) - S(\mathbf{y})\|^2 &= \|(S(\mathbf{x}) - \mathbf{x}) + (\mathbf{x}-\mathbf{y}) + (\mathbf{y} - S(\mathbf{y}))\|^2 \\ &= \|S(\mathbf{x}) - \mathbf{x}\|^2 + \|\mathbf{x}-\mathbf{y}\|^2 + \|\mathbf{y} - S(\mathbf{y})\|^2 \\ &\quad + 2\langle S(\mathbf{x}) - \mathbf{x}, \mathbf{x}-\mathbf{y}\rangle + 2\langle S(\mathbf{x}) - \mathbf{x}, \mathbf{y} - S(\mathbf{y})\rangle + 2\langle \mathbf{x}-\mathbf{y}, \mathbf{y} - S(\mathbf{y})\rangle.\end{aligned}$$

and

$$\|(\mathbf{x} - S(\mathbf{x})) - (\mathbf{y} - S(\mathbf{y}))\|^2 = \|\mathbf{x} - S(\mathbf{x})\|^2 - 2\langle \mathbf{x} - S(\mathbf{x}), \mathbf{y} - S(\mathbf{y})\rangle + \|\mathbf{y} - S(\mathbf{y})\|^2.$$

From

$$\|S(\mathbf{x}) - S(\mathbf{y})\|^2 \le \|\mathbf{x}-\mathbf{y}\|^2 - \|(\mathbf{x} - S(\mathbf{x})) - (\mathbf{y} - S(\mathbf{y}))\|^2,$$

we obtain

$$\begin{aligned}\|\mathbf{x} - S(\mathbf{x})\|^2 + \|\mathbf{y} - S(\mathbf{y})\|^2 &\le \langle \mathbf{x} - S(\mathbf{x}), \mathbf{y} - S(\mathbf{y})\rangle - \langle S(\mathbf{x}) - \mathbf{x}, \mathbf{x}-\mathbf{y}\rangle \\ &\quad - \langle S(\mathbf{x}) - \mathbf{x}, \mathbf{y} - S(\mathbf{y})\rangle - \langle \mathbf{x}-\mathbf{y}, \mathbf{y} - S(\mathbf{y})\rangle \\ &= \langle S(\mathbf{y}) - \mathbf{y}, S(\mathbf{x}) - \mathbf{x}\rangle - \langle \mathbf{x}-\mathbf{y}, S(\mathbf{x}) - \mathbf{x}\rangle \\ &\quad + \langle S(\mathbf{x}) - \mathbf{x}, S(\mathbf{y}) - \mathbf{y}\rangle + \langle \mathbf{x}-\mathbf{y}, S(\mathbf{y}) - \mathbf{y}\rangle \\ &= \langle S(\mathbf{y}) - \mathbf{x}, S(\mathbf{x}) - \mathbf{x}\rangle + \langle S(\mathbf{x}) - \mathbf{y}, S(\mathbf{y}) - \mathbf{y}\rangle,\end{aligned}$$

i.e.,

$$\langle S(\mathbf{y}) - \mathbf{x}, S(\mathbf{x}) - \mathbf{x}\rangle + \langle S(\mathbf{x}) - \mathbf{y}, S(\mathbf{y}) - \mathbf{y}\rangle \ge \|S(\mathbf{x}) - \mathbf{x}\|^2 + \|S(\mathbf{y}) - \mathbf{y}\|^2. \quad (11)$$

Since $S(\mathbf{x}) - \mathbf{x} = \frac{1}{\mu}(T(\mathbf{x}) - \mathbf{x})$ and $S(\mathbf{y}) - \mathbf{y} = \frac{1}{\mu}(T(\mathbf{y}) - \mathbf{y})$, then (11) is equivalent to

$$\langle S(\mathbf{y}) - \mathbf{x}, T(\mathbf{x}) - \mathbf{x}\rangle + \langle S(\mathbf{x}) - \mathbf{y}, T(\mathbf{y}) - \mathbf{y}\rangle \ge \frac{1}{\mu}\left(\|T(\mathbf{x}) - \mathbf{x}\|^2 + \|T(\mathbf{y}) - \mathbf{y}\|^2\right). \quad (12)$$

Let $R = I - T$, then

$$\begin{aligned}\langle S(\mathbf{y}) - \mathbf{x}, T(\mathbf{x}) - \mathbf{x}\rangle &= \langle (S(\mathbf{y}) - \mathbf{y}) + (\mathbf{y} - \mathbf{x}), T(\mathbf{x}) - \mathbf{x}\rangle \\ &= \langle \mathbf{x} - \mathbf{y}, R(\mathbf{x})\rangle + \frac{1}{\mu}\langle R(\mathbf{x}), R(\mathbf{y})\rangle. \end{aligned} \tag{13}$$

Similarly, we have

$$\langle S(\mathbf{x}) - \mathbf{y}, T(\mathbf{y}) - \mathbf{y}\rangle = \langle \mathbf{y} - \mathbf{x}, R(\mathbf{y})\rangle + \frac{1}{\mu}\langle R(\mathbf{x}), R(\mathbf{y})\rangle. \tag{14}$$

By (12) and above two equations (13) and (14), we have

$$\langle \mathbf{x} - \mathbf{y}, R(\mathbf{x}) - R(\mathbf{y})\rangle \geq \frac{1}{\mu}\|R(\mathbf{x}) - R(\mathbf{y})\|^2.$$

C The proof of Proposition 3.4

Since T is α-averaged, by Proposition 3.2 (ii), then we have $\frac{1}{1-d} = \alpha \implies d = \frac{\alpha-1}{\alpha}$, i.e., the operator T is a $-\frac{1-\alpha}{\alpha}$-SPC operator,

$$\|T(\mathbf{x}) - T(\mathbf{y})\|^2 \leq \|\mathbf{x} - \mathbf{y}\|^2 - \frac{1-\alpha}{\alpha}\|(\mathbf{x} - T(\mathbf{x})) - (\mathbf{y} - T(\mathbf{y}))\|^2. \tag{15}$$

Let $\mathbf{y} = \mathbf{z} \in \mathrm{Fix}(T)$ in (15), we finally obtain

$$\|T(\mathbf{x}) - \mathbf{z}\|^2 \leq \|\mathbf{x} - \mathbf{z}\|^2 - \frac{1-\alpha}{\alpha}\|T(\mathbf{x}) - \mathbf{x}\|^2.$$

D The proof of Theorem 4.6

Proof. (i) For any $\mathbf{z} \in \mathrm{Fix}(T)$, since $\mathbf{y}^k = T_w(\mathbf{x}^{k-1})$ and T_w is $\frac{w}{1-d}$-averaged, by (4) we have

$$\left\|\mathbf{y}^k - \mathbf{z}\right\|^2 \leq \left\|\mathbf{x}^{k-1} - \mathbf{z}\right\|^2 - \frac{1-d-w}{w}\left\|\mathbf{x}^{k-1} - T_w(\mathbf{x}^{k-1})\right\|^2,$$

it follows that

$$\begin{aligned}\left\|\mathbf{x}^k - \mathbf{z}\right\|^2 &= \left\|\mathbf{y}^k - \mu_k\nabla f(\mathbf{y}^k) - \mathbf{z}\right\|^2 \\ &= \left\|\mathbf{y}^k - \mathbf{z}\right\|^2 - 2\eta_k\langle \mathbf{y}^k - \mathbf{z}, \nabla f(\mathbf{y}^k)\rangle + \mu_k^2\left\|\nabla f(\mathbf{y}^k)\right\|^2 \\ &\leq \left\|\mathbf{x}^{k-1} - \mathbf{z}\right\|^2 - \frac{1-d-w}{w}\left\|\mathbf{x}^{k-1} - T_w(\mathbf{x}^{k-1})\right\|^2 \\ &\quad - 2\eta_k\langle \mathbf{y}^k - \mathbf{z}, \nabla f(\mathbf{y}^k)\rangle + \mu_k^2\left\|\nabla f(\mathbf{y}^k)\right\|^2. \end{aligned} \tag{16}$$

According to the known conditions, there exist $D_1, D_2 > 0$ such that

$$\left\|\mathbf{y}^k - \mathbf{z}\right\| \leq \left\|\mathbf{x}^{k-1} - \mathbf{z}\right\| \leq D_1, \left\|\nabla f(\mathbf{y}^k)\right\| \leq D_2.$$

If $u_k \leq 0$, then the inequality holds. Otherwise, applying (16) with $\mathbf{z}$ replaced by $\mathbf{x}'$, we obtain, for any $j \geq 1$,

$$\left\|\mathbf{x}^j - \mathbf{x}'\right\|^2 \leq \left\|\mathbf{x}^{j-1} - \mathbf{x}'\right\|^2 - 2\mu_j\langle \nabla f(\mathbf{y}^j), \mathbf{y}^j - \mathbf{x}'\rangle + \mu_j^2 D_2^2,$$

arrange the above inequality, we obtain

$$2\mu_j\langle \nabla f(\mathbf{y}^j), \mathbf{y}^j - \mathbf{x}'\rangle \leq \left\|\mathbf{x}^{j-1} - \mathbf{x}'\right\|^2 - \left\|\mathbf{x}^j - \mathbf{x}'\right\|^2 + \mu_j^2 D_2^2.$$

Since $\mu_k \geq \mu_j$ for all $k \leq j \leq 2k-1$, summing it for all $j = k, k+1, \cdots, 2k-1$, yields

$$\begin{aligned}2\sum_{j=k}^{2k-1}\mu_j\langle \nabla f(\mathbf{y}^j), \mathbf{y}^j - \mathbf{x}'\rangle &\leq \left\|\mathbf{x}^{k-1} - \mathbf{x}'\right\|^2 - \left\|\mathbf{x}^{2k-1} - \mathbf{x}'\right\|^2 + \sum_{j=k}^{2k-1}\mu_j^2 D_2^2 \\ &\leq D_1^2 + D_2^2\sum_{j=k}^{2k-1}\mu_j^2 \\ &\leq D_1^2 + D_2^2(2k-1-k+1)\eta_k^2 \\ &\leq D_1^2 + \frac{D_2^2 c^2 k}{(1+k)^{2\alpha}} \end{aligned}$$

By the definition of u_k and $u_k \geq 0$, it follows that

$$2\sum_{j=k}^{2k-1}\mu_j\langle\nabla f(\mathbf{y}^j),\mathbf{y}^j-\mathbf{x}'\rangle \geq 2u_k\sum_{j=k}^{2k-1}\mu_j \geq 2u_k k\eta_{2k-1} = c(2k)^{1-\alpha}u_k \geq ck^{1-\alpha}u_k,$$

then

$$u_k \leq \frac{D_1^2}{ck^{1-\alpha}} + \frac{cD_2^2k^\alpha}{(1+2k+k^2)^\alpha} \leq \frac{D_1^2}{ck^{1-\alpha}} + \frac{cD_2^2}{k^\alpha}.$$

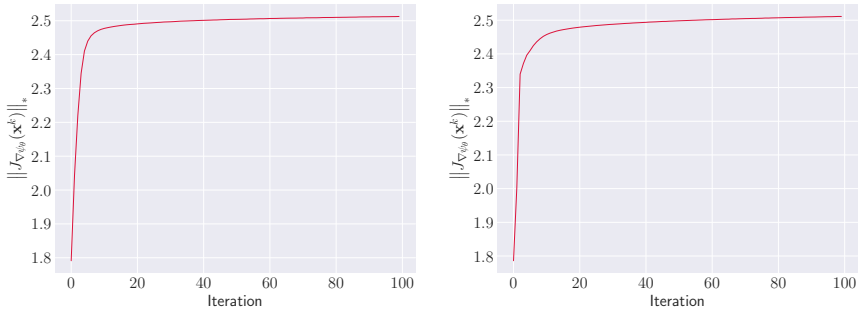
(ii) In order to prove the rate in terms of the outer objective function f, we will use (i) and for simplicity we define $\bar{k} = \arg\min\{\langle\nabla f(\mathbf{y}^j),\mathbf{y}^j-\mathbf{x}'\rangle : k \leq j \leq 2k\}$. We apply the sub-gradient inequality on the convex function f to obtain

$$f(\mathbf{y}_{best}^k) - f(\mathbf{x}') \leq f(\mathbf{y}^{\bar{k}}) - f(\mathbf{x}') \leq \langle\nabla f(\mathbf{y}^{\bar{k}}),\mathbf{y}^{\bar{k}}-\mathbf{x}'\rangle \leq \frac{D_1^2}{ck^{1-\alpha}} + \frac{cD_2^2}{k^\alpha},$$

the first inequality follows from $f(\mathbf{y}_{best}^k) \leq f(\mathbf{y}^{\bar{k}})$, and the last inequality follows from (i). □

E Other experiments about convergence

We evaluate the convergence of Algorithm 1 on the CelebA dataset for the Gaussian deblurring task with $\sigma_{blur} = 1, \sigma_{noise} = 0.02$. We set $K = 100, \mu_k = 9 \times 255\sigma_{noise}(k+1)^{-0.1}$. We first compute the spectral norm $\left\|J_{\nabla\psi_\theta}(\mathbf{x}^k)\right\|_*$, where $\{\mathbf{x}^k\}_{k=0}^K$ is the iterative sequence. Here we run the PI method with 200 iterations. As shown in Figure 6, we can estimate the tight Lipschitz constant $L_\theta = \max_{\mathbf{x}\in\mathbb{R}^n}\left\|J_{\nabla\psi_\theta}(\mathbf{x})\right\|_* \geq 2.5$, then $2/L_\theta \leq 0.8, w \in (0, 0.8)$.



(a) (b)

Figure 6: Spectral norm $\left\|J_{\nabla\psi_\theta}(\mathbf{x}^k)\right\|_*$ at the k-th iterative point $\mathbf{x}^k$ on two images.

We provide additional experiments to demonstrate the convergence of Algorithm 1 on four images in the CelebA dataset. The objective function and PSNR trends during iterations are shown in Figure 7.

F Hyperparameter Setting

We manually set the step size $\mu_k = \frac{c}{(1+k)^\alpha}$ and the weight w of Algorithm 1 to achieve the best performance on the CelebA dataset. All hyperparameters are set to be the same for all images. The hyperparameters are summarized in Table 2.

Table 2: Parameter setting for CelebA dataset.

Parameter	$\sigma_{blur}=1$, $\sigma_{noise}=.02$	$\sigma_{blur}=1$, $\sigma_{noise}=.04$	$\sigma_{blur}=2$, $\sigma_{noise}=.02$	$\sigma_{blur}=1$, $\sigma_{noise}=.04$
c	$9\times255\sigma_{noise}$	$9\times255\sigma_{noise}$	$14\times255\sigma_{noise}$	$20\times255\sigma_{noise}$
α	0.1	0.1	0.09	0.25
w	0.5	0.75	0.75	0.75

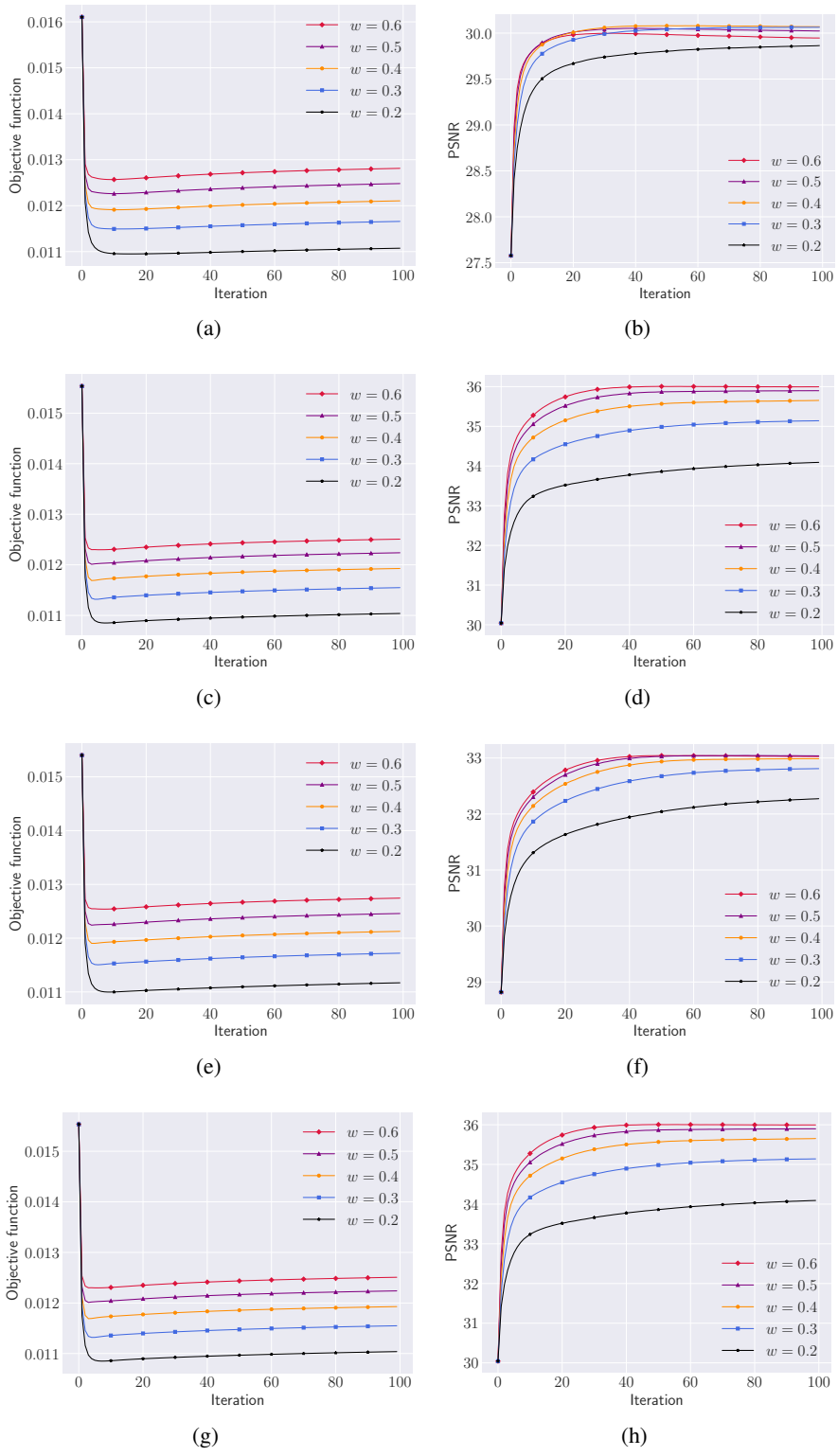


Figure 7: Convergence of Algorithm 1 on other images in the CelebA dataset. (a) Objective function. (b) PSNR.

G Compared with state-of-the-art methods

We first train the SPC denoiser using the CelebA dataset and then apply Algorithm 1 to perform deblurring. We compare with diffusion-based method DiffPIR [70], PnP-PGD [34], and DPIR [67]. As demonstrated in Table 3, we evaluate the efficacy of RED-PRO with the learned truly SPC denoiser across a range of blur intensities, noise levels, and evaluation metrics. We give the hyperparameter setting of RED-PRO in Table 2, Appendix F. We provide visual comparisons in Figure 8. DPIR achieves the best results, our method ranks second, and DiffPIR effectively restores fine details but occasionally alters facial expressions.

Table 3: Deblurring results on CelebA over 20 samples.

METHOD	$\sigma_{blur} = 1$, $\sigma_{noise} = .02$		$\sigma_{blur} = 1$, $\sigma_{noise} = .04$		$\sigma_{blur} = 2$, $\sigma_{noise} = .02$		$\sigma_{blur} = 2$, $\sigma_{noise} = .04$	
	PSNR(↑)	SSIM(↑)	PSNR(↑)	SSIM(↑)	PSNR(↑)	SSIM(↑)	PSNR(↑)	SSIM(↑)
DiffPIR	30.8 ± 2.0	$.86 \pm .03$	29.5 ± 1.8	$.82 \pm .03$	28.6 ± 2.0	$.80 \pm .05$	27.6 ± 1.8	$.77 \pm .05$
PnP-PGD	31.4 ± 1.9	$.87 \pm .02$	27.6 ± 0.9	$.71 \pm .05$	$\underline{29.9 \pm 2.3}$	$.85 \pm .05$	$\underline{28.8 \pm 2.0}$	$.81 \pm .05$
DPIR	$\mathbf{33.2 \pm 3.0}$	$\mathbf{.92 \pm .03}$	$\mathbf{31.8 \pm 2.6}$	$\mathbf{.89 \pm .04}$	$\mathbf{30.1 \pm 2.5}$	$\mathbf{.86 \pm .05}$	$\mathbf{29.1 \pm 2.2}$	$\mathbf{.83 \pm .05}$
RED-PRO	$\underline{32.4 \pm 2.8}$	$\underline{.92 \pm .03}$	30.8 ± 2.3	$\underline{.88 \pm .03}$	29.3 ± 2.3	$\underline{.86 \pm .04}$	28.4 ± 2.0	$\underline{.83 \pm .04}$

(a) Clean image (b) Degraded image (c) DiffPIR (30.07 dB)

(d) PnP-PGD (30.53 dB) (e) DPIR (31.95 dB) (f) Ours (30.93 dB)

Figure 8: Visual comparison on CelebA for Gaussian deblurring with $\sigma_{blur} = 1, \sigma_{noise} = 0.02$.

We consider a sparse-view computed tomography (CT) measurement model:

$$\min_{\mathbf{x}\in \mathrm{Fix}(T_\theta)} \frac{1}{N}\sum_{i=1}^{N} \|\mathbf{A}_i\mathbf{x}-\mathbf{b}_i\|^2,$$

where $\mathbf{b}_i \in \mathbb{R}^m$ is the measured sinogram for the i-th projection, and $\mathbf{A}_i$ is an $m\times n$ discretized Radon transform matrix. For RED-PRO, we use $\mu_k = \frac{2}{\|\mathbf{A}\|^2(1+k)^{0.01}}$, where $\mathbf{A} = [\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_N]^{\mathrm{T}}$, and $w = 0.1$. The GS denoiser is trained on the public Mayo-CT dataset [46]. We simulate CT sinograms using a parallel-beam geometry with 200 angles and 400 detectors. We compare with FBP and the recommended method in [38]. Table 4 presents the results for CT reconstruction. Despite RED-PRO's theoretical guarantees, its empirical performance in CT reconstruction may be inferior to that of PnP-ADMM.

Table 4: Numerical results for CT reconstruction on the Mayo-CT dataset, computed over 128 test images.

Method	PSNR	SSIM
FBP	20.233 ± 0.034	0.1763 ± 0.0138
RED-PRO	30.057 ± 0.488	0.8190 ± 0.0075
PnP-ADMM [28]	**34.216 ± 0.597**	**0.8938 ± 0.0077**