Nonasymptotic Convergence Rates for Plug-and-Play Methods With MMSE Denoisers

Henry Pritchard and Rahul Parhi, *Member, IEEE*

***Abstract*—It is known that the minimum-mean-squared-error (MMSE) denoiser under Gaussian noise can be written as a proximal operator, which suffices for asymptotic convergence of plug-and-play (PnP) methods but does not reveal the structure of the induced regularizer or give convergence rates. We show that the MMSE denoiser corresponds to a regularizer that can be written explicitly as an upper Moreau envelope of the negative log-marginal density, which in turn implies that the regularizer is 1-weakly convex on its domain. Using this property, we derive (to the best of our knowledge) the first sublinear convergence guarantee for PnP proximal gradient descent with an MMSE denoiser. We validate the theory with a one-dimensional synthetic study that recovers the implicit regularizer. We further validate the theory with imaging experiments (deblurring and computed tomography), which exhibit the predicted sublinear behavior.**

***Index Terms*—Moreau envelopes, MMSE denoisers, plug-and-play methods, proximal gradient descent, weak convexity, image reconstruction.**

I. Introduction

In this work, we focus on the setting of *linear inverse problems*, where $\overline{x} \in \mathbb{R}^n$ is the true signal, $y \in \mathbb{R}^m$ is the observed signal obtained by

$$y = A\overline{x} + \epsilon, \tag{1}$$

for linear *forward operator* $A \in \mathbb{R}^{m\times n}$ and measurement noise $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$. The task of recovering $\overline{x}$ from y is extensively studied in statistics and signal processing. A special case is the denoising problem, obtained when $A = I$:

$$z = \overline{x} + \epsilon. \tag{2}$$

Although simpler, denoising is directly useful in solving general inverse problems, often as an intermediate step.

Recovering $\overline{x}$ from y is often posed as the minimization of a regularized objective function

$$F(x) \coloneqq f(x) + g(x), \tag{3}$$

where f is a data-fidelity term (typically $f(x) \coloneqq \frac{1}{2}\|Ax - y\|^2$ under Gaussian noise), and g is a regularization term that encodes prior beliefs about the *class of admissible solutions*. For instance, the ℓ^1-norm ($g(x) \propto \|x\|_1$) promotes sparsity and underlies the theory of compressed sensing [9], [10], [18]. Similarly, total variation (TV), defined as $g(x) \propto \mathrm{TV}(x) \coloneqq \sum_i |x_i - x_{i-1}|$ for 1-D signals, favors piecewise-constant solutions and is widely used in image denoising [11], [12], [38], [46], [52]. Without regularization, the problem is often *ill-posed:* Solutions may be non-unique, unstable, or nonexistent. Since the data-fidelity term is typically well-behaved, the regularizer largely determines the structure of solutions. More broadly, the regularizer largely dictates how optimization algorithms behave, so understanding its structure is crucial.

A. Proximal Gradient Descent

Gradient descent is a classical method for minimizing smooth (Lipschitz gradient) objective functions, with iterations

$$x_{k+1} = x_k - \gamma \nabla F(x_k). \tag{4}$$

With appropriately chosen step size $\gamma > 0$, the iterates x_k converge to a stationary point [34]. However, gradient descent is not well-suited for nonsmooth objectives.

Proximal gradient descent (PGD) addresses this limitation by splitting the objective function as in (3). The data-fidelity term f is typically smooth while the regularizer g is often not (such as ℓ^1 or TV). At each iteration, PGD takes a gradient step on the data-fidelity term followed by a *proximal step* for the regularizer, yielding iterations of the form

$$x_{k+1} = \mathrm{prox}_{\gamma g}\left(x_k - \gamma \nabla f(x_k)\right). \tag{5}$$

Here $\mathrm{prox}_{\gamma g}$ denotes the proximal operator of g (see Definition 4). Convergence of PGD has been widely studied in both convex and nonconvex settings and can be established under various assumptions on f and g with appropriately chosen step size $\gamma > 0$ [36, Table 1].

B. Plug-and-Play Methods

The *plug-and-play* (PnP) framework [58] extends PGD by replacing the proximal step $\mathrm{prox}_{\gamma g}$ with a generic denoiser D_σ. This substitution is motivated by the interpretation of $\mathrm{prox}_{\gamma g}$ as the maximum a posteriori (MAP) solution to the denoising problem (2) under prior $p_X \propto \mathrm{e}^{-\gamma g}$ [15, Eq. (10.13)], the exact problem denoisers are designed to solve. The PnP-PGD iteration

Received 7 November 2025; revised 24 March 2026 and 15 May 2026; accepted 14 June 2026. Date of publication 24 June 2026; date of current version 10 July 2026. The work of Rahul Parhi was supported by a Hellman Fellowship from the University of California. This work was supported by NVIDIA Corporation through the NVIDIA Academic Grant Program, including the donation of an RTX PRO 6000 Blackwell Max-Q Workstation Edition GPU used in this research. The associate editor coordinating the review of this article and approving it for publication was Prof. Andreas Hauptmann. *(Corresponding author: Rahul Parhi.)*

The authors are with the Department of Electrical and Computer Engineering, University of California, San Diego, CA 92093 USA (e-mail: hepritchard@ucsd.edu; rahul@ucsd.edu).

Digital Object Identifier 10.1109/TCI.2026.3707068

TABLE I
SUMMARY OF RELATED CONVERGENCE GUARANTEES FOR PLUG-AND-PLAY PROXIMAL GRADIENT DESCENT (PNP-PGD). NOTE L_f IS THE LIPSCHITZ CONSTANT FOR ∇f, L_{g_σ} IS THE LIPSCHITZ CONSTANT FOR ∇g_σ, γ IS THE STEP SIZE, AND K IS THE ITERATION COUNT.

Work	**Assumptions on data-fidelity term f**	**Assumptions on denoiser D_σ**	**Convergence rate**
Classical PGD [7]	Convex, L_f-smooth	$D_\sigma = \text{prox}_g$, where g is convex	$\mathcal{O}(1/\sqrt{K})$
Lipschitz condition on the denoiser [53]	L_f-smooth, μ-strongly convex	$D_\sigma - I$ is ε-Lipschitz for $\varepsilon < \frac{2\mu}{L_f - \mu}$	Asymptotic
Gradient-step denoisers [32]	L_f-smooth, $\gamma L_f < \frac{L_{g_\sigma}+2}{L_{g_\sigma}+1}$	$D_\sigma = I - \nabla g_\sigma$, where $L_{g_\sigma} < 1$	$\mathcal{O}(1/\sqrt{K})$
Prior work with MMSE denoisers [59]	L_f-smooth, $L_f < 1$	MMSE denoiser	Asymptotic
This paper (Corollary 2)	L_f-smooth, $L_f < 1$	MMSE denoiser	$\mathcal{O}(1/\sqrt{K})$

is given by

$$x_{k+1} = D_\sigma\left(x_k - \gamma \nabla f(x_k)\right), \tag{6}$$

and has been shown to yield state-of-the-art results when combined with powerful image denoisers such as BM3D [17] or a deep neural network (DNN).

A central challenge in the PnP framework is that while a proximal operator is always tied to a well-defined regularizer, many of the most effective denoising methods are not. This substitution therefore breaks the direct connection to an explicit optimization problem, making theoretical interpretations and convergence claims difficult. Much of the literature imposes strict assumptions on the denoiser to obtain convergence guarantees. Many of the most effective denoisers do not satisfy these assumptions, yet still achieve excellent results [2].

C. Related Works

When f and g are convex with L-smooth f, it is well known that PGD converges to a fixed point for appropriate step size γ [47, Section 4.2]. This result can be proven via the machinery of monotone operator theory, the goal being to show that the composite PGD step (5) constitutes an averaged operator, guaranteeing convergence [6], [14], [16].[1]

In particular, convergence of PnP-PGD has been shown under tight assumptions on the denoiser, such as nonexpansiveness [56] or suitable Lipschitz (e.g., residual-Lipschitz) bounds [53]. Many works encourage these properties via, e.g., spectral normalization [53] or Björck orthonormalization [3], [37]. In [49], the authors propose a technique for learning a maximally monotone operator. Convergence of PnP has also been shown for specific denoisers, such as kernel denoisers [4], [24], gradient-step (GS) denoisers [32], [33], bounded denoisers [13], and linear denoisers [23], [43].

A similar line of work learns regularizers explicitly [19], [20], [25], [26], [30], [44], [45], [50], [54], rather than implicitly through the denoiser. This approach benefits from direct interpretability of the regularizer and can inherit convergence guarantees from classical variational theory.

A complementary approach leverages the intrinsic *statistical* structure of the denoiser. A particularly notable example—and the subject of this work—is the MMSE denoiser under Gaussian noise. With the Tweedie/Stein identities [21], [55], the MMSE denoiser can be tied to the score function, endowing it with useful structural properties [41]. Further, [27] and [29] showed that the MMSE denoiser is C^∞ and can be written as the proximal map of a regularizer (which is C^∞) under various forward and noise models, and derived an explicit formula for this regularizer (23). This progress enabled the derivation of asymptotic convergence for both the ADMM [48] and PGD [59] variants of PnP using an MMSE denoiser. However, a gap in the theory remains: The explicit formula offers little insight into the behavior of the induced regularizer, and there are no *nonasymptotic* convergence rates tailored to MMSE denoisers. Table I highlights several closely related results on PnP-PGD, the assumptions they impose on the denoiser, and the convergence guarantees they establish.

D. Contributions

We make the following contributions.

- *Characterization of the implicit regularizer of the MMSE denoiser:* We show in Theorem 4 that the MMSE denoiser is the proximal map of a regularizer that admits an explicit form as an *upper Moreau envelope*, i.e., $\phi_{\text{MMSE}}(x)$ is, up to a constant,

$$\sigma^2 \sup_z \left\{ f_Z(z) - \tfrac{1}{2\sigma^2}\|z - x\|^2 \right\}, \tag{7}$$

  on its domain, where $f_Z = -\log p_Z$ is the negative log-marginal density from (2). This is a novel explicit form of the implicit regularizer in the MMSE denoiser.
- *Nonasymptotic convergence of MMSE-PGD via weak convexity:* In Corollary 1, we prove that ϕ_{MMSE} is 1-weakly convex on its domain. With this property, we establish (to the best of our knowledge) the first *nonasymptotic* convergence guarantees for PnP-PGD with an MMSE denoiser in Theorem 5 and Corollary 2. In particular, we show that, under an L-smooth data-fidelity term f with $L < 1$, both $\|\nabla F(x_k)\|$ and $\|x_{k+1} - x_k\|$ decrease sublinearly in terms of the best iterate.
- *Experiments:* In Section IV, we consider the following experimental setups: (i) a one-dimensional synthetic example validating the upper Moreau envelope form of the implicit regularizer for the MMSE denoiser, (ii) Gaussian deblurring on the MNIST dataset [35], and (iii) computed tomography on the MayoCT dataset [40]. In both imaging tasks we observe the predicted sublinear convergence.

[1] When g is proper, lower-semicontinuous, convex, ∂g is maximally monotone, making $\text{prox}_{\gamma g} = (I + \gamma \partial g)^{-1}$ firmly-nonexpansive, i.e., $1/2$-averaged. The gradient step is also averaged for L-smooth f, making their composition an averaged operator.

II. Preliminaries

In this section, we collect some definitions and results used in the remainder of the paper.

Definition 1 (Lower Moreau Envelope): Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function with $\gamma > 0$. The *lower Moreau envelope* of f is given by

$$M_\gamma f(x) \coloneqq \inf_y \left\{ f(y) + \tfrac{1}{2\gamma}\|y - x\|^2 \right\}. \tag{8}$$

Definition 2 (Upper Moreau Envelope): Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function with $\gamma > 0$. The *upper Moreau envelope* of f is given by

$$M^\gamma f(x) \coloneqq \sup_y \left\{ f(y) - \tfrac{1}{2\gamma}\|y - x\|^2 \right\}. \tag{9}$$

Note that the lower Moreau envelope is the usual Moreau envelope, but we make the distinction in this paper since we use both lower and upper Moreau envelopes. For the infimum defining the lower Moreau envelope to be achieved, f must be lower-semicontinuous. An important property of the lower Moreau envelope $M_\gamma f$ is that it shares global minimizers with f whenever f is lower-semicontinuous, i.e. $\operatorname{argmin}_x f(x) = \operatorname{argmin}_x M_\gamma f(x)$. See [8] for further background.

Definition 3 (Weak convexity): Let $f : \mathbb{R}^n \to \mathbb{R}$, $\gamma > 0$. f is called γ-*weakly convex* if the mapping $x \mapsto f(x) + \frac{\gamma}{2}\|x\|^2$ is convex.

Crucial to our analysis is the following lemma, adapted from [8, Lemma 3].

Lemma 1: Let $f : \mathbb{R}^n \to (-\infty, \infty]$. Then, $M^\gamma M_\gamma f(x) \le f(x)$ for all $x \in \mathbb{R}^n$ and $M^\gamma f$ is $\frac{1}{\gamma}$-weakly convex.

Proof: By definition,

$$\begin{aligned} &M^\gamma M_\gamma f(x) \\ &= \sup_y \inf_z \left\{ f(z) + \tfrac{1}{2\gamma}\|z - y\|^2 - \tfrac{1}{2\gamma}\|y - x\|^2 \right\} \\ &\le \sup_y f(x) = f(x). \end{aligned} \tag{10}$$

Next,

$$\begin{aligned} &M^\gamma f(x) + \tfrac{1}{2\gamma}\|x\|^2 \\ &= \sup_y \left\{ f(y) - \tfrac{1}{2\gamma}\|x - y\|^2 + \tfrac{1}{2\gamma}\|x\|^2 \right\} \\ &= \sup_y \left\{ f(y) - \tfrac{1}{2\gamma}\|y\|^2 + \tfrac{1}{\gamma}\langle x, y\rangle \right\}, \end{aligned} \tag{11}$$

which is a supremum of affine functions, and hence convex. Therefore, $M^\gamma f$ is $\frac{1}{\gamma}$-weakly convex. □

We now turn to the proximal operator, a central tool in optimization. We shall rely on the fact that MMSE denoisers themselves admit a proximal representation.

Definition 4 (Proximal operator): Let $\gamma > 0$, and $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be proper and lower-semicontinuous. The *proximal operator* of f with parameter γ is the mapping

$$\operatorname{prox}_{\gamma f}(x) \coloneqq \operatorname*{argmin}_{z \in \mathbb{R}^n} \left\{ f(z) + \tfrac{1}{2\gamma}\|z - x\|^2 \right\}. \tag{12}$$

The proximal operator is generally set-valued and may be empty unless additional assumptions are imposed. For a proper, lower-semicontinuous function $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$, the infimum defining the lower Moreau envelope is attained at every minimizer of the proximal problem; that is,

$$M_\gamma f(x) = f(z) + \tfrac{1}{2\gamma}\|x - z\|^2, \quad z \in \operatorname{prox}_{\gamma f}(x), \tag{13}$$

for all $x \in \mathbb{R}^n$. When f is a convex, lower-semicontinuous, and proper function, it is well known that $\operatorname{prox}_{\gamma f}$ and $M_\gamma f$ are also related by the following formula due to Moreau [42]:

$$\nabla M_\gamma f(x) = \tfrac{1}{\gamma}(x - \operatorname{prox}_{\gamma f}(x)). \tag{14}$$

Although the proximal operator of a nonconvex function is generally set-valued, when $\operatorname{prox}_{\gamma f}$ is an MMSE denoiser, it is always single-valued and continuous, which is the setting of our main results. This motivates the following extension of the Moreau gradient identity.

Theorem 1 (Extension of the Moreau Gradient Identity): Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be proper and lower-semicontinuous. If $\operatorname{prox}_{\gamma f}(x)$ is single-valued and continuous for all x, then the gradient of the lower Moreau envelope exists and takes the form

$$\nabla M_\gamma f(x) = \tfrac{1}{\gamma}(x - \operatorname{prox}_{\gamma f}(x)), \quad x \in \mathbb{R}^n. \tag{15}$$

Proof: Let $h(x, z) \coloneqq f(z) + \frac{1}{2\gamma}\|z - x\|^2$. We have for all $z \in \mathbb{R}^n$, $t \in \mathbb{R}$,

$$h(x + tu, z) - h(x, z) = \tfrac{t}{\gamma}\langle x - z, u\rangle + \tfrac{t^2}{2\gamma}\|u\|^2. \tag{16}$$

From the definition, $M_\gamma f(x) = \min_z h(x, z)$. Fix some $x_0 \in \mathbb{R}^n$ where there is a unique minimizer $z_0 = \operatorname{prox}_{\gamma f}(x_0)$ by assumption. Take some $u \in \mathbb{R}^n$. For $t > 0$, there is some $z_t = \operatorname{prox}_{\gamma f}(x_0 + tu)$ so that

$$\begin{aligned} &M_\gamma f(x_0 + tu) \\ &= h(x_0 + tu, z_t) \\ &= h(x_0, z_t) + \tfrac{t}{\gamma}\langle x_0 - z_t, u\rangle + \tfrac{t^2}{2\gamma}\|u\|^2 \\ &\ge M_\gamma f(x_0) + \tfrac{t}{\gamma}\langle x_0 - z_t, u\rangle + \tfrac{t^2}{2\gamma}\|u\|^2. \end{aligned} \tag{17}$$

Subtracting and dividing through by t, we get

$$\begin{aligned} &\liminf_{t \downarrow 0} \tfrac{M_\gamma f(x_0 + tu) - M_\gamma f(x_0)}{t} \\ &\ge \liminf_{t \downarrow 0} \left(\tfrac{1}{\gamma}\langle x_0 - z_t, u\rangle + \tfrac{t}{2\gamma}\|u\|^2 \right) \\ &= \tfrac{1}{\gamma}\langle x_0 - z_0, u\rangle, \end{aligned} \tag{18}$$

where the last equality is by the assumption that $\operatorname{prox}_{\gamma f}$ is continuous. Next,

$$\begin{aligned} &M_\gamma f(x_0 + tu) \le h(x_0 + tu, z_0) \\ &= M_\gamma f(x_0) + \tfrac{t}{\gamma}\langle x_0 - z_0, u\rangle + \tfrac{t^2}{2\gamma}\|u\|^2. \end{aligned} \tag{19}$$

Again, subtracting and dividing through by t,

$$\begin{aligned} &\limsup_{t \downarrow 0} \tfrac{M_\gamma f(x_0 + tu) - M_\gamma f(x_0)}{t} \\ &\le \limsup_{t \downarrow 0} \left(\tfrac{1}{\gamma}\langle x_0 - z_0, u\rangle + \tfrac{t}{2\gamma}\|u\|^2 \right) \\ &= \tfrac{1}{\gamma}\langle x_0 - z_0, u\rangle. \end{aligned} \tag{20}$$

Combining (18) and (20), we have that the directional derivative of $M_\gamma f$ at x_0 in direction u satisfies $(M_\gamma f)'(x_0, u) = \frac{1}{\gamma}\langle x_0 - z_0, u\rangle$. Linearity in u gives us a two-sided directional derivative. Therefore, the gradient is

$$\nabla M_\gamma f(x_0) = \tfrac{1}{\gamma}(x_0 - z_0) = \tfrac{1}{\gamma}\left(x_0 - \operatorname{prox}_{\gamma f}(x_0)\right), \quad (21)$$

which completes the proof. □

Many similar versions of this identity are well known in the literature, but are often formulated in terms of *prox-regularity*. We included a simple self-contained proof for completeness. For more detail, see [51, Theorem 13.37].

We next review key results on MMSE denoisers and state their connection to proximal mappings, beginning with Tweedie's formula.

Theorem 2 (Tweedie's Formula): Let $X \sim p_X$, $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$ be independent random variables in $\mathbb{R}^n$, and let $Z = X + \epsilon$, $\psi_\sigma(z) \coloneqq \mathbb{E}[X|Z = z]$. Then,

$$\psi_\sigma(z) = z + \sigma^2 \nabla \log p_Z(z), \quad (22)$$

where $p_Z = p_X * \mathcal{N}(0, \sigma^2 I)$.

For a thorough introduction to Tweedie's formula, we refer the reader to [21]. Note that Tweedie's formula makes no assumptions on the prior distribution p_X, which may even include discrete atoms. We recall the following result on MMSE denoisers due to Gribonval et al. [27], [29].

Theorem 3 (Main result from [29] and [27, Corollary 1]): Let $X \sim p_X$ and $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$ be independent random variables in $\mathbb{R}^n$, and let $Z = X + \epsilon$. Assume p_X is non-degenerate.[2] The MMSE denoiser under (2) given by $\psi_\sigma(z) \coloneqq \mathbb{E}[X|Z = z]$ has the following properties:

1) It is one-to-one from $\mathbb{R}^n$ onto its image.
2) It is C^∞, and so is its inverse.
3) It can be written as $\psi_\sigma(z) = \operatorname{prox}_{\phi_{\text{MMSE}}}(z)$, where

$$\phi_{\text{MMSE}}(z) \coloneqq \begin{cases} -\frac{1}{2}\|\psi_\sigma^{-1}(z) - z\|^2 + \sigma^2 f_Z\left(\psi_\sigma^{-1}(z)\right), & z \in \psi_\sigma(\mathbb{R}^n), \\ \infty, & \text{otherwise}, \end{cases} \quad (23)$$

where $f_Z = -\log p_Z$.

4) ϕ_{MMSE} is C^∞ on $\psi_\sigma(\mathbb{R}^n)$.

III. Main Results

In this section, we present our theoretical contributions. We begin by introducing a new structural representation of the implicit regularizer underlying the MMSE denoiser. This characterization serves as the foundation for deriving novel nonasymptotic convergence guarantees.

A. New Characterization of the MMSE Regularizer

It is known (Theorem 3) that the MMSE denoiser is the proximal map of an infinitely differentiable penalty. Yet beyond this, the structure of the associated penalty ϕ_{MMSE} remains unclear. The only known explicit expression of ϕ_{MMSE} (23) is cumbersome and offers little analytical insight. In this section, we show that ϕ_{MMSE} admits an upper Moreau envelope representation in terms of f_Z, yielding useful insights, namely 1-weak convexity, which we will leverage to prove stronger convergence results for PGD using an MMSE denoiser.

For the following, we will assume the setting of the denoising problem (2), letting $X \sim p_X$, $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$ be independent random variables in $\mathbb{R}^n$, $Z = X + \epsilon$, so that $p_Z = p_X * \mathcal{N}(0, \sigma^2 I)$. We will let $f_Z \coloneqq -\log p_Z$.

Lemma 2: The negative log marginal distribution from the denoising problem (2) can be written as

$$f_Z = \tfrac{1}{\sigma^2} M_1 \phi_{\text{MMSE}} + C_{x_0}, \quad (24)$$

where ϕ_{MMSE} is as in (23) and C_{x_0} takes the form

$$C_{x_0} \coloneqq f_Z(x_0) - \tfrac{1}{\sigma^2} M_1 \phi_{\text{MMSE}}(x_0), \quad (25)$$

for any choice of $x_0 \in \mathbb{R}^n$.

Proof: Pick $x \in \mathbb{R}^n$. By Theorem 2 and Theorem 3,

$$\begin{aligned} x - \operatorname{prox}_{\phi_{\text{MMSE}}}(x) &= -\sigma^2 \nabla \log p_Z(x) \\ &= \sigma^2 \nabla f_Z(x). \end{aligned} \quad (26)$$

Combined with Theorem 1 yields

$$\sigma^2 \nabla f_Z(x) = \nabla M_1 \phi_{\text{MMSE}}(x). \quad (27)$$

Integrating gives the result. □

Theorem 4: The implicit regularizer in the MMSE denoiser can be written as

$$\phi_{\text{MMSE}}(x) = \begin{cases} \sigma^2 M^{\sigma^2} f_Z(x) - \sigma^2 C_{x_0}, & x \in \psi_\sigma(\mathbb{R}^n) \\ +\infty, & \text{otherwise}, \end{cases} \quad (28)$$

where C_{x_0} is defined in (25).

Proof: Pick some $x \in \mathbb{R}^n$. By Lemma 1,

$$\phi_{\text{MMSE}}(x) \geq M^1 M_1 \phi_{\text{MMSE}}(x), \quad (29)$$

which, by the previous result, we can rewrite as

$$\begin{aligned} \phi_{\text{MMSE}}(x) &\geq M^1(\sigma^2 f_Z - \sigma^2 C_{x_0})(x) \\ &= \sigma^2 M^{\sigma^2} f_Z(x) - \sigma^2 C_{x_0}. \end{aligned} \quad (30)$$

For the reverse inequality, take some $x \in \psi_\sigma(\mathbb{R}^n)$. There exists some $z \in \mathbb{R}^n$ so that $x = \psi_\sigma(z) = \operatorname{prox}_{\phi_{\text{MMSE}}}(z)$. The infimum defining $M_1 \phi_{\text{MMSE}}$ is achieved at x (cf. (13)), so that

$$\begin{aligned} \phi_{\text{MMSE}}(x) + \tfrac{1}{2}\|x - z\|^2 &= M_1 \phi_{\text{MMSE}}(z) \\ &= \sigma^2 f_Z(z) - \sigma^2 C_{x_0}, \end{aligned} \quad (31)$$

where the second line holds by Lemma 2. Rearranging, we have

$$\begin{aligned} \phi_{\text{MMSE}}(x) &= \sigma^2 \left(f_Z(z) - \tfrac{1}{2\sigma^2}\|x - z\|^2\right) - \sigma^2 C_{x_0} \\ &\leq \sigma^2 M^{\sigma^2} f_Z(x) - \sigma^2 C_{x_0}, \end{aligned} \quad (32)$$

which completes the proof. □

We illustrate this result in Figs. 1 and 2 under a variety of mixture-of-Gaussian and mixture-of-Laplacian priors. An

[2] i.e., there is no pair $v \in \mathbb{R}^n$, $c \in \mathbb{R}$ such that $\langle X, v\rangle = c$, almost surely.

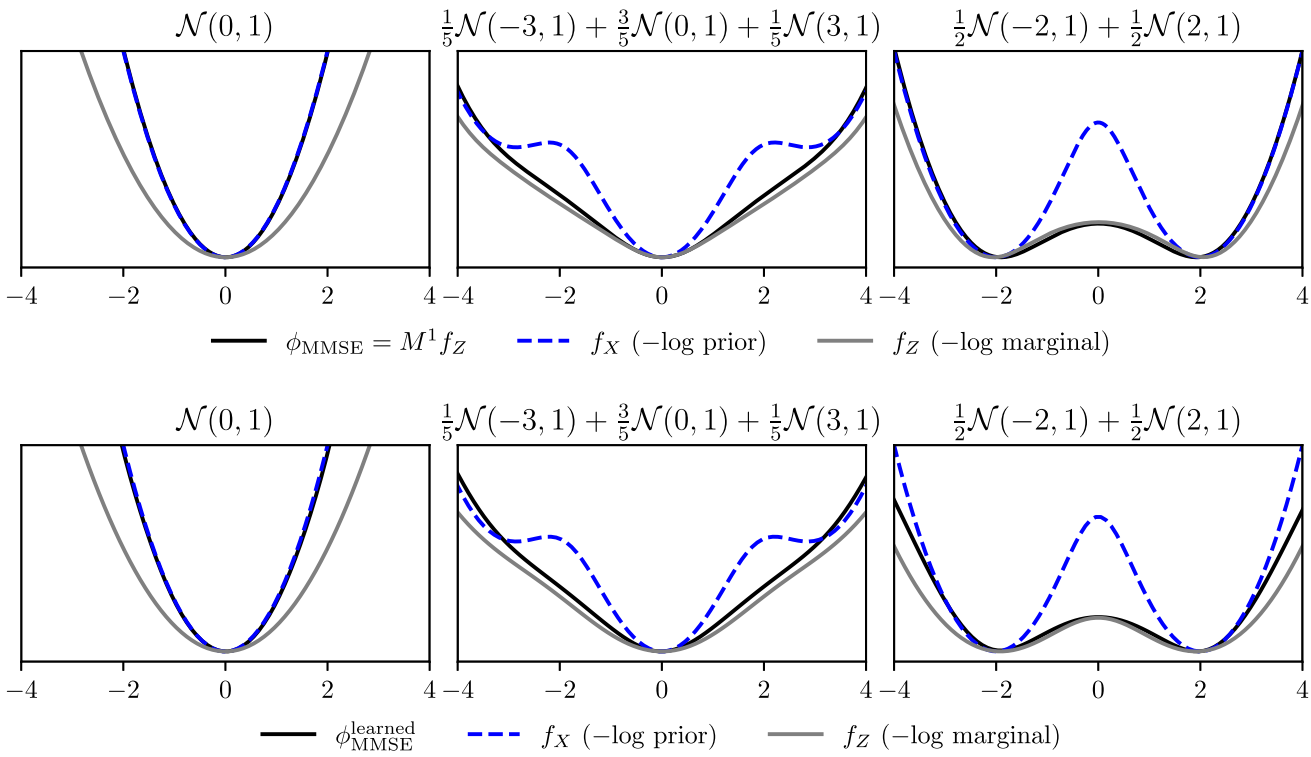

Fig. 1. Top: Calculated implicit regularizer in the MMSE denoiser (ϕ_{MMSE}), negative log-marginal f_Z, and negative log-prior f_X for three mixture-of-Gaussian priors under unit Gaussian noise. Bottom: The learned regularizer $\phi_{\text{MMSE}}^{\text{learned}}$, along with the same reference curves. Details in Section IV-A.

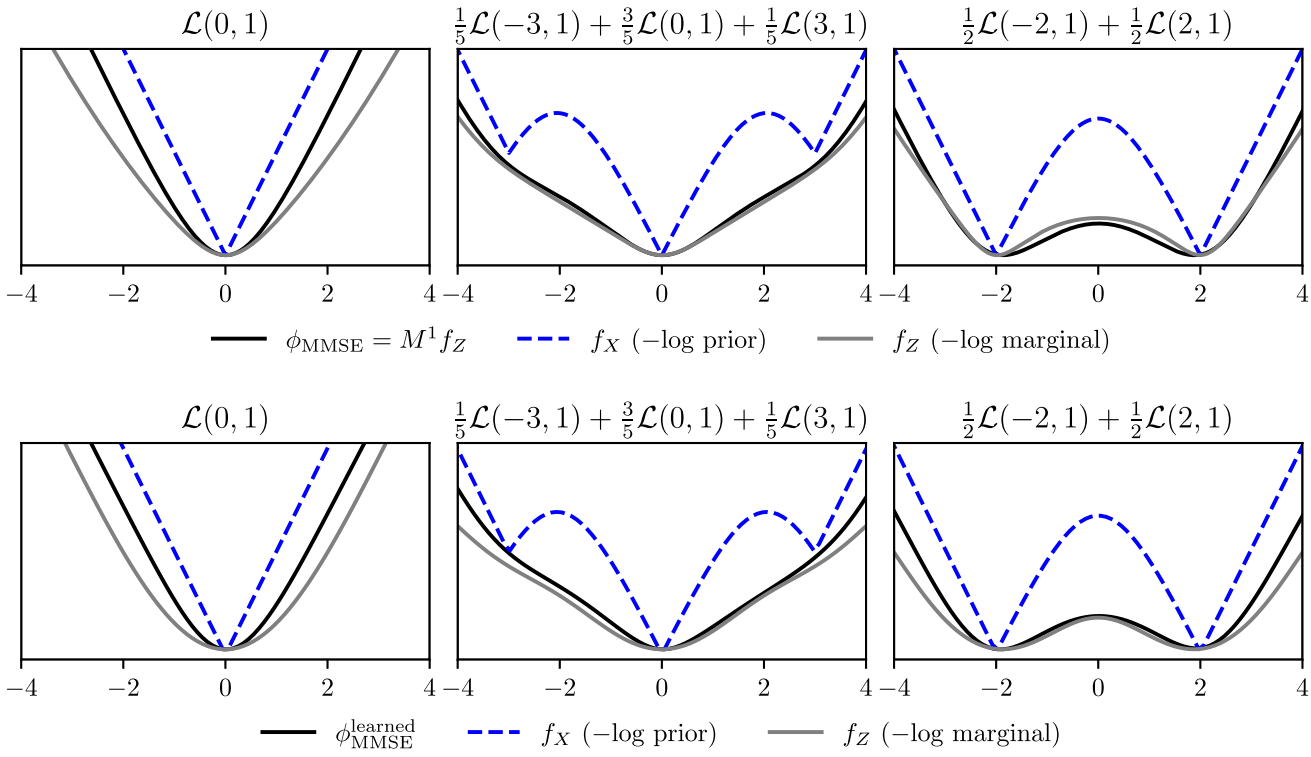

Fig. 2. Top: Calculated implicit regularizer in the MMSE denoiser (ϕ_{MMSE}), negative log-marginal f_Z, and negative log-prior f_X for three mixture-of-Laplacian priors under unit Gaussian noise. Bottom: The learned regularizer $\phi_{\text{MMSE}}^{\text{learned}}$, along with the same reference curves. Details in Section IV-A.

overview of the experimental details is provided in Section IV-A. We conclude with an immediate corollary.

Corollary 1: The implicit regularizer ϕ_{MMSE} is 1-weakly convex on its domain $\text{dom}(\phi_{\text{MMSE}}) = \psi_\sigma(\mathbb{R}^n)$.

Proof: By Theorem 4, ϕ_{MMSE} equals $\sigma^2 M^{\sigma^2} f_Z$ on $\psi_\sigma(\mathbb{R}^n)$. By Lemma 1, $M^{\sigma^2} f_Z$ is $\frac{1}{\sigma^2}$-weakly convex. Hence, $\sigma^2 M^{\sigma^2} f_Z$ is 1-weakly convex. The result follows. □

B. Proximal Gradient Descent Convergence Results

Consider the iteration

$$x_{k+1} = \psi_\sigma(x_k - \nabla f(x_k)), \tag{33}$$

where ψ_σ is the MMSE denoiser under Gaussian noise of variance σ^2 and f is an L-smooth data-fidelity term. The step size is fixed because it is implicit in the MMSE denoiser. The associated objective for these iterates is

$$F(x) \coloneqq f(x) + \phi_{\text{MMSE}}(x), \tag{34}$$

where ϕ_{MMSE} is the implicit regularizer in the MMSE estimator, discussed in Section III-A.

When the objective F is smooth, the map $x \mapsto \nabla F(x)$ is continuous. Therefore, measuring convergence via the gradient norm $\|\nabla F(x_k)\|$ is natural as small gradient norms imply closeness to a stationary point. For nonconvex problems, $\|\nabla F(x_k)\|$ may oscillate even as it trends downward, so nonasymptotic convergence is typically stated in terms of the best iterate up until the current iteration [5]. Since f is smooth and ϕ_{MMSE} is C^∞ on its domain (Theorem 3), the objective F is C^∞ on its domain. We measure convergence via $\min_{1\le i\le K}\|\nabla F(x_i)\|$ at each iteration K. We now state our main convergence result. As we shall see, once weak convexity is established, sublinear convergence follows immediately by specializing existing theory to our context.

Theorem 5 (Convergence of (33)): Let f be proper, bounded from below and L-smooth with $L < 1$. Then for F in (34) and x_k as in (33), the following hold:

1) The sequence $(F(x_k))$ is non-increasing and converges.
2) The sequence has finite squared length, i.e., $\sum_{k=0}^{\infty} \|x_{k+1} - x_k\|^2 < \infty$, and the best-step residual satisfies

$$\min_{1\le i\le K}\|x_{i+1} - x_i\| = \mathcal{O}\left(\tfrac{1}{\sqrt{K}}\right). \tag{35}$$

3) All cluster points of x_k are stationary points of F.

Proof: We have that $\sigma^2 M^{\sigma^2} f_Z - \sigma^2 C_{x_0}$ is a globally-defined 1-weakly convex function that agrees with ϕ_{MMSE} on $\psi_\sigma(\mathbb{R}^n)$ by Theorem 4 and Lemma 1. Further,

$$\begin{aligned}\sigma^2 M^{\sigma^2} f_Z - \sigma^2 C_{x_0} &\ge \sigma^2 f_Z - \sigma^2 C_{x_0} \\ &\ge \sigma^2 \tfrac{n}{2}\log(2\pi\sigma^2) - \sigma^2 C_{x_0},\end{aligned} \tag{36}$$

so ϕ_{MMSE} is bounded below. Next, each iterate of (33) clearly satisfies $x_k \in \psi_\sigma(\mathbb{R}^n)$ for $k \ge 1$. Define

$$\widetilde{\phi} = \sigma^2 M^{\sigma^2} f_Z - \sigma^2 C_{x_0}. \tag{37}$$

By Theorem 4 and Lemma 1, $\widetilde{\phi}$ is globally 1-weakly convex and agrees with ϕ_{MMSE} on $\psi_\sigma(\mathbb{R}^n)$ and hence on the iterates x_k for $k \ge 1$. Applying the descent argument from [31, Theorem 1] with $\tau = 1$, $M = 1$, and $L_f < 1$ yields the result.

Corollary 2: Under the assumptions of Theorem 5,

$$\min_{1\le i\le K}\|\nabla F(x_i)\| = \mathcal{O}\left(\tfrac{1}{\sqrt{K}}\right). \tag{38}$$

Proof: From Theorem 3 and Theorem 4,

$$x_{k+1} = \operatorname*{argmin}_z G_k(z), \tag{39}$$

where

$$\begin{aligned}G_k(z) \coloneqq{}& \sigma^2(M^{\sigma^2} f_Z(z) - C_{x_0}) \\ &+ \tfrac{1}{2}\|z - (x_k - \nabla f(x_k))\|^2.\end{aligned} \tag{40}$$

By Lemma 1, $\sigma^2 M^{\sigma^2} f_Z$ is 1-weakly convex. Hence, G_k is convex and we can conclude that its minimizer satisfies first-order optimality, i.e., $\nabla G_k(x_{k+1}) = 0$. We can write this as $\nabla\phi_{\text{MMSE}}(x_{k+1}) + x_{k+1} - (x_k - \nabla f(x_k)) = 0$, because $\phi_{\text{MMSE}}(z)$ and $\sigma^2 M^{\sigma^2} f_Z(z) - \sigma^2 C_{x_0}$ agree on the open set $\psi_\sigma(\mathbb{R}^n)$, and $x_{k+1} \in \psi_\sigma(\mathbb{R}^n)$. Therefore,

$$\nabla F(x_{k+1}) = x_k - x_{k+1} + \nabla f(x_{k+1}) - \nabla f(x_k). \tag{41}$$

From this, we have that

$$\begin{aligned}\|\nabla F(x_{k+1})\| &\le \|x_k - x_{k+1}\| + \|\nabla f(x_{k+1}) - \nabla f(x_k)\| \\ &\le (1+L)\|x_{k+1} - x_k\| < 2\|x_{k+1} - x_k\|.\end{aligned} \tag{42}$$

The result then follows from Theorem 5(2). □

Remark 1: The requirement $L < 1$ may seem restrictive. However, any L-smooth data-fidelity term can be scaled appropriately to meet this condition. For example, in our computed tomography experiment (Section IV-D), the forward operator is a discretized Radon transform, for which f is L-smooth, $L \approx 140$. We simply scale the data-fidelity term to obtain empirical results that verify our theoretical convergence rate.

Remark 2: The *asymptotic* convergence of PGD using an MMSE denoiser has previously been established for a range of step sizes $\gamma > 0$. In particular, [59] showed that PnP-MMSE converges for any $\gamma \le 1/L$. Though the MMSE denoiser can be written as $\text{prox}_{\phi_{\text{MMSE}}}$, we note that $\text{prox}_{\gamma\phi_{\text{MMSE}}}$ cannot be evaluated for any $\gamma \ne 1$. Consequently, any change in the step size can be realized by a rescaling of the data-fidelity term. For this reason, we assume an implicit step size of 1 and scale the Lipschitz constant of the data-fidelity term accordingly.

IV. Experiments

In this section, we numerically verify the theoretical results presented in Section III. The code to reproduce our experiments is publicly available on GitHub.[3] All training and testing were performed on an NVIDIA RTX PRO 6000 Blackwell Max-Q Workstation Edition GPU using Ubuntu 24.04.

Central to our experiments is the task of estimating the MMSE denoiser of a given distribution under Gaussian noise. Following the work of [22], we choose to model the MMSE

[3] https://github.com/sparsity-group/pnpmmse

TABLE II
DNN ARCHITECTURAL DETAILS FOR INVERSE PROBLEMS EXPERIMENTS (SECTION IV-C AND SECTION IV-D). H DENOTES THE HIDDEN DIMENSION AND L DENOTES THE NUMBER OF LAYERS.

	Gaussian Deblurring	**Computed Tomography**
Input size	28×28	512×512
Dimensions	$H \times 28 \times 28,\ H \times 14 \times 14,\ H \times 14 \times 14,$ $H \times 7 \times 7,\ H \times 7^2 \times 1 \times 1,\ 64 \times 1 \times 1$	$H \times 512 \times 512,\ H \times 256 \times 256,\ H \times 256 \times 256,\ H \times 128 \times 128$ $H \times 128 \times 128,\ H \times 64 \times 64,\ H \times 64 \times 64,\ H \times 32 \times 32,$ $H \times 16 \times 16,\ 64 \times 1 \times 1,\ 64 \times 1 \times 1$
L	6	11
H	64	256
Direct connections (W_k)	4 Conv2d (3×3 kernel), 2 Linear	9 Conv2d (3×3 kernel), 1 Conv2d (16×16 kernel), 1 Linear
Skip connections (A_k)	3 Conv2d, 1 Linear	8 Conv2d (3×3 kernel), 1 Conv2d (16×16 kernel)

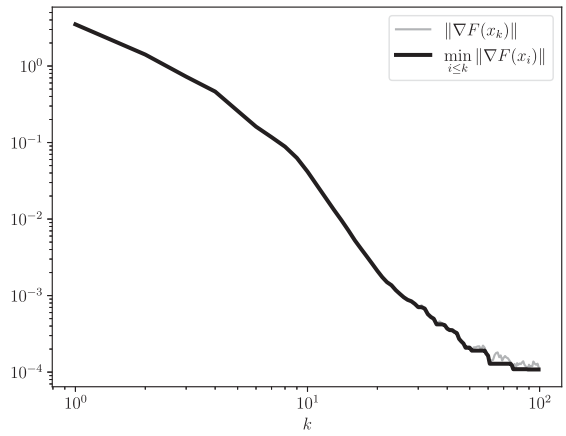

Fig. 3. Best-iterate gradient norm over 100 iterations of PnP-PGD with an MMSE denoiser for the MNIST Gaussian deblurring experiment in Section IV-C, plotted on a log-log scale. The quantity decays sublinearly, consistent with a $\mathcal{O}(1/\sqrt{K})$ rate.

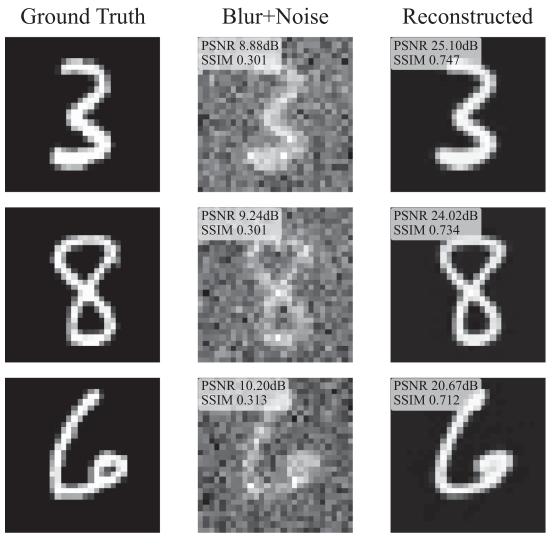

Fig. 4. PnP-PGD reconstruction results on three selections from the MNIST dataset under Gaussian blur (Section IV-C). From left to right: clean reference, blurred and noisy input, and reconstruction after 100 iterations. PSNR gains are roughly 10–16 dB across all three examples.

denoiser as the gradient of a softplus deep neural network (DNN). We found that this architecture performed better than rectified linear unit (ReLU) alternatives. We also incorporate skip connections at each layer, as they have been shown to aid in

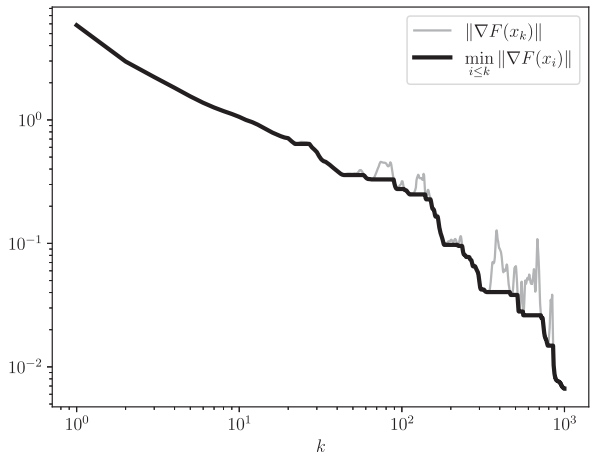

Fig. 5. Best-iterate gradient norm over 1000 iterations of PnP-PGD with an MMSE denoiser for the MayoCT computed tomography experiment in Section IV-D, plotted on a log-log scale. The quantity decays sublinearly, consistent with a $\mathcal{O}(1/\sqrt{K})$ rate.

image denoising tasks [39]. Finally, by Tweedie's formula (Theorem 2), the *true* MMSE denoiser can be written as the gradient of $\frac{1}{2}\|\cdot\|^2 + \sigma^2 \log p_Z$, where p_Z is the marginal distribution under Gaussian noise. For this reason, our choice of modeling the MMSE denoiser as the gradient of a DNN is well-motivated.

A. Numerical Illustration of Theorem 4

We first consider the denoising problem (2) with simple 1-dimensional mixture-of-Gaussian and mixture-of-Laplacian priors, denoted by p_X, with additive zero-mean, unit-variance Gaussian noise ϵ. In each case, p_Z denotes the marginal distribution of z. We now describe the DNN architecture used to learn the MMSE denoiser. Let

$$h_1 := \sigma(W_1 x), \quad W_1 \in \mathbb{R}^H,$$
$$h_k := \sigma(W_k h_{k-1} + a_{k-1} x + b_{k-1}), \quad k = 2, \dots, L, \tag{43}$$

and $g_\theta(x) := w_{\text{out}}^{\mathrm{T}} h_L + a_{\text{out}} x + b_{\text{out}}$, where $W_k \in \mathbb{R}^{H \times H}$ are the weights, $a_k, b_k \in \mathbb{R}^H$ are the skip connections and biases, and $w_{\text{out}}, a_{\text{out}}, b_{\text{out}} \in \mathbb{R}$ are the output weights and biases.

We use a softplus activation function

$$\sigma(x) := \frac{1}{\beta} \log(1 + \exp(\beta x)), \tag{44}$$

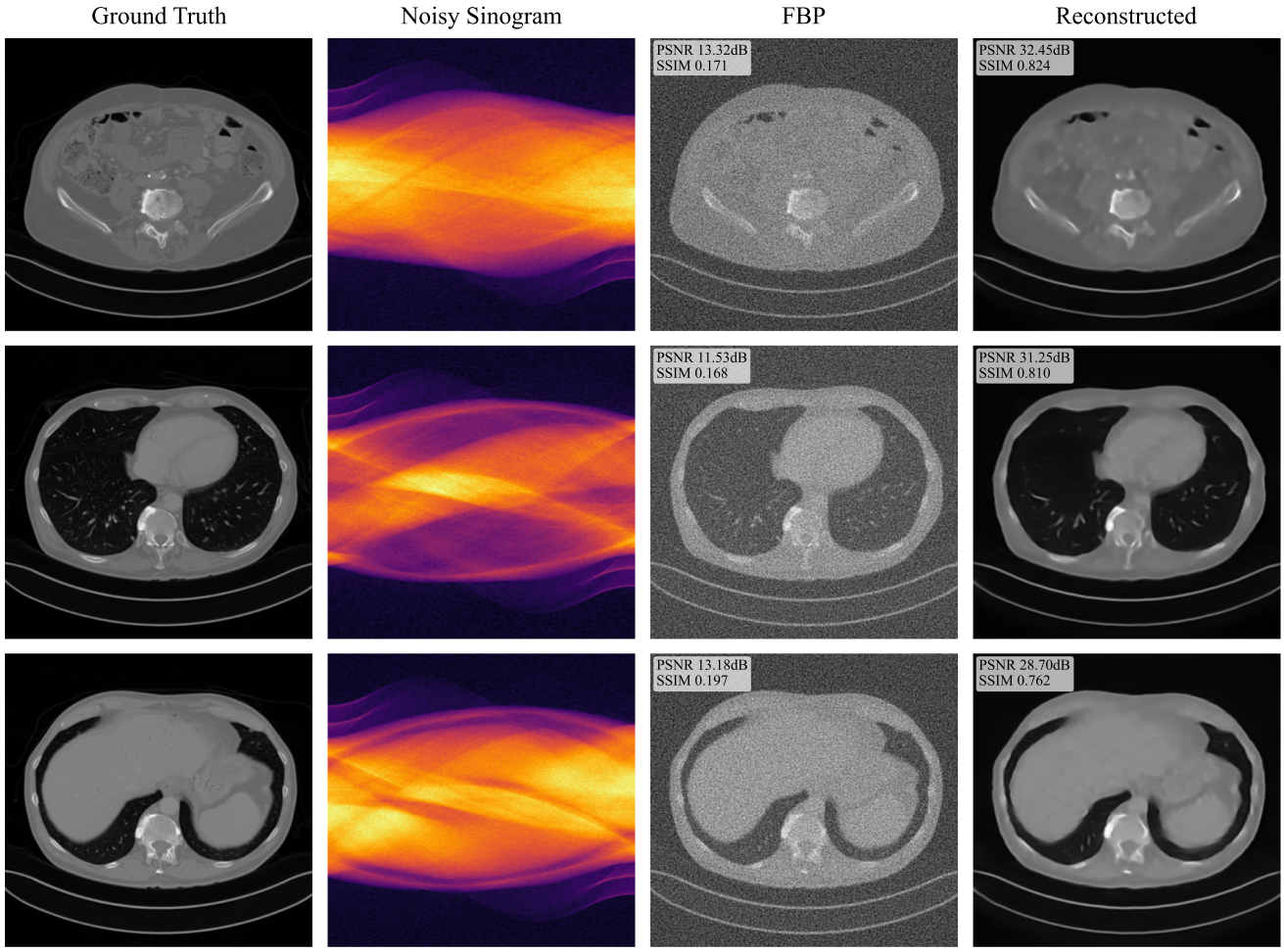

Fig. 6. PnP-PGD reconstruction results on three axial slices from the MayoCT dataset. From left to right: clean reference, noisy sinogram, filtered backprojection of the noisy sinogram, and PnP-PGD output using an MMSE denoiser with early stopping. PSNR and SSIM are reported for the FBP and reconstructed images. Reconstructions achieve PSNR gains of roughly 15–20 dB over the FBP baseline, with SSIM improvements around 0.65.

with $\beta = 10$. There are $L = 4$ hidden layers, with a hidden dimension of $H = 50$. We model our MMSE denoiser as the gradient of the DNN, i.e., $\psi_\theta := \nabla g_\theta$. To learn the denoiser, we optimize the objective

$$\mathcal{L}(\theta) := \frac{1}{B}\sum_{i=1}^{B} \|\psi_\theta(z_i) - x_i\|^2, \tag{45}$$

with a batch size $B = 2000$, $x_i \sim p_X$, and $z_i = x_i + \epsilon_i$, $\epsilon_i \sim \mathcal{N}(0,1)$. We consider the cases where p_X is either a mixture-of-Gaussian or mixture-of-Laplacian distribution. We trained the DNNs using the Adam optimizer for 500 epochs with a learning rate 10^{-3} followed by 500 epochs with a learning rate of 10^{-4}.

As discussed in Section III, the MMSE denoiser ψ_σ can be written as $\mathrm{prox}_{\phi_{\text{MMSE}}}$, where ϕ_{MMSE} is the implicit regularizer. To approximate this, we take advantage of [28, Eq. (8)] and write

$$\phi^{\text{learned}}_{\text{MMSE}}(x) := \langle \psi_\theta^{-1}(x), x\rangle - g_\theta(\psi_\theta^{-1}(x)) - \tfrac{1}{2}x^2, \tag{46}$$

where we compute ψ_θ^{-1} numerically via least-squares. When ψ_θ is the *bona fide* MMSE denoiser, $\phi^{\text{learned}}_{\text{MMSE}}$ is the *exact* implicit regularizer [28] up to a constant.

Finally, since we can find the *true* MMSE denoiser ψ_σ via Tweedie's formula (Theorem 2), we compute the corresponding implicit regularizer ϕ_{MMSE}. From (13) and Lemma 2, we have the closed-form expression

$$\phi_{\text{MMSE}}(x) = f_Z(\psi_\sigma^{-1}(x)) - \tfrac{1}{2}(x - \psi_\sigma^{-1}(x))^2 - C_{x_0}, \tag{47}$$

since $\sigma^2 = 1$ in this setting. Again, ψ_σ is inverted numerically via least-squares.

In Figs. 1 and 2 we compare the learned vs. calculated implicit regularizers for three mixture-of-Gaussian and three mixture-of-Laplacian priors, along with the reference curves $f_X := -\log p_X$ and $f_Z := -\log(p_X * \mathcal{N}(0,1))$. Since the purpose of this experiment is visual inspection, we ignore the constant C_{x_0} in Lemma 2 and shift the curves vertically so they can be easily compared.

B. DNN Architecture for Inverse Problems

We now detail the DNN architecture for the inverse problems experiments that appear in Section IV-C (Gaussian deblurring)

and Section IV-D (Computed Tomography). In both cases, we define $h_1 \coloneqq \sigma(W_1 x)$, and let

$$h_k \coloneqq \sigma(W_k h_{k-1} + A_{k-1}\mathcal{I}_k(x) + b_k), \tag{48}$$

where $k = 2, \dots, (L-1)$ and $\mathcal{I}_k(\cdot)$ denotes bilinear interpolation of x so that the dimensions match for matrix-vector multiplication. We let the scalar output be $g_\theta(x)$, defined as

$$w_{\text{out}}^{\mathrm{T}}\sigma(W_L \operatorname{vec}(h_{L-1}) + A_L \operatorname{vec}(\mathcal{I}_L(x)) + b_L). \tag{49}$$

The DNNs use a softplus activation (44) with $\beta = 10$. Each layer is either a convolutional or fully connected layer. There is a skip connection at every layer except the first and last, and we provide the exact architectural details in Table II. We train with the mean-squared error loss (45) with $z_i = x_i + \epsilon_i$ for zero-mean Gaussian noise ϵ_i with variance $\sigma^2 = 0.04$.

C. Gaussian Deblurring

We next consider the Gaussian deblurring inverse problem (1), where x is randomly sampled from the MNIST dataset of handwritten digits [35]. The forward model A is a Gaussian blurring operator, computed by circular convolution with a (3×3) Gaussian kernel with $\sigma^2 = 1$. The kernel is normalized to have unit mass. We normalize the blur operator to have operator norm less than 1 to satisfy the assumptions in Theorem 5.

We use the DNN architecture described in Section IV-B, trained using the Adam optimizer with a cosine annealing learning rate schedule. The network is trained with batches of 500 for 50000 epochs with a final learning rate of 3×10^{-5}. We let $y = Ax + \epsilon$, with ϵ zero-mean Gaussian noise with variance $\sigma^2 = 0.04$. We use the data-fidelity term $f(x) \coloneqq \frac{1}{2}\|Ax - y\|^2$, and plot $\min_{i \le K}\|\nabla F(x_i)\|$ for the first 100 iterations in Fig. 3, which we computed via (41). We display the reconstruction results in Fig. 4. Observe that there is clear sublinear decay, consistent with our theoretical results in Theorem 5 and Corollary 2.

D. Computed Tomography

We now consider the computed tomography inverse problem. The forward operator A is given by a parallel beam Radon transform discretized via the Operator Discretization Library (ODL) [1]. We scale the operator by $1/(1 + \|A\|_{\text{op}})$ in accordance with Theorem 5. We use 200 projection angles, uniformly spaced over $[0, \pi)$, with 400 detector bins, implemented using the Astra toolbox [57] with CUDA acceleration.

We use the DNN architecture described in Section IV-B, trained using the Adam optimizer with a cosine annealing learning rate schedule. The network is trained with batches of 64 for 40000 epochs with a final learning rate of 10^{-6}. We use training and testing data from the MayoCT dataset [40] (resolution of 512×512, pixel values ranging from 0 to 1). We let $y = Ax + \epsilon$, with ϵ zero-mean Gaussian noise with variance $\sigma^2 = 0.04$. We use the data-fidelity term $f(x) \coloneqq \frac{1}{2}\|Ax - y\|^2$ and plot $\min_{i \le K}\|\nabla F(x_i)\|$ for the first 1000 iterations in Fig. 5, computed via (41). We display the reconstruction results for three slices of the MayoCT dataset in Fig. 6. There is clear sublinear decay, consistent with our theoretical result in Theorem 5 and Corollary 2.

V. Conclusion

We presented a novel representation of the MMSE denoiser's implicit regularizer in Theorem 4 and used it to derive nonasymptotic convergence results for PnP-PGD using an MMSE denoiser in Theorem 5 and Corollary 2. Our analysis is strictly tied to the Gaussian-noise MMSE setting. Tweedie's formula, for example, is not the same under other noise models [21], so we suspect this flavor of result is unique to the Gaussian setting. However, we expect analogous nonasymptotic guarantees to be attainable for other PnP schemes that use MMSE denoisers under Gaussian noise, such as PnP-ADMM or the proximal point method. Exploration of these extensions constitutes future work.

Acknowledgment

The authors would like to thank Stanislas Ducotterd and Guillaume Lauga for helpful feedback on this work.

References

[1] J. Adler, H. Kohr, and O. Öktem, "Operator discretization library (ODL)," *Zenodo*, 2017.
[2] R. Ahmad et al., "Plug-and-play methods for magnetic resonance imaging: Using denoisers for image recovery," *IEEE Signal Process. Mag.*, vol. 37, no. 1, pp. 105–116, Jan. 2020.
[3] C. Anil, J. Lucas, and R. Grosse, "Sorting out Lipschitz function approximation," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 291–301.
[4] C. D. Athalye, K. N. Chaudhury, and B. Kumar, "On the contractivity of plug-and-play operators," *IEEE Signal Process. Lett.*, vol. 30, pp. 1447–1451, 2023.
[5] H. Attouch, J. Bolte, and B. F. Svaiter, "Convergence of descent methods for semi-algebraic and tame problems: Proximal algorithms, forward–backward splitting, and regularized Gauss–Seidel methods," *Math. Program.*, vol. 137, no. 1, pp. 91–129, 2013.
[6] H. H. Bauschke and P. L. Combettes, "Convex analysis and monotone operator theory in Hilbert spaces," *CMS Books Math., Ouvrages de mathématiques de la SMC*, 2011.
[7] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imag. Sci.*, vol. 2, no. 1, pp. 183–202, 2009.
[8] P. Bernard, "Lasry-lions regularization and a lemma of ilmanen," *Rendiconti del Seminario Matematico della Università di Padova*, vol. 124, pp. 221–229, 2010.
[9] A. Bora, A. Jalal, E. Price, and A. G. Dimakis, "Compressed sensing using generative models," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 537–546.
[10] E. J. Candes, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.
[11] A. Chambolle, "An algorithm for total variation minimization and applications," *J. Math. Imag. Vis.*, vol. 20, no. 1, pp. 89–97, 2004.
[12] A. Chambolle and T. Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging," *J. Math. Imag. Vis.*, vol. 40, no. 1, pp. 120–145, 2011.
[13] S. H. Chan, X. Wang, and O. A. Elgendy, "Plug-and-play ADMM for image restoration: Fixed-point convergence and applications," *IEEE Trans. Comput. Imag.*, vol. 3, no. 1, pp. 84–98, Mar. 2017.
[14] P. L. Combettes, "Solving monotone inclusions via compositions of nonexpansive averaged operators," *Optimization*, vol. 53, no. 5–6, pp. 475–504, 2004.
[15] P. L. Combettes and J.-C. Pesquet, "Proximal splitting methods in signal processing," in *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*. New York, NY, USA: Springer, 2011, pp. 185–212.
[16] P. L. Combettes and V. R. Wajs, "Signal recovery by proximal forward-backward splitting," *Multiscale Model. Simul.*, vol. 4, no. 4, pp. 1168–1200, 2005.

[17] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3-D transform-domain collaborative filtering," *IEEE Trans. Image Process.*, vol. 16, no. 8, pp. 2080–2095, Aug. 2007.
[18] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.
[19] S. Ducotterd, S. Neumayer, and M. Unser, "Learning of patch-based smooth-plus-sparse models for image reconstruction," in *Proc. 2nd Conf. Parsimony Learn.*, 2025, pp. 89–104.
[20] S. Ducotterd and M. Unser, "Multivariate fields of experts for convergent image reconstruction," *IEEE Trans. Comput. Imag.*, vol. 12, pp. 827–838, 2026.
[21] B. Efron, "Tweedie's formula and selection bias," *J. Amer. Stat. Assoc.*, vol. 106, no. 496, pp. 1602–1614, 2011.
[22] Z. Fang, S. Buchanan, and J. Sulam, "What's in a prior? Learned proximal networks for inverse problems," in *Proc. Int. Conf. Learn. Representations*, 2024, pp. 55445–55485.
[23] R. G. Gavaskar, C. D. Athalye, and K. N. Chaudhury, "On plug-and-play regularization using linear denoisers," *IEEE Trans. Image Process.*, vol. 30, pp. 4802–4813, 2021.
[24] R. G. Gavaskar and K. N. Chaudhury, "Plug-and-play ISTA converges with kernel denoisers," *IEEE Signal Process. Lett.*, vol. 27, pp. 610–614, 2020.
[25] A. Goujon, S. Neumayer, P. Bohra, S. Ducotterd, and M. Unser, "A neural-network-based convex regularizer for inverse problems," *IEEE Trans. Comput. Imag.*, vol. 9, pp. 781–795, 2023.
[26] A. Goujon, S. Neumayer, and M. Unser, "Learning weakly convex regularizers for convergent image-reconstruction algorithms," *SIAM J. Imag. Sci.*, vol. 17, no. 1, pp. 91–115, 2024.
[27] R. Gribonval and P. Machart, "Reconciling 'priors' & 'priors' without prejudice?," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 26, 2013, pp. 1–9.
[28] R. Gribonval and M. Nikolova, "A characterization of proximity operators," *J. Math. Imag. Vis.*, vol. 62, no. 6, pp. 773–789, 2020.
[29] R. Gribonval, "Should penalized least squares regression be interpreted as maximum a posteriori estimation?," *IEEE Trans. Signal Process.*, vol. 59, no. 5, pp. 2405–2410, May 2011.
[30] J. Hertrich et al., "Learning regularization functionals for inverse problems: A comparative study," Ser. Handbook of Numerical Analysis, Elsevier, 2026, doi: 10.1016/bs.hna.2026.04.001.
[31] S. Hurault, A. Chambolle, A. Leclaire, and N. Papadakis, "Convergent plug-and-play with proximal denoiser and unconstrained regularization parameter," *J. Math. Imag. Vis.*, vol. 66, no. 4, pp. 616–638, 2024.
[32] S. Hurault, A. Leclaire, and N. Papadakis, "Gradient step denoiser for convergent plug-and-play," in *Proc. Int. Conf. Learn. Representations*, 2022, pp. 1–30.
[33] S. Hurault, A. Leclaire, and N. Papadakis, "Proximal denoiser for convergent plug-and-play optimization with nonconvex regularization," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 9483–9505.
[34] S. M. Kakade, "Dimension free optimization and non-convex optimization," CSE 547 Machine Learning for Big Data Lecture Slides, University of Washington, 2018, (lecture Notes). New, York, NY, USA: Spring, 2018.
[35] Y. LeCun, C. Cortes, and C. Burges, "MNIST handwritten digit database," 2010.
[36] H. Li and Z. Lin, "Accelerated proximal gradient methods for nonconvex programming," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 1–9.
[37] Q. Li, S. Haque, C. Anil, J. Lucas, R. B. Grosse, and J.-H. Jacobsen, "Preventing gradient attenuation in Lipschitz constrained convolutional networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–13.
[38] J. Liu, Y. Sun, X. Xu, and U. S. Kamilov, "Image restoration using total variation regularized deep image prior," in *Proc. ICASSP IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 7715–7719.
[39] X.-J1-4 Mao, C. Shen, and Y.-B. Yang, "Image restoration using convolutional auto-encoders with symmetric skip connections," 2016, *arXiv:1606.08921*.
[40] C. McCollough, "TU-FG-207A-04: Overview of the low dose CT grand challenge," *Med. Phys.*, vol. 43, pp. 3759–3760, Jun. 2016.
[41] P. Milanfar and M. Delbracio, "Denoising: A powerful building block for imaging, inverse problems and machine learning," *Philos. Trans. A*, vol. 383, no. 2299, 2025, Art. no. 20240326.
[42] J.-J. Moreau, "Proximité et dualité dans un espace hilbertien," *Bull. de la Société Mathématique de France*, vol. 93, pp. 273–299, 1965.
[43] P. Nair, R. G. Gavaskar, and K. N. Chaudhury, "Fixed-point and objective convergence of plug-and-play algorithms," *IEEE Trans. Comput. Imag.*, vol. 7, pp. 337–348, 2021.
[44] G. Ongie, A. Jalal, C. A. Metzler, R. G. Baraniuk, A. G. Dimakis, and R. Willett, "Deep learning techniques for inverse problems in imaging," *IEEE J. Sel. Areas Inf. Theory*, vol. 1, no. 1, pp. 39–56, May 2020.
[45] R. Parhi and R. D. Nowak, "Deep learning meets sparse regularization: A signal processing perspective," *IEEE Signal Process. Mag.*, vol. 40, no. 6, pp. 63–74, Sep. 2023.
[46] R. Parhi and M. Unser, "The sparsity of cycle spinning for wavelet-based solutions of linear inverse problems," *IEEE Signal Process. Lett.*, vol. 30, pp. 568–572, 2023.
[47] N. Parikh and S. Boyd, "Proximal algorithms," *Found. Trends Optim.*, vol. 1, no. 3, pp. 127–239, 2014.
[48] C. Park, S. Shoushtari, W. Gan, and U. S. Kamilov, "Convergence of nonconvex PnP-ADMM with MMSE denoisers," in *Proc. IEEE 9th Int. Workshop Comput. Adv. Multi-Sensor Adaptive Process.*, 2023, pp. 511–515.
[49] J.-C. Pesquet, A. Repetti, M. Terris, and Y. Wiaux, "Learning maximally monotone operators for image recovery," *SIAM J. Imag. Sci.*, vol. 14, no. 3, pp. 1206–1237, 2021.
[50] M. Pourya, E. Kobler, M. Unser, and S. Neumayer, "DEALing with image reconstruction: Deep attentive least squares," in *Proc. Int. Conf. Mach. Learn.*, 2025, pp. 49689–49708.
[51] R. T. Rockafellar and R. J. Wets, *Variational Analysis*. New York, NY, USA: Springer, 1998.
[52] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D: Nonlinear Phenomena*, vol. 60, no. 1-4, pp. 259–268, 1992.
[53] E. Ryu, J. Liu, S. Wang, X. Chen, Z. Wang, and W. Yin, "Plug-and-play methods provably converge with properly trained denoisers," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 5546–5557.
[54] Z. Shumaylov, J. Budd, S. Mukherjee, and C.-B. Schönlieb, "Weakly convex regularisers for inverse problems: Convergence of critical points and primal-dual optimisation," in *Proc. 41st Int. Conf. Mach. Learn.*, vol. 235, 2024, pp. 45286–45314.
[55] C. Stein, "A bound for the error in the normal approximation to the distribution of a sum of dependent random variables," in *Proc. 6th Berkeley Symp. Math. Statist. Probability, Volume 2: Probability Theory*, 1972, vol. 6, pp. 583–603.
[56] Y. Sun, B. Wohlberg, and U. S. Kamilov, "An online plug-and-play algorithm for regularized image reconstruction," *IEEE Trans. Comput. Imag.*, vol. 5, no. 3, pp. 395–408, Sep. 2019.
[57] W. Van Aarle et al., "Fast and flexible X-ray tomography using the ASTRA toolbox," *Opt. Exp.*, vol. 24, no. 22, pp. 25129–25147, 2016.
[58] S. V. Venkatakrishnan, C. A. Bouman, and B. Wohlberg, "Plug-and-play priors for model based reconstruction," in *Proc. IEEE Glob. Conf. Signal Inf. Process.*, 2013, pp. 945–948.
[59] X. Xu, Y. Sun, J. Liu, B. Wohlberg, and U. S. Kamilov, "Provable convergence of plug-and-play priors with MMSE denoisers," *IEEE Signal Process. Lett.*, vol. 27, pp. 1280–1284, 2020.