Proximal Denoiser for Convergent Plug-and-Play Optimization with Nonconvex Regularization

Samuel Hurault [1] **Arthur Leclaire** [1] **Nicolas Papadakis** [1]

Abstract

Plug-and-Play (PnP) methods solve ill-posed inverse problems through iterative proximal algorithms by replacing a proximal operator by a denoising operation. When applied with deep neural network denoisers, these methods have shown state-of-the-art visual performance for image restoration problems. However, their theoretical convergence analysis is still incomplete. Most of the existing convergence results consider nonexpansive denoisers, which is non-realistic, or limit their analysis to strongly convex data-fidelity terms in the inverse problem to solve. Recently, it was proposed to train the denoiser as a gradient descent step on a functional parameterized by a deep neural network. Using such a denoiser guarantees the convergence of the PnP version of the Half-Quadratic-Splitting (PnP-HQS) iterative algorithm. In this paper, we show that this gradient denoiser can actually correspond to the proximal operator of another scalar function. Given this new result, we exploit the convergence theory of proximal algorithms in the nonconvex setting to obtain convergence results for PnP-PGD (Proximal Gradient Descent) and PnP-ADMM (Alternating Direction Method of Multipliers). When built on top of a smooth gradient denoiser, we show that PnP-PGD and PnP-ADMM are convergent and target stationary points of an explicit functional. These convergence results are confirmed with numerical experiments on deblurring, super-resolution and inpainting. [1]

[1]Univ. Bordeaux, Bordeaux INP, CNRS, IMB, UMR 5251,F-33400 Talence, France. Correspondence to: Samuel Hurault <samuel.hurault@math.u-bordeaux.fr>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

[1]Code is available at https://github.com/samuro95/Prox-PnP.

1. Introduction

Many image restoration (IR) tasks can be addressed by solving an optimization problem

$$x^* \in \arg\min_x \lambda f(x) + g(x) \tag{1}$$

where f is a data-fidelity term, g a regularization term and $\lambda > 0$ a parameter that controls the strength of the regularization. In particular, we will consider the case $f(x) = \frac{1}{2\sigma^2}||Ax - y||^2$ that corresponds to the linear observation model $y = Ax^* + \xi$ where y is the degraded image, x^* is the clean image that we want to recover, A is a linear operator and ξ a white Gaussian noise of standard deviation σ. In several applications (like deblurring or inpainting), the operator A is non-invertible, thus leading to an ill-posed inverse problem. Including a term g allows to cope with the ill-posed nature of this problem by assuming a-priori that x must be regular in some sense. A long-standing problem consists in designing functions $g(x)$ that reflect a relevant regularity prior on x while allowing for efficient numerical schemes to solve (1).

For early image models, based e.g. on Fourier spectrum (Ruderman, 1994), total variation (Rudin et al., 1992), wavelet sparsity (Mallat, 2009) or patch-based Gaussian mixtures (Zoran & Weiss, 2011), the problem (1) can be solved either explicitly (with linear filtering) or with provably convergent algorithms (like iterative thresholding (Daubechies et al., 2004)). More recently, tremendous progress has been made in IR by adopting deep image models that are learned from a database of clean images (Lunz et al., 2018; Prost et al., 2021; González et al., 2021). It is still an open question to understand how deep models can be encoded in a regularization term $g(x)$ that has sufficient properties to apply well-known optimization techniques.

One fruitful way to tackle this problem is to rely on first-order proximal splitting algorithms (Combettes & Pesquet, 2011). These algorithms operate individually on f and g via the proximity operator defined for a stepsize $\tau > 0$ as

$$\mathrm{Prox}_{\tau f}(x) = \arg\min_z \frac{1}{2\tau}||x - z||^2 + f(z), \tag{2}$$

Among these algorithms, the Proximal Gradient Descent (PGD) (also called Forward-Backward Splitting) al-

ternates between a proximal operation on g and a gradient descent step on f, when f is differentiable. In more general settings, Half-Quadratic-Splitting (HQS), Alternating Direction Method of Multipliers (ADMM) or Douglas-Rachford Splitting (DRS) use proximal operations for both f and g. Plug-and-Play (PnP) methods (Venkatakrishnan et al., 2013) draw an elegant connexion between proximal methods and deep image models by replacing the proximity operator of g with an image denoiser. State-of-the-art results for various IR problems have then been obtained with PnP methods (Meinhardt et al., 2017; Buzzard et al., 2018; Ahmad et al., 2020; Yuan et al., 2020; Zhang et al., 2021), but with scarce theoretical guarantees. Indeed, since a generic deep denoiser cannot be expressed as a proximal mapping (Moreau, 1965) in general, convergence results relying on properties of the proximal operator do not follow easily. Moreover, the regularizer g is only made implicit via the denoising operation. Therefore, PnP algorithms do not seek the minimization of an explicit objective functional which strongly limits their interpretation and numerical control.

In this paper, we tackle these issues by shedding a new light on the gradient-step denoiser proposed in (Hurault et al., 2022; Cohen et al., 2021), which writes

$$D_\sigma = \mathrm{Id} - \nabla g_\sigma \tag{3}$$

where $g_\sigma : \mathbb{R}^n \to \mathbb{R}$ is a scalar function parameterized by a differentiable neural network. (Hurault et al., 2022) have shown that, if this denoiser is used in PnP-HQS, one can ensure convergence of the iterates towards a stationary point of an explicit functional. This trick works specifically for PnP-HQS and does not extend to other PnP frameworks.

Contributions. In this work, we propose to plug a denoiser trained as a proximal mapping, *i.e.* an *implicit* gradient step. We first make use of the results from (Gribonval, 2011; Gribonval & Nikolova, 2020) on the characterization of proximity operators, to demonstrate that the gradient-step denoiser (3) can actually be the exact proximal operator of some nonconvex smooth function. With this result, we show that PnP-PGD, PnP-ADMM and PnP-DRS algorithms are guaranteed to converge to stationary points of an explicit functional.

2. Related works

PnP methods were applied successfully on several IR tasks by using off-the-shelf deep denoisers such as DnCNN (Zhang et al., 2017a) or DRUNet (Zhang et al., 2021). These denoisers are used in various proximal algorithms such as HQS (Zhang et al., 2017b; 2021), ADMM and DRS (Romano et al., 2017; Ryu et al., 2019), Proximal Gradient Descent (PGD) (Terris et al., 2020). While convergence of PnP methods using pseudo-linear denoisers is established (Sreehari et al., 2016; Gavaskar & Chaudhury, 2020; Nair et al., 2021; Chan, 2019), little is known about convergence with deep denoisers.

Nonexpansive denoisers Most of existing works intend to get convergence via contractive PnP fixed-point iterations. For instance, convergence of various proximal algorithms can be obtained by assuming the denoiser averaged (Sun et al., 2019), firmly nonexpansive (Sun et al., 2021; Terris et al., 2020) or simply nonexpansive (Reehorst & Schniter, 2018; Liu et al., 2021). As native off-the-shelf deep denoisers are generally not 1-Lipschitz, several authors proposed to train deep denoisers with constrained Lipschitz constants. For example, Ryu et al. (2019) normalize each layer individually by its spectral norm, but with such a strategy, one cannot use residual skip connections, which are widely uses in deep denoisers. Terris et al. (2020) propose instead to enrich the training loss of the denoisers with a penalization on the spectral norm of the Jacobian.

However, imposing nonexpansiveness to the denoiser is very likely to hurt its denoising performance (see Appendix F). Ryu et al. (2019) argue that it is more realistic to suppose nonexpansiveness of the residual. This nevertheless comes at the cost of imposing strong convexity on the data term f in the IR problem (1), which excludes many tasks like deblurring, super-resolution and inpainting. Finally, let us recall that all the strategies based on nonexpansiveness have another major limitation. These methods target fixed points of some (firmly) nonexpansive or contractive operators but do not minimize an explicit functional, thus making convergence difficult to monitor.

Gradient step denoisers With the Regularization by Denoising framework (RED), (Romano et al., 2017) shows that under homogeneity, nonexpansiveness and Jacobian symmetry conditions, a denoiser can be written as a gradient descent step realized on a convex potential (as in (3)). However, as shown in (Reehorst & Schniter, 2018), such conditions are unrealistic for deep denoisers. As a consequence, it is proposed in (Cohen et al., 2021) and (Hurault et al., 2022) to directly train the denoiser as an explicit gradient step performed on a possibly nonconvex function. Plugging this denoiser in the PnP-HQS algorithm, it provides a convergent scheme that targets a stationary point of an explicit function. However, this strategy does not extend to other PnP frameworks such as PnP-PGD or PnP-ADMM.

Proximal denoisers As PnP methods were built by replacing a proximal operation by a denoiser, it is interesting to investigate under which conditions a denoiser can actually be a proximal map. The theorem of Moreau (Moreau, 1965) states that we get a proximal operator of some *convex* potential if and only if it is nonexpansive and the sub-gradient of

a convex function. (Sreehari et al., 2016) exploit this result to show the convergence of PnP-ADMM, but the analysis is thus limited to nonexpansive (and pseudo-linear) denoisers. The purpose of this paper is to extend such results to deep image denoisers that are not nonexpansive. To do so, we will focus on learning a deep image denoiser as the proximal operator of a *nonconvex* functional. Under this condition, we provide new proofs of convergence of the iterates of PnP-PGD, PnP-ADMM and PnP-DRS, towards stationary points of explicit functions.

3. Proximal gradient step denoiser

Recently introduced in (Hurault et al., 2022) and (Cohen et al., 2021) the Gradient Step (GS) Denoiser (3) writes

$$D_\sigma = \nabla h_\sigma, \tag{4}$$

with a potential

$$h_\sigma : x \to \frac{1}{2}||x||^2 - g_\sigma(x), \tag{5}$$

obtained from a scalar function $g_\sigma : \mathbb{R}^n \to \mathbb{R}$ parameterized by a differentiable neural network. This denoiser D_σ is trained to denoise images degraded with Gaussian noise of level σ. In (Hurault et al., 2022), it is shown that, although constrained to be an exact conservative field (4), it can realize state-of-the-art denoising.

In this section, we make use of the recent works (Gribonval, 2011; Gribonval & Nikolova, 2020) on the characterization of proximity operators to show that this gradient denoiser can be written as a proximal operator. These results generalize the theorem of Moreau (Moreau, 1965) to denoisers that may not be nonexpansive. Applying Theorem 1 in (Gribonval & Nikolova, 2020), we first have that if h_σ is convex, the GS denoiser $D_\sigma = \nabla h_\sigma$ is linked to the proximal operator of some function $\phi_\sigma : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$: $\forall x \in \mathbb{R}^n$, $D_\sigma(x) \in \operatorname{Prox}_{\phi_\sigma}(x) = \arg\min_z \frac{1}{2}||x-z||^2 + \phi_\sigma(z)$. The next proposition shows that, if the residual $\mathrm{Id} - D_\sigma$ is contractive, there exists a closed-form and smooth ϕ_σ such that $D_\sigma = \operatorname{Prox}_{\phi_\sigma}$ is single-valued.

Proposition 3.1 (proof in Appendix A)**.** *Let $\mathcal{X}$ be an open convex subset of $\mathbb{R}^n$ and $g_\sigma : \mathcal{X} \to \mathbb{R}$ a $\mathcal{C}^{k+1}$ function with $k \geq 1$ and ∇g_σ L-Lipschitz with $L < 1$. Then, for $h_\sigma : x \to \frac{1}{2}||x||^2 - g_\sigma(x)$ and $D_\sigma := \nabla h_\sigma = \mathrm{Id} - \nabla g_\sigma$,*

(i) h_σ is $(1-L)$-strongly convex and $\forall x \in \mathcal{X}$, $J_{D_\sigma}(x)$ is positive definite

(ii) D_σ is injective, $\operatorname{Im}(D_\sigma)$ is open and, $\forall x \in \mathcal{X}$, $D_\sigma(x) = \operatorname{Prox}_{\phi_\sigma}(x)$, with $\phi_\sigma : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ defined by

$$\phi_\sigma(x) := \begin{cases} g_\sigma({D_\sigma}^{-1}(x))) - \frac{1}{2}\left|\left|{D_\sigma}^{-1}(x) - x\right|\right|^2 & \text{if } x \in \operatorname{Im}(D_\sigma), \\ +\infty & \text{otherwise}, \end{cases} \tag{6}$$

(iii) $\forall x \in \mathcal{X}$, $\phi_\sigma(x) \geq g_\sigma(x)$ and for $x \in \operatorname{Fix}(D_\sigma)$, $\phi_\sigma(x) = g_\sigma(x)$.

(iv) ϕ_σ is $\mathcal{C}^k$ on $\operatorname{Im}(D_\sigma)$ and $\forall x \in \operatorname{Im}(D_\sigma)$, $\nabla\phi_\sigma(x) = {D_\sigma}^{-1}(x) - x = \nabla g_\sigma({D_\sigma}^{-1}(x))$

(v) $\nabla\phi_\sigma$ is $\frac{L}{1-L}$-Lipschitz on $\operatorname{Im}(D_\sigma)$

Remark 3.2. Note that despite ϕ_σ being possibly nonconvex, $D_\sigma = \operatorname{Prox}_{\phi_\sigma}$ is one-to-one. In the literature, this is referred to as prox-regularity of ϕ_σ. Also note that D_σ is possibly not nonexpansive.

PnP algorithms, such as PnP-PGD or PnP-ADMM, were built by replacing one proximal operation with a denoiser D_σ. When used with a denoiser D_σ satisfying the assumptions of Proposition 3.1, PnP algorithms become classical proximal algorithms, with the notable difference that a nonconvex function is involved.

4. Convergence analysis of PnP methods with proximal denoisers

In this section, we study the convergence of the three algorithms PnP-PGD, PnP-ADMM and PnP-DRS, with a plugged denoiser $D_\sigma = \operatorname{Prox}_{\phi_\sigma}$ that corresponds to the proximal operator of a nonconvex regularization function ϕ_σ. For that purpose, we target the objective function

$$F_{\lambda,\sigma} := \lambda f + \phi_\sigma. \tag{7}$$

where f is a (possibly nonconvex) data-fidelity term, λ is a regularization parameter and ϕ_σ is defined as in Proposition 3.1 from the function g_σ satisfying $D_\sigma = \mathrm{Id} - \nabla g_\sigma$. In all this analysis, to use Proposition 3.1, g_σ is assumed $\mathcal{C}^2$ with contractive gradient ($L < 1$). We also assume f and g_σ bounded from below. From Proposition 3.1 (iii), we get that ϕ_σ and thus $F_{\lambda,\sigma}$ are also bounded from below.

We show, under some conditions on the parameters, that all three algorithms give convergence in terms of function values, and also convergence of the iterates if the generated sequences are bounded and $F_{\lambda,\sigma}$ verifies the Kurdyka-Lojasiewicz (KL) property. The boundedness of the generated sequences is discussed in Appendix D. The KL property has been widely used to study the convergence of optimization algorithms in the nonconvex setting (Attouch et al., 2010; 2013; Ochs et al., 2014). Very large classes of functions, in particular all the real semi-algebraic functions, satisfy this property.

4.1. Convergence of PnP-PGD

The PnP-PGD algorithm is adapted from the standard proximal gradient descent (PGD) algorithm where the proximal step is replaced by a denoising operation. In the usual explicit gradient step $\mathrm{Id} - \tau\lambda\nabla f$ used in PGD, we fix the

stepsize $\tau = 1$. With a regularization parameter $\lambda > 0$, PnP-PGD then writes

$$\begin{cases} z_{k+1} = x_k - \lambda \nabla f(x_k) \\ x_{k+1} = D_\sigma(z_{k+1}) \end{cases} \tag{8}$$

At each iteration, $x_k \in \mathrm{Im}(D_\sigma)$, and we get from relations (6), (7) and (8) the value of the objective function

$$F_{\lambda,\sigma}(x_k) = \lambda f(x_k) + g_\sigma(z_k) - \frac{1}{2}\left|\left|z_k - x_k\right|\right|^2. \tag{9}$$

To study the convergence of PnP-PGD, we make use of the literature on the convergence analysis (Attouch et al., 2013; Beck, 2017; Li & Lin, 2015) of the PGD for minimizing the sum of two possibly nonconvex functions. We assume here the data-fidelity term f differentiable with Lipschitz gradient, but not necessarily convex.

Theorem 4.1 (Convergence of PnP-PGD, proof in Appendix B)**.** *Let $g_\sigma : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ of class $\mathcal{C}^2$ with L-Lipschitz gradient, $L < 1$, and $D_\sigma := \mathrm{Id} - \nabla g_\sigma$. Let ϕ_σ defined from g_σ and D_σ as in (6). Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ differentiable with L_f-Lipschitz gradient. Assume that f and g_σ are bounded from below. Then, for $\lambda L_f < 1$, the iterates x_k given by the iterative scheme (8) verify*

(i) $(F_{\lambda,\sigma}(x_k))$ *is non-increasing and converges.*

(ii) The residual $||x_{k+1} - x_k||$ *converges to* 0 *at rate* $\min_{k \leq K} ||x_{k+1} - x_k||^2 = \mathcal{O}(1/K)$

(iii) All cluster points of the sequence (x_k) *are stationary points of* $F_{\lambda,\sigma}$.

(iv) Additionally suppose that f and g_σ are respectively KL and semi-algebraic, then if the sequence (x_k) is bounded, it converges, with finite length, to a stationary point of $F_{\lambda,\sigma}$.

Remark 4.2. If ∇g_σ has Lipschitz constant $L > 1$, similar to (Hurault et al., 2022), we can introduce $0 < \alpha < \frac{1}{L}$ and relax the denoising operation by replacing D_σ with $D_\sigma^\alpha = \alpha D_\sigma + (1-\alpha)\,\mathrm{Id} = \mathrm{Id} - \alpha \nabla g_\sigma$ in (8). We can then define ϕ_σ^α from $g_\sigma^\alpha = \alpha g_\sigma$ as in (6) and using the modified PnP-PGD algorithm $x_{k+1} = D_\sigma^\alpha(x_k - \lambda \nabla f(x_k))$, we get the same convergence results than in Theorem 4.1.

Notice that in Theorem 4.1, the usual condition on the stepsize becomes a condition on the regularization parameter $\lambda L_f < 1$. The regularization trade-off parameter is then limited by the value of L_f. Even if this is not usual in optimization, we argue that this is a not a problem as the regularization strength is also regulated by the σ parameter which we are free to tune manually.

4.2. Convergence of PnP-ADMM and PnP-DRS

The classical PnP version of ADMM, PnP-ADMM, for a given stepsize $\tau > 0$, can be written as

$$\begin{cases} y_{k+1} = \mathrm{Prox}_{\tau\lambda f}(x_k) \\ z_{k+1} = D_\sigma(y_{k+1} + x_k) \\ x_{k+1} = x_k + (y_{k+1} - z_{k+1}) \end{cases} \tag{10}$$

Following (Eckstein & Fukushima, 1994), one can show that PnP-ADMM is equivalent to PnP-DRS

$$\begin{cases} y_{k+1} &= \mathrm{Prox}_{\tau\lambda f}(x_k) \\ z_{k+1} &= D_\sigma(2y_{k+1} - x_k) \\ x_{k+1} &= x_k + \beta(z_{k+1} - y_{k+1}) \end{cases} \tag{11}$$

with a relaxation parameter $\beta = 1$.

The authors of (Li & Pong, 2016) and (Themelis & Patrinos, 2020) propose convergence proofs of the DRS algorithm for the minimization of the sum of two nonconvex functions, one of the two functions being differentiable with Lipschitz gradient. We will adapt their results to obtain convergence results of PnP-DRS. By equivalence, the convergence of PnP-ADMM follows immediately.

4.2.1. Differentiable data-fidelity term

We first consider f convex differentiable with Lipschitz gradient. Like for PnP-PGD, we fix the stepsize at $\tau = 1$ in (11) and use a regularization parameter $\lambda > 0$. PnP-DRS with $\beta = 1$ then realizes

$$\begin{cases} y_{k+1} = \mathrm{Prox}_{\lambda f}(x_k) \\ z_{k+1} = D_\sigma(2y_{k+1} - x_k) = \mathrm{Prox}_{\phi_\sigma}(2y_{k+1} - x_k) \\ x_{k+1} = x_k + (z_{k+1} - y_{k+1}) \end{cases} \tag{12}$$

As in (Themelis & Patrinos, 2020), we consider for the Lyapunov function the Douglas-Rachford Envelope

$$F_{\lambda,\sigma}^{DR,1}(x) = \phi_\sigma(z) + \lambda f(y) + \langle y - x, y - z\rangle + \frac{1}{2}\left|\left|y - z\right|\right|^2 \tag{13}$$

where, for a given x, y and z are obtained by the two first steps of the DRS iterations (12).

Under the same assumptions as Theorem 4.1, we now show that we also have convergence of PnP-DRS and PnP-ADMM towards stationary points of $F_{\lambda,\sigma}$.

Theorem 4.3 (Convergence of PnP-DRS with f differentiable, proof in Appendix C.1)**.** *Let $g_\sigma : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ of class $\mathcal{C}^2$ with L-Lipschitz gradient with $L < 1$. Let $D_\sigma := \mathrm{Id} - \nabla g_\sigma$. Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be convex and differentiable with L_f-Lipschitz gradient. Assume that f and g_σ are bounded from below. Then, for $\lambda L_f < 1$, the iterates (y_k, z_k, x_k) given by the iterative scheme (12) verify*

(i) $(F_{\lambda,\sigma}^{DR,1}(x_k))$ *is nonincreasing and converges.*

(ii) The residual $||y_k - z_k||$ *converges to* 0 *at rate* $\min_{k \leq K} ||y_k - z_k||^2 = \mathcal{O}(1/K)$

(iii) (y_k) *and* (z_k) *have the same cluster points, all of them being stationary for* $F_{\lambda,\sigma}$ *with the same value of* $F_{\lambda,\sigma}$.

(iv) Additionally suppose that f *and* g_σ *are respectively KL and semi-algebraic, then if the sequence* (y_k, z_k, x_k) *is bounded, the whole sequence converges, and* y_k *and* z_k *converge to the same stationary point of* $F_{\lambda,\sigma}$.

4.2.2. Non differentiable data-fidelity term

To cope with a possibly non-differentiable data-fidelity term, and to get rid of the restriction on the trade-off parameter λ, we inverse the denoising and proximal steps in (12):

$$\begin{cases} y_{k+1} = D_\sigma(x_k) \\ z_{k+1} = \operatorname{Prox}_{\lambda f}(2y_{k+1} - x_k) \\ x_{k+1} = x_k + (z_{k+1} - y_{k+1}) \end{cases} \tag{14}$$

and adapt the Douglas-Rachford Envelope accordingly

$$F^{DR,2}_{\lambda,\sigma}(x) = \phi_\sigma(y) + \lambda f(z) + \langle y - x, y - z\rangle + \frac{1}{2}||y - z||^2 \tag{15}$$

where, for a given x, y and z are obtained by the two first steps of the DRS iterations (14).

Convergence of this algorithm would be ensured for ϕ_σ Lipschitz differentiable on $\mathbb{R}^n$. However, Proposition 3.1 shows that ϕ_σ is differentiable only on $\operatorname{Im}(D_\sigma)$. We now show that, if the set $\operatorname{Im}(D_\sigma)$ is convex and $L < \frac{1}{2}$, we still have convergence of the iterates in (14).

Theorem 4.4 (Convergence of PnP-DRS with f not differentiable, proof in Appendix C.2)**.** *Let* $g_\sigma : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ *of class* $\mathcal{C}^2$ *with L-Lipschitz gradient with* $L < \frac{1}{2}$. *Let* $D_\sigma := \mathrm{Id} - \nabla g_\sigma$. *Assume that* $\operatorname{Im}(D_\sigma)$ *is convex. Let* $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ *be proper lower-semicontinuous. Assume that* f *and* g_σ *are bounded from below. Then,* $\forall \lambda > 0$, *the iterates* (y_k, z_k, x_k) *given by the iterative scheme* (14) *verify*

(i) $(F^{DR,2}_{\lambda,\sigma}(x_k))$ *is nonincreasing and converging.*

(ii) The residual $||y_k - z_k||$ *converges to* 0 *at rate* $\min_{k \leq K} ||y_k - z_k||^2 = \mathcal{O}(1/K)$

(iii) For any cluster point (y^*, z^*, x^*), y^* *and* z^* *coincides to a stationary point of* $F_{\lambda,\sigma}$.

(iv) Additionally suppose that f *and* g_σ *are respectively KL and semi-algebraic, then if the sequence* (y_k, z_k, x_k) *is bounded, the whole sequence converges, and* y_k *and* z_k *converge to the same stationary point of* $F_{\lambda,\sigma}$.

Remark 4.5. Theorem 4.4 requires $\operatorname{Im}(D_\sigma)$ to be convex. We underline that this assumption is difficult to verify in practice. On the other hand, Theorem 4.4 has two advantages with respect to PnP-PGD in Theorem 4.1 and PnP-DRS in Theorem 4.3. First, it encompasses nonconvex and non differentiable data-fidelity terms f. Next, the choice of the regularization parameter λ is no more constrained (see the discussion at the end of Section 4.1) and automatic parameter tuning methods (Wei et al., 2020) could be used.

5. Experiments

5.1. Proximal GS denoiser

In this section, we learn a gradient denoiser (4) $D_\sigma = \mathrm{Id} - \nabla g_\sigma$ that verifies the conditions of Proposition 3.1 and thus that can be written as a proximal mapping. We first need to choose a relevant parametrization for the function g_σ. Several options are explored in (Cohen et al., 2021). As in (Hurault et al., 2022), we design the regularization function g_σ as

$$g_\sigma(x) = \frac{1}{2}||x - N_\sigma(x)||^2, \tag{16}$$

where $N_\sigma : \mathbb{R}^n \to \mathbb{R}^n$ is a $\mathcal{C}^2$ neural network. This choice makes the function g_σ bounded from below, as required in Theorems 4.1, 4.3 and 4.4. Next, it allows to take benefit of existing efficient denoising architectures. Indeed, relation (16) leads to the following expression for the denoiser

$$\begin{aligned} D_\sigma(x) &= \mathrm{Id} - \nabla g_\sigma(x) \\ &= N_\sigma(x) + J^\top_{N_\sigma(x)}(x - N_\sigma(x)), \end{aligned} \tag{17}$$

where $J_{N_\sigma(x)}$ is the Jacobian of N_σ at point x. This corresponds to applying the neural network N_σ with an additive correction that makes the denoiser a conservative field.

Denoising Network Architecture Similar to (Hurault et al., 2022), we choose to parameterize N_σ with the architecture DRUNet (Zhang et al., 2021) (represented in Appendix E), a U-Net in which residual blocks are integrated. DRUNet takes the noise level σ as input, which is consistent with our formulation. In order to ensure continuous differentiability w.r.t. the input, we change RELU activations to Softplus, which is $\mathcal{C}^\infty$. We also limit the number of residual blocks to 2 at each scale to lower the computational burden.

Training details We first train the GS denoiser, in the same conditions as (Hurault et al., 2022), with the L^2 loss

$$\mathcal{L}(\sigma) = \mathbb{E}_{x \sim p, \xi_\sigma \sim \mathcal{N}(0,\sigma^2)} \left[||D_\sigma(x + \xi_\sigma) - x||^2 \right], \tag{18}$$

where p is the distribution of a database of clean images. However, the resulting denoiser does not verify

$\nabla g_\sigma = \mathrm{Id} - D_\sigma$ contractive *i.e.* with $L < 1$ Lipschitz gradient, as required by Proposition 3.1.

The link between the Lipschitz constant of N_σ and the one of ∇g_σ being difficult to establish, following (Pesquet et al., 2021), we enforce $L < 1$ by regularizing the training loss of D_σ with the spectral norm of the Hessian of g_σ that reads $\nabla^2 g_\sigma = J_{(\mathrm{Id} - D_\sigma)}$. More specifically, we fine-tune the previously trained GS denoiser with the following loss:

$$\mathcal{L}_S(\sigma) = \mathbb{E}_{x\sim p, \xi_\sigma \sim \mathcal{N}(0,\sigma^2)} \Big[||D_\sigma(x+\xi_\sigma) - x||^2 + \mu \max(\left|\left|J_{(\mathrm{Id}-D_\sigma)}(x+\xi_\sigma)\right|\right|_S, 1-\epsilon) \Big] \tag{19}$$

where $||.||_S$ denotes the spectral (or operator) norm. During training, it is estimated with 50 power iterations. In practice, we fine-tune during 10 epochs, with σ ranging in $[0, 25]$, using various penalization parameter μ and with $\epsilon = 0.1$.

Denoising results We evaluate the PSNR performance of the proposed Prox-DRUNet denoiser (17), trained with the loss (19) with different values of μ. In Table 1, we compare, for various noise levels σ, our model with DRUNet equipped with the same architecture N_σ as our prox-DRUNet (2 residual blocks and softplus activations) and trained with an L^2 loss as in (18). Next, we present the results obtained with GS-DRUNet (Hurault et al., 2022), which corresponds to the denoiser (17) trained with the loss (18). We also indicate the performance of the classical FFDNet (Zhang et al., 2018) and DnCNN (Zhang et al., 2017a) denoisers.

σ(./255)	5	10	15	20	25
FFDNet	39.95	35.81	33.53	31.99	30.84
DnCNN	39.80	35.82	33.55	32.02	30.87
DRUNet	40.19	36.10	33.85	32.34	31.21
GS-DRUNet	40.27	36.16	33.92	32.41	31.28
Prox-DRUNet ($\mu = 10^{-3}$)	40.12	35.93	33.60	32.01	30.82
Prox-DRUNet ($\mu = 10^{-2}$)	40.04	35.86	33.51	31.88	30.64

Table 1. Average denoising PSNR performance of our prox-denoiser and compared methods on 256×256 center-cropped images from the CBSD68 dataset (Martin et al., 2001), for various noise levels σ.

As can be seen in Table 1, the Lipschitz fine-tuning step inevitably worsens the denoising performance, especially for larger values of μ and higher noise levels. Nevertheless for $\mu = 10^{-2}$ prox-DRUNet shows comparable performance with FFDNet and DnCNN. The decrease in PSNR when constraining the Lipschitz constant of $I - D_\sigma$ is thus limited (compared to imposing nonexpansiveness of D_σ, see Appendix F).

Lipschitz constant In our experiments, the Lipschitz constant of $g_\sigma = \mathrm{Id} - D_\sigma$ is not hardly constrained to satisfy $L < 1$. This property is rather softly enforced by penalization with the loss function (19). We now investigate the potential gap between the theoretical assumption $L < 1$ and the practical considerations. We evaluate in Table 2 the maximum value of $\left|\left|J_{(\mathrm{Id}-D_\sigma)}(x)\right|\right|_S$ while denoising noisy images from the CBSD68 testing dataset. Here, the value $\left|\left|J_{(\mathrm{Id}-D_\sigma)}(x)\right|\right|_S$ is computed by running the power method until convergence. Table 2 first illustrates that GS-DRUNet ($\mu = 0$) does not check the $L < 1$ Lipschitz property, especially for very noisy images. Next, for a large enough penalization parameter ($\mu = 10^{-2}$), prox-DRUNet satisfies the $L < 1$ constraint, on the CBSD68 testing dataset and at all studied noise levels.

Tables 1 and 2 exhibit a clear trade-off, controlled by μ, between denoising performance and Lipschitz constant. Contrary to nonexpansive denoisers (see Appendix F), our constrained prox denoiser provides denoising of high quality while satisfying the Lipschitz constraint.

σ(./255)	0	5	10	15	20	25
GS-DRUNet ($\mu = 0$)	0.94	1.26	2.47	1.96	2.50	3.27
Prox-DRUNet ($\mu = 10^{-2}$)	0.87	0.92	0.95	0.99	0.96	0.96
Prox-DRUNet ($\mu = 10^{-3}$)	0.86	0.94	0.97	0.98	0.99	1.19

Table 2. Maximal value of $\left|\left|J_{(\mathrm{Id}-D_\sigma)}(x)\right|\right|_S$ obtained with proximal denoisers (17) on 256×256 center-cropped CBSD68 dataset, for various noise levels σ.

We precise here that after each PnP experiment conducted with Prox-DRUNet in the following section, we empirically verified that we still have $\left|\left|\nabla^2 g_\sigma(x_k)\right|\right|_S < 1$ on all the iterates x_k where D_σ was evaluated.

5.2. PnP restoration

In this section, we apply, with the proximal denoiser Prox-DRUNET, the PnP algorithms PnP-PGD (8), PnP-DRSdiff (12) (*diff* specifies that this PnP-DRS is dedicated to differentiable data-fidelity terms f) and PnP-DRS (14) on deblurring super-resolution and inpainting experiments. We seek an estimate x of a clean image $x^* \in \mathbb{R}^n$, from a degraded observation obtained as $y = Ax^* + \xi_\nu \in \mathbb{R}^m$, with A a $m \times n$ degradation matrix and ξ_ν a white Gaussian noise with zero mean and standard deviation ν. With this formulation the data-fidelity term takes the form $f(x) = \frac{1}{2\nu^2} ||Ax - y||^2$ and the Lipschitz constant of ∇f is $L_f = \frac{1}{\nu^2} \left|\left|A^T A\right|\right|_S$. Convergence of PnP-PGD (8) and PnP-DRSdiff (12) are guaranteed by Theorems 4.1 and 4.3 for $\lambda L_f < 1$ and, with Theorem 4.4, PnP-DRS (12) converges without condition on the value of λ.

We will use for evaluation Gaussian noise with 3 noise levels $\nu \in \{2.55, 7.65, 12.75\}/255$ *i.e.* $\nu \in \{0.01, 0.03, 0.05\}$. For each noise level, we propose default values for the parameters σ and λ that we keep for both deblurring and

super-resolution. These values are explicitly given in Appendix G.2. Note that for both PnP-PGD and PnP-DRSdiff, λ is set to its maximal possible value for theoretical convergence and σ is adjusted to the same value for both algorithms. Therefore PnP-PGD and PnP-DRSdiff target a stationary point of the same objective function $F_{\lambda,\sigma}$. The algorithm terminates when the relative difference between consecutive values of the objective function is less than $\epsilon = 10^{-8}$ or the number of iterations exceeds $K = 1000$.

The convergence Theorem 4.4 of Prox-PnP-DRS requires $L < 1/2$. As D_σ is trained to ensure $L < 1$, as suggested in Remark 4.2, we relax the denoising operation replacing D_σ by $D_\sigma^\alpha = \alpha D_\sigma + (1-\alpha)\,\mathrm{Id} = \mathrm{Id} - \alpha\nabla g_\sigma$, with $\alpha = 1/2$. To illustrate the relavance of PnP-DRS, we also provide in Appendix G.6, inpaiting experiments which involve a non differentiable data-fidelity term.

5.2.1. Deblurring

For image deblurring, the degradation operator $A = H$ is a convolution performed with circular boundary conditions. The proximal operator of f can be effciently calculated using the discrete Fourier transform.

We demonstrate the effectiveness of our method on a large variety of blur kernels (represented in Appendix G.1) and noise levels. As in (Zhang et al., 2017b; Pesquet et al., 2021; Zhang et al., 2021; Hurault et al., 2022), we use the 8 real-world camera shake kernels of Levin et al. (2009) as well as the 9×9 uniform kernel and the 25×25 Gaussian kernel with standard deviation 1.6 (as in (Romano et al., 2017)). The algorithms are initialized with the observed blurred image and robustness to initialization is discussed in Appendix G.4.

Figure 1 first illustrates, on the image 'starfish', that the three methods give consistent deblurring results with sharp edges. The convergence of the curves $\min_{0\leq i\leq k} ||x_{i+1} - x_i||^2$ (l) and $F_{\lambda,\sigma}(x_k)$ (i,j,k) empirically confirm the theoretical convergence results. Additional illustrations are provided in Appendix G.3.

Next we numerically evaluate in Table 3 the PSNR performance of our three PnP algorithms on CBSD68. We give comparisons with the deep state-of-the-art PnP methods IRCNN (Zhang et al., 2017b) and DPIR (Zhang et al., 2021) which both apply the PnP-HQS algorithm with decreasing stepsize but without convergence guarantees. We also provide comparisons with the GS-PnP-HQS method (Hurault et al., 2022) which corresponds to PnP-HQS with the denoiser GS-DRUNet from Table 1. We finally indicate the deblurring performance of *nonexp-PnP-PGD*, the PnP-PGD algorithm applied with the denoiser *nonexp-DRUNet* trained to be nonexpansive (see Appendix F).

Observe that, among Prox-PnP methods, Prox-PnP-DRS gives the best results. Compared to the two other algorithms, convergence is guaranteed whatever be the value of λ, which can thus be tuned to optimize performance. Prox-PnP-DRS then shows very competitive performance with respect to the state-of-the-art deep PnP methods IRCNN and DPIR. Moreover, contrary to the proposed algorithms, IRCNN and DPIR do not have any guarantee of convergence and we empirically show in Appendix G.5 that DPIR actually fails to converge in practice. Note also that, Prox-PnP-PGD and Prox-PnP-DRS, which share the same regularization parameter λ in the objective function, almost always converge towards the same local minima, with very similar convergence curves (see Figure 1).

Method	2.55	7.65	12.75
IRCNN	31.42	28.01	26.40
DPIR	31.93	28.30	26.82
GS-PnP-HQS	31.70	28.28	26.86
Prox-PnP-PGD	30.57	27.80	26.61
Prox-PnP-DRSdiff	30.57	27.78	26.61
Prox-PnP-DRS	31.54	28.07	26.60
nonexp-PnP-PGD	30.25	27.06	25.30

Table 3. PSNR (dB) of deblurring methods on CBSD68. PSNR are averaged over 10 blur kernels for various noise levels ν. *Prox-PnP* stands for the PnP algorithm applied with the proximal denoiser.

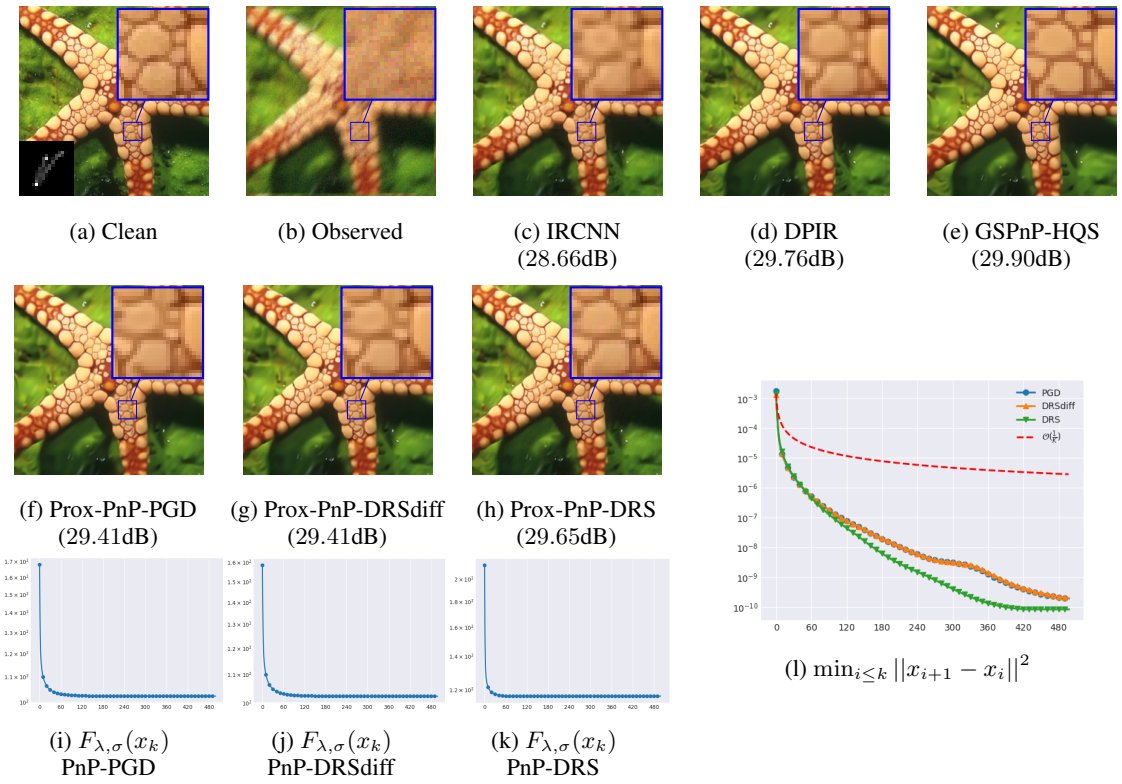
5.2.2. Super-resolution

For single image super-resolution (SR), the low-resolution image $y \in \mathbb{R}^m$ is obtained from the high-resolution one $x \in \mathbb{R}^n$ via $y = SHx + \xi_\nu$ where $H \in \mathbb{R}^{n\times n}$ is the convolution with anti-aliasing kernel. The matrix S is the standard s-fold downsampling matrix of size $m \times n$ and $n = s^2 \times m$. An efficient closed-form calculation of the proximal map for the data-fidelity term $f(x) = \frac{1}{2\nu^2}||SHx - y||^2$ is given by Zhao et al. (2016).

As in Zhang et al. (2021), we evaluate SR performance on 4 isotropic Gaussian blur kernels with different standard deviations (0.7, 1.2, 1.6 and 2.0) represented in Appendix G.1.

Method	$s=2$			$s=3$		
	2.55	7.65	12.75	2.55	7.65	12.75
IRCNN	26.97	25.86	25.45	25.60	24.72	24.38
DPIR	27.79	26.58	25.83	26.05	25.27	24.66
GS-PnP	27.88	26.81	26.01	25.97	25.35	24.74
Prox-PnP-PGD	27.44	26.57	25.82	25.75	25.20	24.63
Prox-PnP-DRSdiff	27.44	26.58	25.82	25.75	25.19	24.63
Prox-PnP-DRS	27.93	26.61	25.79	26.13	25.29	24.67
nonexp-PnP-PGD	27.13	26.20	25.40	23.83	24.57	24.01

Table 4. PSNR (dB) of super-resolution methods on CBSD68. PSNR averaged over 4 blur kernels for various scales s and noise levels ν.



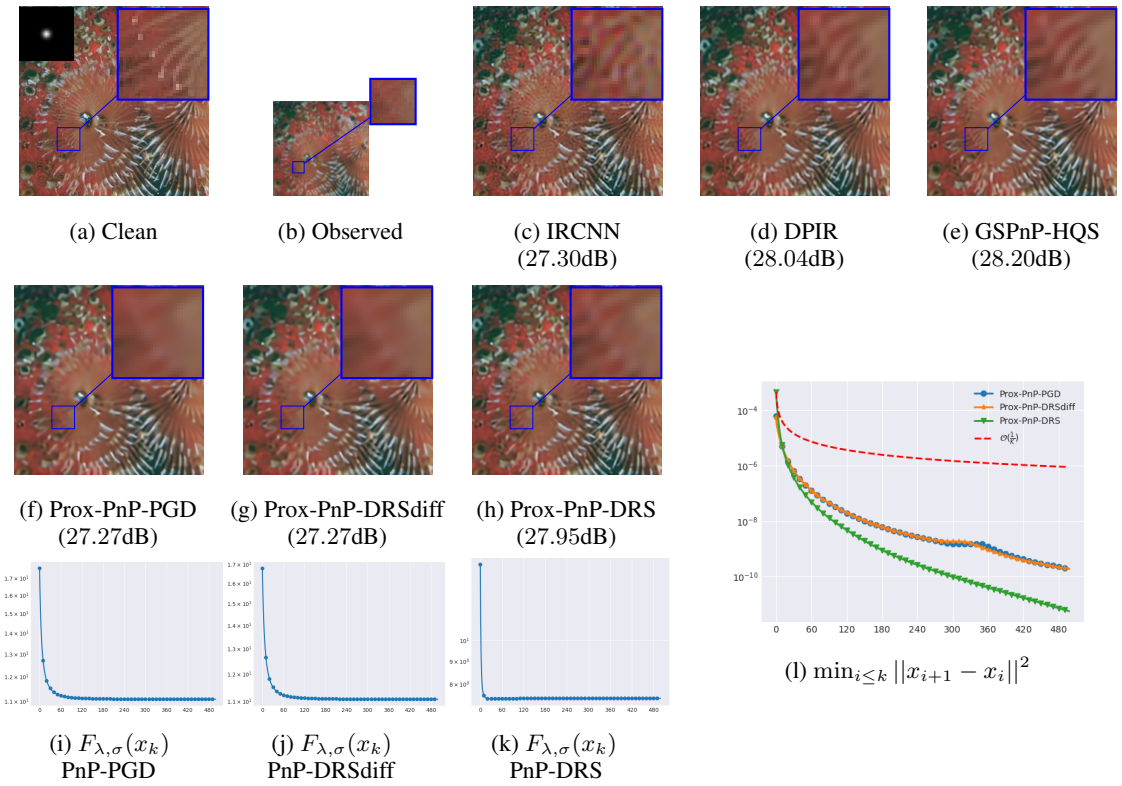
(a) Clean (b) Observed (c) IRCNN (28.66dB) (d) DPIR (29.76dB) (e) GSPnP-HQS (29.90dB)

(f) Prox-PnP-PGD (29.41dB) (g) Prox-PnP-DRSdiff (29.41dB) (h) Prox-PnP-DRS (29.65dB)

(l) $\min_{i \leq k} ||x_{i+1} - x_i||^2$

(i) $F_{\lambda,\sigma}(x_k)$ PnP-PGD (j) $F_{\lambda,\sigma}(x_k)$ PnP-DRSdiff (k) $F_{\lambda,\sigma}(x_k)$ PnP-DRS

Figure 1. Deblurring of "starfish" degraded with the indicated blur kernel and input noise level $\nu = 0.03$.



(a) Clean (b) Observed (c) IRCNN (27.30dB) (d) DPIR (28.04dB) (e) GSPnP-HQS (28.20dB)

(f) Prox-PnP-PGD (27.27dB) (g) Prox-PnP-DRSdiff (27.27dB) (h) Prox-PnP-DRS (27.95dB)

(l) $\min_{i \leq k} ||x_{i+1} - x_i||^2$

(i) $F_{\lambda,\sigma}(x_k)$ PnP-PGD (j) $F_{\lambda,\sigma}(x_k)$ PnP-DRSdiff (k) $F_{\lambda,\sigma}(x_k)$ PnP-DRS

Figure 2. Super-resolution on an image from CBSD68 downscaled with the indicated blur kernel, scale 2 and input noise level $\nu = 0.01$.

We consider downsampled images at scale $s = 2$ and $s = 3$. On a SR example (Figure 2), we observe the convergence of the iterates and of the function values for all three Prox-PnP algorithms. The methods are numerically compared in Table 4 against IRCNN, DPIR, GS-PnP and nonexp-PnP-PGD. Observe that, despite being trained to guarantee convergence, the three algorithms reach the performance of DPIR and that Prox-PnP-DRS realizes, with GS-PnP, the best super-resolution performance across the variety of scales and noise levels. Finally note that, as for deblurring, the PnP performance is significantly reduced when plugging a nonexpansive denoiser rather than a nonexpansive residual.

6. Conclusion

In this paper we provide new convergence results for PnP schemes able to produce state-of-the-art results for image restoration. We propose to learn a denoiser D_σ as the proximal operator of a nonconvex regularization function. Our proximal denoiser is not limited by nonexpansiveness and competes with unconstrained state-of-the-art denoisers. We show that, by plugging this denoiser, PnP-PGD, PnP-ADMM and PnP-DRS algorithms are guaranteed to converge to stationary points of an explicit functional. PnP-PGD enables to treat data-fidelity terms for which the proximal mapping cannot be computed in closed-form. The convergence studies for PnP-PGD/DRSdiff impose a constraint on the regularization parameter, on the contrary to PnP-DRS which thus yields better performance, at the cost of assuming convexity of $\mathrm{Im}(D_\sigma)$. Extensive quantitative experiments on ill-posed IR tasks, including data-fidelity terms that are not strongly convex, confirm the relevance of this approach, and show that it allows to precisely monitor convergence of the numerical schemes. To better understand the applicability of theoretical results, it is important to understand the geometry of the denoiser image $\mathrm{Im}(D_\sigma)$, a key issue that remains to be investigated.

Acknowledgements

This work was funded by the French ministry of research through a CDSN grant of ENS Paris-Saclay. This study has also been carried out with financial support from the French Research Agency through the PostProdLEAP and Mistic projects (ANR-19-CE23-0027-01 and ANR-19-CE40-005).

References

Ahmad, R., Bouman, C. A., Buzzard, G. T., Chan, S., Liu, S., Reehorst, E. T., and Schniter, P. Plug-and-play methods for magnetic resonance imaging: Using denoisers for image recovery. *IEEE signal processing magazine*, 37(1): 105–116, 2020.

Attouch, H., Bolte, J., Redont, P., and Soubeyran, A. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the kurdyka-łojasiewicz inequality. *Mathematics of operations research*, 35(2):438–457, 2010.

Attouch, H., Bolte, J., and Svaiter, B. F. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized gauss–seidel methods. *Mathematical Programming*, 137(1-2):91–129, 2013.

Beck, A. *First-order methods in optimization*. SIAM, 2017.

Buzzard, G. T., Chan, S. H., Sreehari, S., and Bouman, C. A. Plug-and-play unplugged: Optimization-free reconstruction using consensus equilibrium. *SIAM Journal on Imaging Sciences*, 11(3):2001–2020, 2018.

Chan, S. H. Performance analysis of plug-and-play admm: A graph signal processing perspective. *IEEE Transactions on Computational Imaging*, 5(2):274–286, 2019.

Cohen, R., Blau, Y., Freedman, D., and Rivlin, E. It has potential: Gradient-driven denoisers for convergent solutions to inverse problems. *Advances in Neural Information Processing Systems*, 34, 2021.

Combettes, P. L. and Pesquet, J.-C. Proximal splitting methods in signal processing. In *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, pp. 185–212. Springer, 2011.

Daubechies, I., Defrise, M., and De Mol, C. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 57(11):1413–1457, 2004.

Eckstein, J. and Fukushima, M. Some reformulations and applications of the alternating direction method of multipliers. In *Large scale optimization*, pp. 115–134. Springer, 1994.

Gavaskar, R. G. and Chaudhury, K. N. Plug-and-play ista converges with kernel denoisers. *IEEE Signal Processing Letters*, 27:610–614, 2020.

González, M., Almansa, A., and Tan, P. Solving inverse problems by joint posterior maximization with autoencoding prior. *arXiv preprint arXiv:2103.01648*, 2021.

Gribonval, R. Should penalized least squares regression be interpreted as maximum a posteriori estimation? *IEEE Transactions on Signal Processing*, 59(5):2405–2410, 2011.

Gribonval, R. and Nikolova, M. A characterization of proximity operators. *Journal of Mathematical Imaging and Vision*, 62(6):773–789, 2020.

Hurault, S., Leclaire, A., and Papadakis, N. Gradient step denoiser for convergent plug-and-play. In *International Conference on Learning Representations*, 2022.

Levin, A., Weiss, Y., Durand, F., and Freeman, W. T. Understanding and evaluating blind deconvolution algorithms. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1964–1971. IEEE, 2009.

Li, G. and Pong, T. K. Douglas–rachford splitting for nonconvex optimization with application to nonconvex feasibility problems. *Mathematical programming*, 159(1): 371–401, 2016.

Li, H. and Lin, Z. Accelerated proximal gradient methods for nonconvex programming. *Advances in neural information processing systems*, 28:379–387, 2015.

Liu, J., Asif, S., Wohlberg, B., and Kamilov, U. Recovery analysis for plug-and-play priors using the restricted eigenvalue condition. *Advances in Neural Information Processing Systems*, 34, 2021.

Lunz, S., Öktem, O., and Schönlieb, C.-B. Adversarial regularizers in inverse problems. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 8516–8525, 2018.

Mallat, S. *A Wavelet Tour of Signal Processing, The Sparse Way*. Academic Press, Elsevier, 3rd edition edition, 2009. ISBN 978-0-12-374370-1.

Martin, D., Fowlkes, C., Tal, D., and Malik, J. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proc. 8th Int'l Conf. Computer Vision*, volume 2, pp. 416–423, July 2001.

Meinhardt, T., Moller, M., Hazirbas, C., and Cremers, D. Learning proximal operators: Using denoising networks for regularizing inverse imaging problems. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1781–1790, 2017.

Moreau, J.-J. Proximité et dualité dans un espace hilbertien. *Bulletin de la Société Mathématique de France*, 93:273–299, 1965.

Nair, P., Gavaskar, R. G., and Chaudhury, K. N. Fixed-point and objective convergence of plug-and-play algorithms. *IEEE Transactions on Computational Imaging*, 2021.

Ochs, P., Chen, Y., Brox, T., and Pock, T. ipiano: Inertial proximal algorithm for nonconvex optimization. *SIAM Journal on Imaging Sciences*, 7(2):1388–1419, 2014.

Pesquet, J.-C., Repetti, A., Terris, M., and Wiaux, Y. Learning maximally monotone operators for image recovery. *SIAM Journal on Imaging Sciences*, 14(3):1206–1237, 2021.

Prost, J., Houdard, A., Almansa, A., and Papadakis, N. Learning local regularization for variational image restoration. *arXiv preprint arXiv:2102.06155*, 2021.

Reehorst, E. T. and Schniter, P. Regularization by denoising: Clarifications and new interpretations. *IEEE transactions on computational imaging*, 5(1):52–67, 2018.

Romano, Y., Elad, M., and Milanfar, P. The little engine that could: Regularization by denoising (red). *SIAM Journal on Imaging Sciences*, 10(4):1804–1844, 2017.

Ruderman, D. L. The statistics of natural images. *Network: computation in neural systems*, 5(4):517, 1994.

Rudin, L. I., Osher, S., and Fatemi, E. Nonlinear total variation based noise removal algorithms. *Phys. D*, 60: 259–268, 1992.

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

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

Sun, Y., Wohlberg, B., and Kamilov, U. S. An online plug-and-play algorithm for regularized image reconstruction. *IEEE Transactions on Computational Imaging*, 5(3):395–408, 2019.

Sun, Y., Wu, Z., Xu, X., Wohlberg, B., and Kamilov, U. S. Scalable plug-and-play admm with convergence guarantees. *IEEE Transactions on Computational Imaging*, 7: 849–863, 2021.

Terris, M., Repetti, A., Pesquet, J.-C., and Wiaux, Y. Building firmly nonexpansive convolutional neural networks. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8658–8662. IEEE, 2020.

Themelis, A. and Patrinos, P. Douglas–rachford splitting and ADMM for nonconvex optimization: Tight convergence results. *SIAM Journal on Optimization*, 30(1):149–181, 2020.

Venkatakrishnan, S. V., Bouman, C. A., and Wohlberg, B. Plug-and-play priors for model based reconstruction. In

2013 IEEE Global Conference on Signal and Information Processing, pp. 945–948. IEEE, 2013.

Wei, K., Aviles-Rivero, A., Liang, J., Fu, Y., Huang, H., and Schönlieb, C.-B. Tfpnp: Tuning-free plug-and-play proximal algorithm with applications to inverse imaging problems. *arXiv preprint arXiv:2012.05703*, 2020.

Yuan, X., Liu, Y., Suo, J., and Dai, Q. Plug-and-play algorithms for large-scale snapshot compressive imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Zhang, K., Zuo, W., Chen, Y., Meng, D., and Zhang, L. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Transactions on Image Processing*, 26(7):3142–3155, 2017a.

Zhang, K., Zuo, W., Gu, S., and Zhang, L. Learning deep cnn denoiser prior for image restoration. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3929–3938, 2017b.

Zhang, K., Zuo, W., and Zhang, L. Ffdnet: Toward a fast and flexible solution for cnn-based image denoising. *IEEE Transactions on Image Processing*, 27(9):4608–4622, 2018.

Zhang, K., Li, Y., Zuo, W., Zhang, L., Van Gool, L., and Timofte, R. Plug-and-play image restoration with deep denoiser prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

Zhao, N., Wei, Q., Basarab, A., Dobigeon, N., Kouamé, D., and Tourneret, J.-Y. Fast single image super-resolution using a new analytical solution for $\ell_2 - \ell_2$ problems. *IEEE Transactions on Image Processing*, 25(8):3683–3697, 2016.

Zoran, D. and Weiss, Y. From learning models of natural image patches to whole image restoration. In *2011 International Conference on Computer Vision*, pp. 479–486. IEEE, 2011.

A. Proof of Proposition 3.1

Our proof uses results from (Gribonval & Nikolova, 2020) and borrows proofs from (Gribonval, 2011). The latter shows a similar result with $D_\sigma = D_{mmse}$ the MMSE Gaussian noise denoiser. With Tweedie's formula, $D_{mmse} = \mathrm{Id} - \nabla g_{mmse}$ with $g_{mmse} = -\sigma^2 \log p * \mathcal{N}(0, \sigma^2 \mathrm{Id})$ the log-probability distribution of the noisy observation, and $g_{mmse} \in \mathcal{C}^\infty$. As g_σ does not write as $-\sigma^2 \log p * \mathcal{N}(0, \sigma^2 \mathrm{Id})$ for a prior p, the GS denoiser $D_\sigma = \mathrm{Id} - \nabla g_\sigma$ does not correspond to Tweedie's formula and is not a MMSE. The real MMSE estimator D_σ^* cannot be computed in closed-form and D_σ only approximates D_σ^* via MSE minimization. Then, the fact that D_σ^* is a prox (Gribonval, 2011) does not directly extend to D_σ. We show here that we can still express D_σ as the proximal mapping of an explicit *nonconvex* function.

Proof. (i) We first recall that in our setting (relations (3), (4) and (5)) , $D_\sigma = \mathrm{Id} - \nabla g_\sigma$ can be written as $D_\sigma = \nabla h_\sigma$. We show that if $\nabla g_\sigma = \mathrm{Id} - \nabla h_\sigma$ is L-Lipschitz with $L < 1$, then h_σ is $(1-L)$-strongly convex. For $x, y \in \mathcal{X}$, we have

$$\begin{aligned} \langle \nabla h_\sigma(x) - \nabla h_\sigma(y), x - y \rangle &= ||x-y||^2 - \langle \nabla g_\sigma(x) - \nabla g_\sigma(y), x - y \rangle \\ &\geq ||x-y||^2 - ||\nabla g_\sigma(x) - \nabla g_\sigma(y)||\, ||x-y|| \\ &\geq ||x-y||^2 - L\, ||x-y||^2 \\ &= (1-L)\, ||x-y||^2 , \end{aligned} \tag{20}$$

where the two inequalities respectively follow from Cauchy-Schwarz and from the Lipschitz continuity of ∇g_σ.

As h_σ is $\mathcal{C}^2$ and strongly convex, it follows that $\forall x \in \mathcal{X}$, the Jacobian $J_{D_\sigma}(x) = \nabla^2 h_\sigma(x)$ is positive definite.

(ii) Let us first show that D_σ is injective. Assume by contradiction that $\exists x, x' \in \mathcal{X}$ such that $x \neq x'$ and $D_\sigma(x) = D_\sigma(x')$. Let $v = (x' - x)/\,||x' - x||$. As $\mathcal{X}$ is convex, $\forall t \in [0, ||x' - x||]$, $x + tv \in \mathcal{X}$. The function $r : t \to \langle v, D_\sigma(x + tv) \rangle$ is $\mathcal{C}^1$ and satisfies $r(0) = r(||x' - x||)$. By Rolle's theorem, $\exists t_0 \in (0, ||x' - x||)$ such that $r'(t_0) = \langle v, J_{D_\sigma}(x + t_0 v)^T v \rangle = 0$, which is impossible because $J_{D_\sigma}(x + t_0 v)$ is positive definite. Therefore D_σ is injective and we can consider its inverse D_σ^{-1} on $\mathrm{Im}(D_\sigma)$. Also, the inverse function theorem ensures that $\mathrm{Im}(D_\sigma)$ is open in $\mathbb{R}^n$.

Let ϕ_σ defined as

$$\phi_\sigma(x) := \begin{cases} -\frac{1}{2} \left|\left| {D_\sigma}^{-1}(x) - x \right|\right|^2 + g_\sigma({D_\sigma}^{-1}(x))) & \text{if } x \in \mathrm{Im}(D_\sigma), \\ +\infty \quad \text{otherwise} \end{cases} \tag{21}$$

We now show that we also have $D_\sigma = \mathrm{Prox}_{\phi_\sigma}$.

For $y \in \mathcal{X}$, we thus look for the minimum of

$$\theta(x) = \phi_\sigma(x) + \frac{1}{2} ||y - x||^2 . \tag{22}$$

First, the definition of ϕ_σ ensures that $\mathrm{Prox}_{\phi_\sigma}$ takes its values in $\mathrm{Im}(D_\sigma)$. We let $x = D_\sigma(u)$ for $u \in \mathcal{X}$ and define

$$\begin{aligned} \Psi(u) = \theta(D_\sigma(u)) &= \phi_\sigma(D_\sigma(u)) + \frac{1}{2} ||y - D_\sigma(u)||^2 \\ &= -\frac{1}{2} ||u - D_\sigma(u)||^2 + g_\sigma(u) + \frac{1}{2} ||y - D_\sigma(u)||^2 \\ &= -\frac{1}{2} ||\nabla g_\sigma(u)||^2 + g_\sigma(u) + \frac{1}{2} ||y - D_\sigma(u)||^2 . \end{aligned} \tag{23}$$

The function Ψ is $\mathcal{C}^1$ in $\mathcal{X}$. As $D_\sigma(u) = u - \nabla g_\sigma(u)$, we get $J_{D_\sigma}(u) = \mathrm{Id} - \nabla^2 g_\sigma(u)$ (which is symmetric) and

$$\begin{aligned} \nabla \Psi(u) &= -\nabla^2 g_\sigma(u).\nabla g_\sigma(u) + \nabla g_\sigma(u) - J_{D_\sigma}(u).(y - D_\sigma(u)) \\ &= J_{D_\sigma}(u).(\nabla g_\sigma(u) - y + D_\sigma(u)) \\ &= J_{D_\sigma}(u).(u - y). \end{aligned} \tag{24}$$

Let $r(t) := \Psi(y + t(u - y))$ for $t \in [0, 1]$. Then,

$$\begin{aligned} r'(t) &= \langle \nabla\Psi(y + t(u - y)), u - y\rangle \\ &= t\langle J_{D_\sigma}(y + t(u - y)).(u - y), u - y\rangle, \end{aligned} \tag{25}$$

which is 0 if and only if $t = 0$ because $J_{D_\sigma}(y + t(u - y))$ is positive definite. Then r admits a unique stationary point at $t = 0$. This is a global minimum because $r'(t)$ has the sign of t. Therefore, Ψ admits a unique stationary point at $y+0.(u-y) = y$, which is a global minimum. By the definition of ϕ_σ, we can deduce that $\forall y$, $x \to \phi_\sigma(x)+\frac{1}{2}\left|\left|y - x\right|\right|^2$ admits a unique global minimum at $x = D_\sigma(y)$ *i.e.* $D_\sigma = \mathrm{Prox}_{\phi_\sigma}$.

(iii) For any $x \in \mathcal{X}$ we have

$$\begin{aligned} \phi_\sigma(x) &= \frac{1}{2}\left|\left|x - x\right|\right|^2 + \phi_\sigma(x) \\ &\geq \frac{1}{2}\left|\left|x - D_\sigma(x)\right|\right|^2 + \phi_\sigma(D_\sigma(x)) \\ &= g_\sigma(x) \end{aligned} \tag{26}$$

where the first inequality comes from the definition of the proximal operator $D_\sigma = \mathrm{Prox}_{\phi_\sigma}$ and the last equality is given by the definition (21) of ϕ_σ.

(iv) By the inverse function theorem, ϕ_σ is immediately $\mathcal{C}^k$ ($k \geq 1$) on $\mathrm{Im}(D_\sigma)$. We recall that $D_\sigma = \mathrm{Id} - \nabla g_\sigma$. From the definition of the proximal operator of the differentiable function ϕ_σ, that is single-valued and satisfies $D_\sigma = \mathrm{Prox}_{\phi_\sigma}$ (point(ii)), we have $\forall y \in \mathrm{Im}(D_\sigma)$

$$D_\sigma(y) = (\mathrm{Id} + \nabla\phi_\sigma)^{-1}(y) \quad \text{i.e.} \quad \nabla\phi_\sigma(y) = D_\sigma^{-1}(y) - y = \nabla g_\sigma(D_\sigma^{-1}(y)). \tag{27}$$

(v) We recall again that the denoiser is defined in (4) as $D_\sigma = \nabla h_\sigma$, with the function $h_\sigma : x \to \frac{1}{2}\left|\left|x\right|\right|^2 - g_\sigma(x)$. Let $x, y \in \mathrm{Im}(D_\sigma)$, there exists $u, v \in \mathbb{R}^n$ such that $x = D_\sigma(u)$ and $y = D_\sigma(v)$. Hence we have

$$\begin{aligned} \left|\left|\nabla\phi_\sigma(x) - \nabla\phi_\sigma(y)\right|\right| &= \left|\left|{D_\sigma}^{-1}(x) - {D_\sigma}^{-1}(y) - (x - y)\right|\right| \\ &= \left|\left|u - D_\sigma(u) - (v - D_\sigma(v))\right|\right| \\ &= \left|\left|\nabla g_\sigma(u) - \nabla g_\sigma(v)\right|\right| \\ &\leq L\left|\left|u - v\right|\right| \\ &\leq \frac{L}{1 - L}\left|\left|\nabla h_\sigma(u) - \nabla h_\sigma(v)\right|\right| \\ &= \frac{L}{1 - L}\left|\left|D_\sigma(u) - D_\sigma(v)\right|\right| \\ &= \frac{L}{1 - L}\left|\left|x - y\right|\right|, \end{aligned}$$

where the first and second inequalities respectively follow from the facts that ∇g_σ is L-Lipschitz and h_σ is $(1 - L)$-strongly convex (from point (i)).

□

B. Proof of Theorem 4.1

To prove this convergence theorem, we use the literature of convergence analysis (Attouch et al., 2013; Beck, 2017; Li & Lin, 2015) of the PGD algorithm in the nonconvex setting. The two first points follow exactly the same arguments as (Beck, 2017; Hurault et al., 2022). The third point requires an adaptation to handle the nonconvexity of ϕ_σ. Finally, the last point of the theorem directly follows from (Attouch et al., 2013).

Proof.

(i) We denote the proximal gradient fixed point operator $T_{\lambda,\sigma} = \operatorname{Prox}_{\phi_\sigma} \circ (\operatorname{Id} - \lambda \nabla f)$, the objective function $F_{\lambda,\sigma} = \lambda f + \phi_\sigma$ and we introduce

$$Q(x,y) = \lambda f(y) + \lambda \langle x - y, \nabla f(y) \rangle + \frac{1}{2} ||x-y||^2 + \phi_\sigma(x). \tag{28}$$

We have

$$Q(x,x) = F_{\lambda,\sigma}(x) \tag{29}$$

and

$$\begin{aligned} \arg\min_x Q(x,y) &= \arg\min_x \lambda \langle x - y, \nabla f(y) \rangle + \frac{1}{2} ||x-y||^2 + \phi_\sigma(x) \\ &= \arg\min_x \phi_\sigma(x) + \frac{1}{2} ||x - (y - \lambda \nabla f(y)||^2 \\ &= \operatorname{Prox}_{\phi_\sigma} \circ (\operatorname{Id} - \lambda \nabla f)(y) = T_{\lambda,\sigma}(y). \end{aligned} \tag{30}$$

The $\arg\min$ is unique by Proposition 3.1 and by definition of the $\arg\min$, $x_{k+1} = T_{\lambda,\sigma}(x_k)$ implies that

$$Q(x_{k+1}, x_k) \leq Q(x_k, x_k). \tag{31}$$

Moreover, with f being L_f-smooth, we have by the descent lemma, for any $t \leq \frac{1}{L_f}$ and any $x, y \in \mathbb{R}^n$,

$$f(x) \leq f(y) + \langle x - y, \nabla f(y) \rangle + \frac{1}{2t} ||x-y||^2 . \tag{32}$$

Hence for every $x, y \in \mathbb{R}^n$ and taking $t = \lambda < \frac{1}{L_f}$ in relation (32), we get from relation (28)

$$Q(x,y) \geq F_{\lambda,\sigma}(x). \tag{33}$$

Therefore, combining from (29), (31) and (33), we get at iteration k,

$$F_{\lambda,\sigma}(x_{k+1}) \leq Q(x_{k+1}, x_k) \leq Q(x_k, x_k) = F_{\lambda,\sigma}(x_k). \tag{34}$$

The sequence $(F_{\lambda,\sigma}(x_k))$ is thus non-increasing and lower-bounded by assumption. $(F_{\lambda,\sigma}(x_k))$ thus converges to a limit $F^*_{\lambda,\sigma}$.

(ii) Note that $Q(x_{k+1}, x_k) \leq Q(x_k, x_k)$ in (31) implies

$$\phi_\sigma(x_{k+1}) \leq \phi_\sigma(x_k) - \lambda \langle x_{k+1} - x_k, \nabla f(x_k) \rangle - \frac{1}{2} ||x_{k+1} - x_k||^2 . \tag{35}$$

Using again relation (32) with stepsize $t = \frac{1}{L_f}$, we get

$$\begin{aligned} F_{\lambda,\sigma}(x_{k+1}) &= \lambda f(x_{k+1}) + \phi_\sigma(x_{k+1}) \\ &\leq \phi_\sigma(x_k) - \lambda \langle x_{k+1} - x_k, \nabla f(x_k) \rangle - \frac{1}{2} ||x_{k+1} - x_k||^2 \\ &+ \lambda f(x_k) + \lambda \langle x_{k+1} - x_k, \nabla f(x_k) \rangle + \frac{L_f}{2} ||x_{k+1} - x_k||^2 \\ &= F_{\lambda,\sigma}(x_k) - \frac{1}{2}(1 - L_f) ||x_{k+1} - x_k||^2 . \end{aligned} \tag{36}$$

Summing over $k = 0, 1, ..., m$ gives

$$\begin{aligned} \sum_{k=0}^{m} ||x_{k+1} - x_k||^2 &\leq \frac{2}{1 - L_f} \left(F_{\lambda,\sigma}(x_0) - F_{\lambda,\sigma}(x_{m+1}) \right) \\ &\leq \frac{2}{1 - L_f} \left(F_{\lambda,\sigma}(x_0) - F^*_{\lambda,\sigma} \right). \end{aligned} \tag{37}$$

Therefore, $\lim_{k \to \infty} ||x_{k+1} - x_k|| = 0$ with the convergence rate $\gamma_k = \min_{0 \leq i \leq k} ||x_{i+1} - x_i||^2 \leq \frac{2}{k} \frac{F_{\lambda,\sigma}(x_0) - \lim F_{\lambda,\sigma}(x_k)}{1 - L_f}$

(iii) We begin by the two following lemmas characterizing the proximal gradient descent operator $T_{\lambda,\sigma}$.

Lemma B.1. *With the assumptions of Theorem 4.1 (in particular that* $\operatorname{Prox}_{\phi_\sigma} = (\operatorname{Id} + \partial\phi_\sigma)^{-1}$ *is single-valued), for* $x^* \in \mathbb{R}^n$, x^* *is a fixed point of the proximal gradient descent operator* $T_{\lambda,\sigma} = \operatorname{Prox}_{\phi_\sigma} \circ (\operatorname{Id} - \lambda\nabla f)$, i.e. $T_{\lambda,\sigma}(x^*) = x^*$, *if and only if* x^* *is a stationary point of* $F_{\lambda,\sigma}$, i.e. $-\lambda\nabla f(x^*) \in \partial\phi_\sigma(x^*)$.

Proof. By definition of the proximal operator, we have

$$\begin{aligned} T_{\lambda,\sigma}(x^*) = x^* &\Leftrightarrow x^* = \operatorname{Prox}_{\phi_\sigma} \circ (\operatorname{Id} - \lambda\nabla f)(x^*) \\ &\Leftrightarrow x^* = (\operatorname{Id} + \partial\phi_\sigma)^{-1}(x^* - \lambda\nabla f(x^*)) \\ &\Leftrightarrow -x^* + x^* - \lambda\nabla f(x^*) \in \partial\phi_\sigma(x^*) \\ &\Leftrightarrow -\lambda\nabla f(x^*) \in \partial\phi_\sigma(x^*). \end{aligned} \tag{38}$$

□

Lemma B.2. *With the assumptions of Theorem 4.1,* $T_{\lambda,\sigma}$ *is* $(1 + \lambda L_f)(1 + L)$ *Lipschitz-continuous.*

Proof. The fonction $\operatorname{Id} - \lambda\nabla f$ is $1 + \lambda L_f$-Lipschitz and $\operatorname{Prox}_{\phi_\sigma} = D_\sigma = \operatorname{Id} - \nabla g_\sigma$ is $1 + L$-Lipschitz. By composition, $T_{\lambda,\sigma}$ is $(1 + \lambda L_f)(1 + L)$-Lipschitz. □

We can now turn to the proof of (iii). Let x^* be a cluster point of $(x_k)_{k\geq 0}$. Then there exists a subsequence $(x_{k_j})_{j\geq 0}$ converging to x^*. We have $\forall j \geq 0$,

$$\begin{aligned} ||x^* - T_{\lambda,\sigma}(x^*)|| &\leq \left|\left|x^* - x_{k_j}\right|\right| + \left|\left|x_{k_j} - T_{\lambda,\sigma}(x_{k_j})\right|\right| + \left|\left|T_{\lambda,\sigma}(x_{k_j}) - T_{\lambda,\sigma}(x^*)\right|\right| \\ &\leq (1 + (1 + \lambda L_f)(1 + L)) \left|\left|x^* - x_{k_j}\right|\right| + \left|\left|x_{k_j} - T_{\lambda,\sigma}(x_{k_j})\right|\right| \text{ by Lemma B.2.} \end{aligned} \tag{39}$$

Using (ii), the right-hand side of the inequality tends to 0 as $j \to \infty$. Thus $||x^* - T_{\lambda,\sigma}(x^*)|| = 0$ and $x^* = T_{\lambda,\sigma}(x^*)$, which by Lemma B.1 means that x^* is a stationary point of $F_{\lambda,\sigma}$.

(iv) The convergence of the iterates follows Theorem 5.1 from (Attouch et al., 2013). To apply this result, we only need to ensure that ϕ_σ verifies the Kurdyka-Lojasiewicz (KL) inequality on $\operatorname{Im}(D_\sigma)$, since the stationary points x^* of $F_{\lambda,\sigma}$ are fixed points of $T_{\lambda,\sigma}$ and belong to $\operatorname{Im}(D_\sigma)$. We supposed g_σ to be a semi-algebraic mapping. Semi-algebraic mappings are KL. As mentioned in (Attouch et al., 2013), by the Tarski–Seidenberg theorem, the composition and inverse of semi-algebraic mappings are semi-algebraic mappings. Therefore, by the definition of ϕ_σ, we get that ϕ_σ is a semi-algebraic and thus KL on $\operatorname{Im}(D_\sigma)$.

□

C. Proofs of PnP-DRS convergence theorems

Theorems 4.3 and 4.4 come from the convergence results presented in Theorems 4.1, 4.3 and 4.4 from (Themelis & Patrinos, 2020). In their paper, the following minimization problem is considered

$$\arg\min_x F(x) = f_1(x) + f_2(x) \tag{40}$$

where f_1 is assumed $\mathcal{C}^1$ on $\mathbb{R}^n$, M_{f_1} semiconvex, with gradient L_{f_1}-Lipschitz.

We first define the notion of *semiconvexity* (also refered as *weak convexity* in the literature).

Definition C.1 (Semiconvex functions)**.** f proper lower semicontinuous is said to be M-semiconvex with $M \in \mathbb{R}$ if $x \to f(x) + \frac{M}{2}||x||^2$ is convex.

In this setting, the authors of (Themelis & Patrinos, 2020) show the convergence of DRS iterations

$$\begin{cases} y_{k+1} = \operatorname{Prox}_{\tau f_1}(x_k) \\ z_{k+1} = \operatorname{Prox}_{\tau f_2}(2y_{k+1} - x_k) \\ x_{k+1} = x_k + (z_{k+1} - y_{k+1}) \end{cases} \tag{41}$$

by using the Douglas–Rachford envelope,

$$F^{DR}_{\lambda,\sigma}(x) = f_1(y) + f_2(z) + \frac{1}{\tau}\langle y - x, y - z\rangle + \frac{1}{2\tau}\left|\left|y - z\right|\right|^2, \tag{42}$$

where y and z are obtained from x by the two first relations of (41).

Theorem C.2 (Convergence results of DRS from (Themelis & Patrinos, 2020))**.** *Let $f_1 : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ and $f_2 : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be proper, lower semicontinuous functions. Suppose that f_1 is M_{f_1}-semiconvex and $\mathcal{C}^1$ on $\mathbb{R}^n$, with gradient L_{f_1}-Lipschitz. Then, for a stepsize*

$$\tau < \min\left(\frac{1}{2M_{f_1}}, \frac{1}{L_{f_1}}\right), \tag{43}$$

we have

(a) ***(Descent lemma)*** $F^{DR}_{\lambda,\sigma}(x_k) - F^{DR}_{\lambda,\sigma}(x_{k+1}) \geq \frac{c}{(1+\tau L_{f_1})^2}\left|\left|x_k - x_{k+1}\right|\right|^2$ *and* $F^{DR}_{\lambda,\sigma}(x_k) - F^{DR}_{\lambda,\sigma}(x_{k+1}) \geq c\left|\left|y_k - y_{k+1}\right|\right|^2$ *for some constant $c > 0$ (Theorem 4.1).*

(b) $(y_k - z_k)$ *vanishes with rate* $\min_{k\leq K}\left|\left|y_k - z_k\right|\right| = \mathcal{O}(\frac{1}{\sqrt{K}})$ *(Theorem 4.3, (i)).*

(c) (y_k) *and* (z_k) *have the same cluster points, all of them being stationary for F (from equation (40)), and on which F has the same (finite) value, being the limit of $(F^{DR}_{\lambda,\sigma}(x_k))$ (Theorem 4.3, (ii))*

(d) If the sequence (y_k, z_k, x_k) is bounded, and f_1 and f_2 are KL, then the sequences (y_k) and (z_k) converge to (the same) stationary point of F (Theorem 4.4).

C.1. Proof of Theorem 4.3

If f is differentiable, we can directly apply Theorem C.2 with $f_1 = \lambda f$ and $f_2 = \phi_\sigma$. As f is convex, the condition (43) on the stepsize becomes $\lambda\tau < \frac{1}{L_f}$ *i.e.* $\lambda L_f < 1$ for $\tau = 1$.

C.2. Proof of Theorem 4.4

We seek to apply Theorem C.2 with the data-fidelity term $f_2 = \lambda f$ and the regularization $f_1 = \phi_\sigma$. With this setup, the DRS updates (41) correspond to the iterations (12) and the Douglas–Rachford envelope leads to equation (13). We have to adapt Theorem C.2, since the function ϕ_σ we consider is only smooth on the subspace $\mathrm{Im}(D_\sigma)$ and not on the whole space $\mathbb{R}^n$ as in (Themelis & Patrinos, 2020). This is the purpose of the four points (i-iv) of Theorem 4.3 that correspond to the four points (a-d) of Theorem C.2 and are shown below.

Proof.

(i) The proof of the descent lemma Theorem C.2 (a) uses successively the following properties:

1. The gradient descent lemma of a smooth and semiconvex function.
2. The characterisation of the proximal mapping of differentiable functions : for f differentiable $x = \mathrm{Prox}_f(y)$ iif $y = x + \nabla f(x)$
3. The strong monotonicity of the proximal operator of a smooth function.

We consider $\mathrm{Im}(D_\sigma)$ convex. We first show that ϕ_σ is semiconvex on $\mathrm{Im}(D_\sigma)$.

D_σ is $(L+1)$-Lipschitz. We directly apply Proposition 2 from (Gribonval & Nikolova, 2020) to get that there is $\tilde{\phi}_\sigma$ such that $\forall x \in \mathbb{R}^n, D_\sigma(x) \in \mathrm{Prox}_{\tilde{\phi}_\sigma}(x)$ and $x \to \tilde{\phi}_\sigma(x) + \frac{L}{2(L+1)}\left|\left|x\right|\right|^2$ is convex. As $\frac{L}{L+1} < 1, \forall x \in \mathbb{R}^n, y \to \tilde{\phi}_\sigma(y) + \frac{1}{2}\left|\left|x - y\right|\right|^2$ is strongly convex and $\mathrm{Prox}_{\tilde{\phi}_\sigma}(x)$ is single valued. Therefore

$$\forall x \in \mathbb{R}^n, D_\sigma(x) = \mathrm{Prox}_{\tilde{\phi}_\sigma}(x) = \mathrm{Prox}_{\phi_\sigma}(x). \tag{44}$$

For any $x \in \mathrm{Im}(D_\sigma)$ with $y \in \mathcal{X}$ such that $x = D_\sigma(y)$, the optimality condition of the proximal operator gives

$$\begin{cases} y - x \in \partial\phi_\sigma(x) \\ y - x \in \partial\tilde{\phi}_\sigma(x) \end{cases} \tag{45}$$

and then $\partial\phi_\sigma(x) \cap \partial\tilde{\phi}_\sigma(x) \neq \emptyset$. Moreover, $\mathrm{Im}(D_\sigma)$ is convex, thus connected and thus polygonally connected. By (Gribonval & Nikolova, 2020) Corollary 9, $\exists K \in \mathbb{R}$ such that $\phi_\sigma = \tilde{\phi}_\sigma + K$ on $\mathrm{Im}(D_\sigma)$. As $\tilde{\phi}_\sigma$ is $\frac{L}{L+1}$-semiconvex, we get that ϕ_σ is $\frac{L}{L+1}$-semiconvex on the convex set $\mathrm{Im}(D_\sigma)$.

The three points 1. 2. and 3. are then still verified on $\mathrm{Im}(D_\sigma)$ and Theorem C.2(a) is still valid. From (a) we get immediately that $F_{\lambda,\sigma}(x_k)$ is nonincreasing. With our hypothesis, $F_{\lambda,\sigma}(x_k)$ is bounded from below, thus converges.

As given in Proposition 3.1, $\nabla\phi_\sigma$ is $\frac{L}{1-L}$ Lipschitz on $\mathrm{Im}(D_\sigma)$ and ϕ_σ is $\frac{L}{L+1}$ semiconvex, with $0 < L < 1$. As we fix the stepsize to $\tau = 1$ in our PnP-DRS algorithm (12), the condition (43) on the stepsize in the statement of Theorem C.2 becomes

$$\begin{aligned} & 1 < \min\left(\frac{L+1}{2L}, \frac{1-L}{L}\right) \\ \Leftrightarrow\ & L < \frac{1}{2} \end{aligned} \tag{46}$$

(ii) As $F_{\lambda,\sigma}$ is bounded from below, the proof of Theorem C.2 (b) is still valid when ϕ_σ is smooth only on $\mathrm{Im}(D_\sigma)$ and we have $\min_{k \leq K} ||y_k - z_k|| = \mathcal{O}(\frac{1}{\sqrt{K}})$.

(iii) In our setting, the limit of y_k is not guaranteed to belong to $\mathrm{Im}(D_\sigma)$. Hence we can not follow the proof of Theorem C.2 (c). We rather rely on the convergence results from (Li & Pong, 2016), Theorem 1, and show that for any cluster point (y^*, z^*, x^*), we have that $y^* = z^*$ is a stationary point of $F_{\lambda,\sigma}$

Assuming that the whole sequence (y_k, z_k, x_k) has a cluster point (y^*, z^*, x^*), there exists a subsequence $(y_{k_j}, z_{k_j}, x_{k_j})$ converging to (y^*, z^*, x^*).

First, from (ii) we have that $y^* = z^*$. Let us now show that $y^* = D_\sigma(x^*)$. For $F^{DR}_{\lambda,\sigma}$ defined in (42), we have from (a) that

$$F^{DR}_{\lambda,\sigma}(x_k) - F^{DR}_{\lambda,\sigma}(x_{k+1}) \geq c\,||y_k - y_{k+1}||^2 \tag{47}$$

Summing from $k = 1$ to $k = N - 1 \geq 1$, we get

$$c\sum_{k=1}^{N-1} ||y_k - y_{k+1}||^2 \leq F^{DR}_{\lambda,\sigma}(x_1) - F^{DR}_{\lambda,\sigma}(x_N). \tag{48}$$

Taking the limit $j \to +\infty$ with $N = k_j$ and recalling that $F^{DR}_{\lambda,\sigma}$ is lower semicontinuous, we obtain

$$c\sum_{k=1}^{+\infty} ||y_k - y_{k+1}||^2 \leq F^{DR}_{\lambda,\sigma}(x_1) - F^{DR}_{\lambda,\sigma}(x_*) < +\infty, \tag{49}$$

where the second inequality comes from the fact that $\inf F^{DR}_{\lambda,\sigma} = \inf F_{\lambda,\sigma} > +\infty$ (Theorem 3.4 from (Themelis & Patrinos, 2020)).

Therefore $||y_k - y_{k+1}|| \to 0$ when $k \to +\infty$, in particular, $\lim_{j\to+\infty} y_{k_j+1} = y^*$. However, by continuity of D_σ, $y_{k_j+1} = D_\sigma(x_{k_j}) \to D_\sigma(x^*)$. Therefore,

$$y^* = D_\sigma(x^*). \tag{50}$$

We now show that $\lim_{j\to\infty} f(z_{k_j}) = f(z^*)$. As $z_k = \mathrm{Prox}_{\lambda f}(y_k)$ is defined as a minimizer, we have

$$\lambda f(z_k) + \frac{1}{2}\,||2y_k - x_{k-1} - z_k||^2 \leq \lambda f(z^*) + \frac{1}{2}\,||2y_k - x_{k-1} - z^*||^2\,. \tag{51}$$

Taking the limit along the convergent subsequence gives

$$\overline{\lim_{j\to\infty}} f(z_{k_j}) \leq f(z^*). \tag{52}$$

From the lower semicontinuity of f, we also have $\underline{\lim}_{j\to\infty} f(z_{k_j}) \geq f(z^*)$ and therefore $\lim_{j\to\infty} f(z_{k_j}) = f(z^*)$.

Following (Li & Pong, 2016), the subdifferential of a proper, nonconvex function f is defined as the limiting subdifferential

$$\partial f(x) := \left\{ v \in \mathbb{R}^n, \exists x_k, f(x_k) \to f(x), v_k \to v, \lim_{z\to x_k} \frac{f(z) - f(x_k) - \langle v_k, z - x_k\rangle}{||z - x_k||} \geq 0 \; \forall k \right\} \tag{53}$$

which verifies

$$\{v \in \mathbb{R}^n, \exists x_k, f(x_k) \to f(x), v_k \to v, v_k \in \partial f(x_k)\} \subseteq \partial f(x) \tag{54}$$

From the definition of the y and z updates in (12), we have

$$\begin{cases} 0 &= \nabla\phi_\sigma(y_{k+1}) + (y_{k+1} - x_k), \\ 0 &\in \lambda\partial f(z_{k+1}) + (z_{k+1} - 2y_{k+1} + x_k). \end{cases} \tag{55}$$

Hence summing both equations, $\forall k \geq 1$, we get

$$\nabla\phi_\sigma(y_k) + (z_k - y_k) \in -\lambda\partial f(z_k). \tag{56}$$

Passing to the limit $j \to +\infty$ in the converging subsequence and using the fact that $z_k - y_k \to 0$ and $y^* = z^*$, with the property (54) of the subdifferencial, we get

$$\nabla\phi_\sigma(y^*) \in -\lambda\partial f(y^*) \tag{57}$$

that is to say $y^* = z^*$ is a stationary point of $F_{\lambda,\sigma}$.

(iv) Point (d) is still valid when ϕ_σ is smooth only on $\mathrm{Im}(D_\sigma)$. □

D. On the boundedness of the generates sequences

In order to obtain convergence of the iterates, in the Theorems 4.1, 4.3 and 4.4 the generated sequences are assumed to be bounded. As we detail below, we observe in our experiments that boundedness is always verified. Nevertheless, we can make sure that this property is satisfied. A sufficient condition for the boundedness of the iterates (x_k) is the coercivity of the objective function $F_{\lambda,\sigma}$, that is, $\lim_{|x|\to\infty} F_{\lambda,\sigma}(x) = +\infty$ (because the non-increasing property gives $F_{\lambda,\sigma}(x_k) \leq F_{\lambda,\sigma}(x_0)$). In the case of Theorem 4.3, boundedness of (y_k) and (z_k) then follow respectively from the fact that $\mathrm{Prox}_{\lambda f}$ is Lipschitz ((Themelis & Patrinos, 2020), Proposition 2.3) and that $y_k - z_k \to 0$. In the case of Theorem 4.4, boundedness of (y_k) and (z_k) also follow respectively from the fact that D_σ is Lipschitz and that $y_k - z_k \to 0$.

Similar to (Hurault et al., 2022), we can constrain $F_{\lambda,\sigma}$ to be coercive by adding a penalization that vanishes in the neighborhood of $[0,1]^n$ and which tends to $+\infty$ at ∞. For that, we consider the ball $B(a,r)$ with $a = (1/2,\dots,1/2)$ and $r = \sqrt{n}$, and the penalization

$$p_{a,r}(x) = \rho(\|x - a\|^2 - r) \quad \text{with } \rho(t) = \begin{cases} t^3 \text{ if } t > 0 \\ 0 \text{ otherwise} \end{cases}. \tag{58}$$

It is clear that $p_{a,r}$ vanishes on $B(a,r)$ and tends to $+\infty$ at ∞. Besides, one can verify that $p_{a,r}$ is $\mathcal{C}^2$ (because ρ is). Therefore, we can add an extra term to g_σ, thus leading to:

$$\hat{g}_\sigma(x) = g_\sigma(x) + \gamma p_{a,r}(x) = \frac{1}{2}\,||x - N_\sigma(x)||^2 + \gamma p_{a,r}(x) \tag{59}$$

and $\gamma > 0$ controls the penalization strength.

The denoising operation becomes

$$\hat{D}_\sigma(x) = (\mathrm{Id} - \nabla\hat{g}_\sigma)(x) = D_\sigma(x) - \gamma\nabla p_{a,r}(x). \tag{60}$$

In practice we observe that all the iterates always remain in $B(a,r)$ and that the penalization $p_{a,r}(x)$ is never activated.

E. DRUNet architecture

The architecture of the DRUNet *light* denoiser of ((Zhang et al., 2021)) is given in Figure 3.

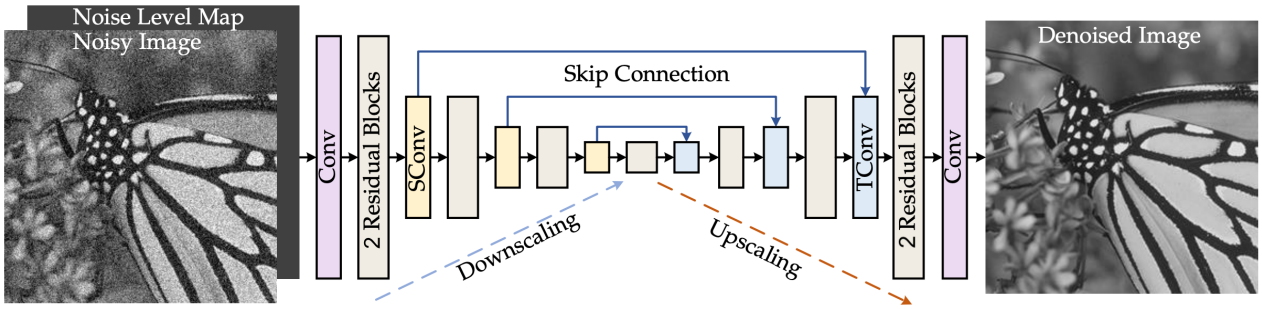


Figure 3. Architecture of the DRUNet *light* denoiser ((Zhang et al., 2021)) used to parameterize N_σ.

F. On the performance of nonexpansive denoisers

We provide in this appendix denoising performance comparisons with a nonexpansive denoiser. To do so, we use the strategy exposed in Section 5.1 to train DRUNet directly to be nonexpansive. The pretrained DRUNet denoiser N_σ is fine-tuned to denoise while being 1-Lipschitz with the following loss:

$$\mathbb{E}_{x\sim p,\xi_\sigma\sim\mathcal{N}(0,\sigma^2)}\left[||N_\sigma(x+\xi_\sigma)-x||^2+\mu\max(||J_{N_\sigma}(x+\xi_\sigma)||_S,1-\epsilon)\right] \tag{61}$$

with $\epsilon=0.1$ and different values of μ.

As in Section 5.1, we analyse the denoising performance of the resulting denoiser nonexp-DRUNet (Table 5) as well as the maximal value of the spectral norm $||J_{N_\sigma}(x)||_S$, on the CBSD68 testset (Table 6). Table 6 illustrates that the penalization parameter has to be set to $\mu=10^{-2}$ to make the Lipschitz constant lower than 1. On the other hand, as can be observed in Table 5, such setting severely degrades denoising performances.

From this observation, we argue that is less harmful for the denoising performance to impose nonexpansiveness on the denoiser residual $\mathrm{Id}-D_\sigma$ than on the denoiser itself.

$\sigma(./255)$	5	10	15	20	25
GS-DRUNet	40.27	36.16	33.92	32.41	31.28
prox-DRUNet ($\mu=10^{-2}$)	40.04	35.86	33.51	31.88	30.64
prox-DRUNet ($\mu=10^{-3}$)	40.12	35.93	33.60	32.01	30.82
nonexp-DRUNet ($\mu=10^{-2}$)	34.92	32.90	31.42	30.30	29.42
nonexp-DRUNet ($\mu=10^{-3}$)	39.71	35.71	33.50	32.00	30.89

Table 5. Average PSNR denoising performance of our prox-denoiser and compared methods on 256×256 center-cropped images from the CBSD68 dataset, for various noise levels σ.

$\sigma(./255)$	0	5	10	15	20	25
DRUNet ($\mu=0$)	1.13	1.73	2.36	2.67	2.76	3.22
nonexp-DRUNet ($\mu=10^{-2}$)	0.97	0.98	0.98	0.98	0.98	0.98
nonexp-DRUNet ($\mu=10^{-3}$)	1.06	1.07	1.07	1.10	1.13	1.20

Table 6. Maximal value of $\left|\left|J_{D_\sigma(x)}\right|\right|_S$ obtained with contractive denoisers $D_\sigma=N_\sigma$ on 256×256 center-cropped CBSD68 dataset, for various noise levels σ.

G. Experiments

G.1. Blur and anti-aliasing kernels

We show in Figures 4 and 5 the kernels respectively used for the evaluation of the deblurring and SR methods.

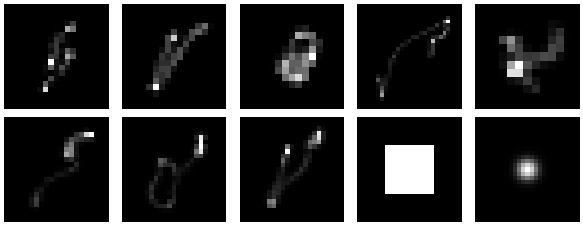


Figure 4. The 10 blur kernels used for deblurring evaluation.

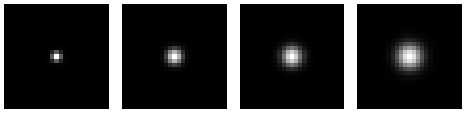


Figure 5. The 4 blur kernels used for super-resolution evaluation.

G.2. Regularization parameters

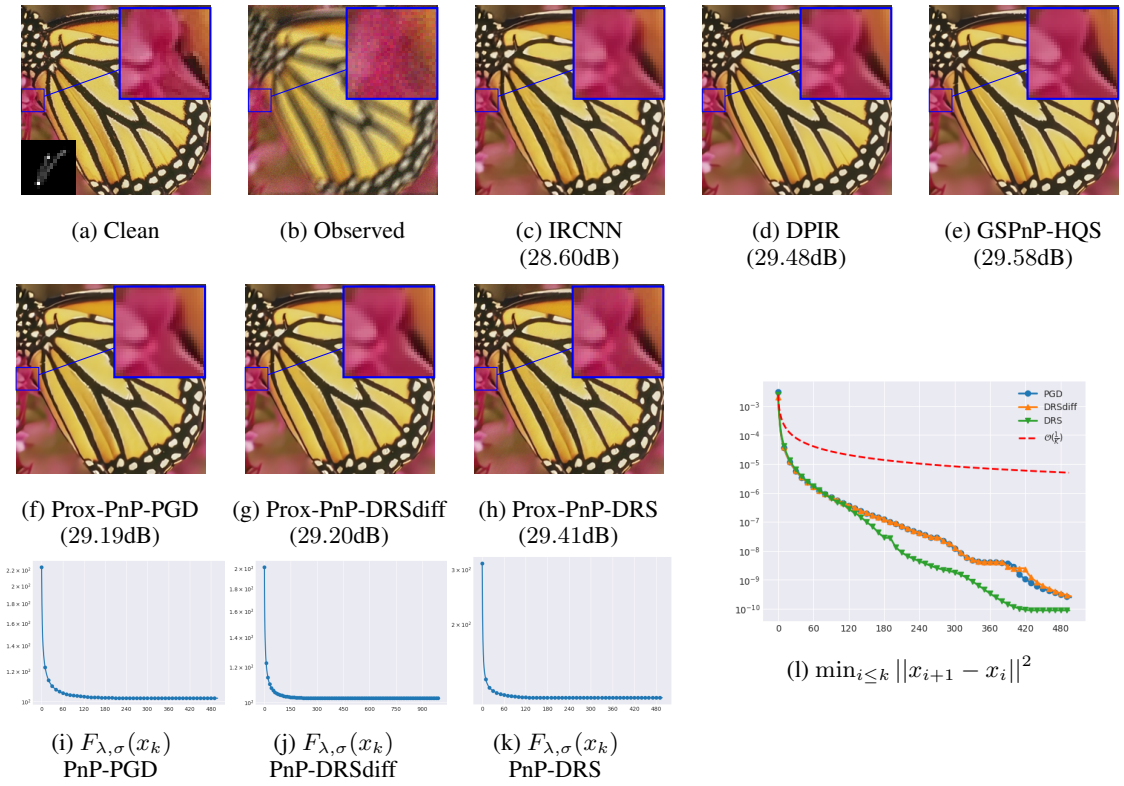
We give in Table 7 the value of the parameters to obtain the results presented in Section 5.2. Note that the same parameters are taken for both deblurring and super-resolution, and for all the variety of considered kernels and down-sampling scales. They are just scaled with respect to the noise level ν in the input image. We keep the same parameters for PnP-PGD and PnP-DRSdiff. Both algorithms thus target a stationary points of the same objective function $F_{\lambda,\sigma}$. The blur kernels are normalized so that the Lipschitz constant $L_f = \frac{1}{2\nu^2}\left|\left|H^T H\right|\right|_S$ of ∇f is equal to $\frac{1}{2\nu^2}$. For PnP-PGD and PnP-DRSdiff, λ is set as close as possible to its maximal possible value $\frac{1}{L_f} = \nu^2$ required for convergence in Theorems 4.1 and 4.3.

$\nu(./255)$		2.55	7.65	12.75
PGD & DRSdiff	λ/ν^2	0.99	0.99	0.99
	σ/ν	0.75	0.5	0.5
DRS	λ/ν^2	5	1.5	0.75
	σ/ν	2	1	0.5

Table 7. Choice of parameters (λ, σ) for both debluring and super-resolution experiments of Section 5.2. These parameters are only scaled with respect to the input noise level ν

G.3. Additional deblurring and super-resolution experiments

We complement the numerical analysis with additional visual results and convergence curves. Figures 6 and 7 respectively give the deblurring and super-resolution results on the two images "buttefly" and "leaves". We also provide visual and PSNR comparison with competitive PnP methods. Note that on both experiments, all Prox-PnP algorithm converge towards a visually coherent output. The Prox-PnP-DRS reaches or outperforms the performance of the DPIR algorithm and restores sharp images.



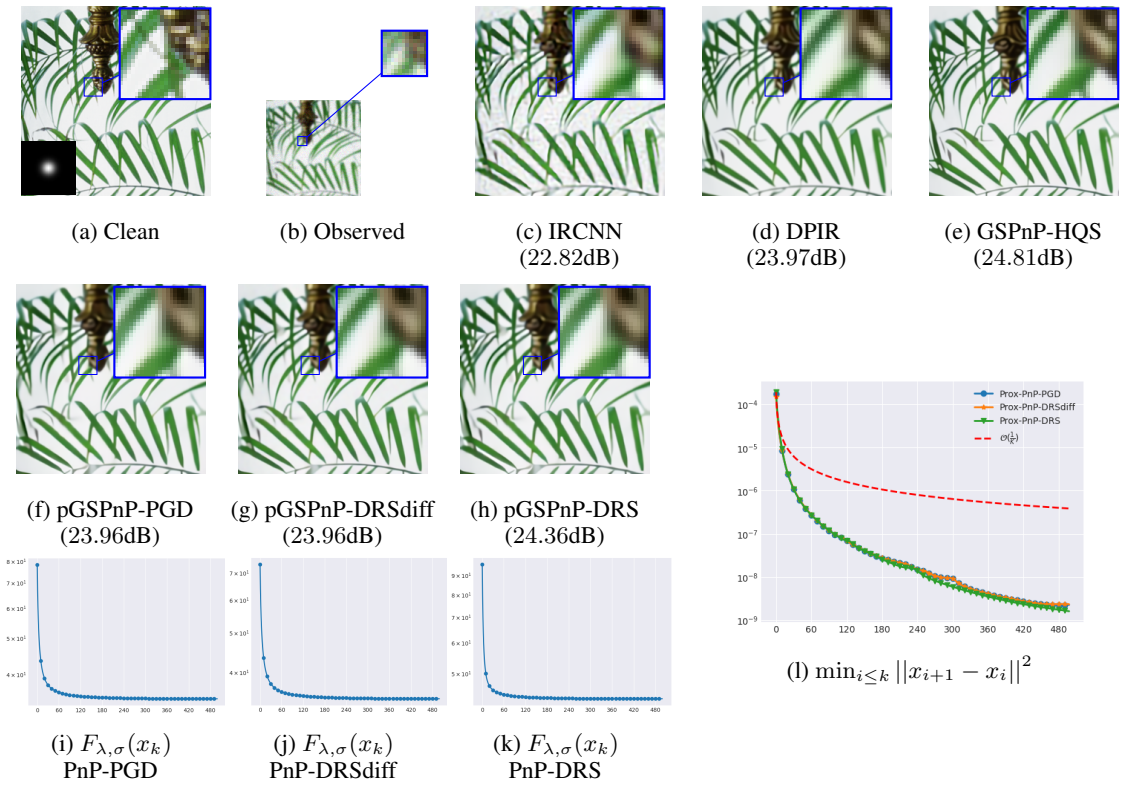
(a) Clean (b) Observed (c) IRCNN (28.60dB) (d) DPIR (29.48dB) (e) GSPnP-HQS (29.58dB)

(f) Prox-PnP-PGD (29.19dB) (g) Prox-PnP-DRSdiff (29.20dB) (h) Prox-PnP-DRS (29.41dB)

(i) $F_{\lambda,\sigma}(x_k)$ PnP-PGD (j) $F_{\lambda,\sigma}(x_k)$ PnP-DRSdiff (k) $F_{\lambda,\sigma}(x_k)$ PnP-DRS

(l) $\min_{i\leq k} ||x_{i+1} - x_i||^2$

Figure 6. Deblurring with various methods of "butterfly" degraded with the indicated blur kernel and input noise level $\nu = 0.03$.



(a) Clean (b) Observed (c) IRCNN (22.82dB) (d) DPIR (23.97dB) (e) GSPnP-HQS (24.81dB)

(f) pGSPnP-PGD (23.96dB) (g) pGSPnP-DRSdiff (23.96dB) (h) pGSPnP-DRS (24.36dB)

(i) $F_{\lambda,\sigma}(x_k)$ PnP-PGD (j) $F_{\lambda,\sigma}(x_k)$ PnP-DRSdiff (k) $F_{\lambda,\sigma}(x_k)$ PnP-DRS

(l) $\min_{i\leq k} ||x_{i+1} - x_i||^2$

Figure 7. Super-resolution with various methods of "leaves" degraded with the indicated blur kernel, scale 2 and input noise level $\nu = 0.03$.

G.4. Robustness to initialization

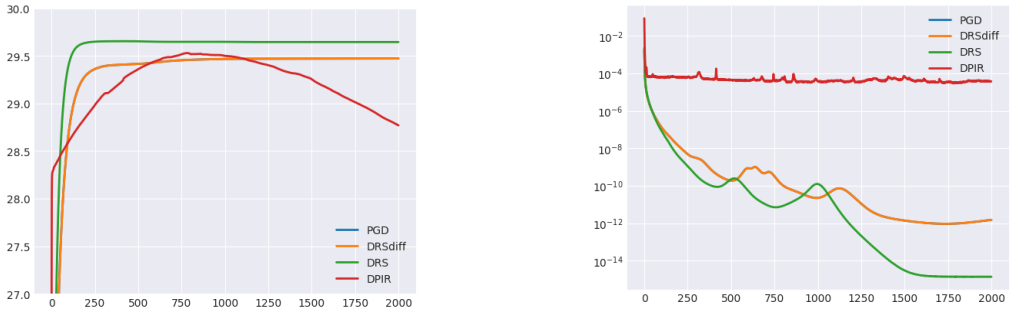
Table 8 shows the behavior of the three Prox-PnP methods initialized with the degraded image y + random Gaussian noise. The methods are robust for small noise perturbations and converge to different local minima for stronger noise. PnP-DRS remains consistent for stronger noise as it is run with a larger λ value, which tends to "convexify" the objective function.

σ_{init} $(./255)$	PGD	DRSdiff	DRS
0	29.41	29.41	29.65
10	29.41	29.41	29.65
20	27.34	25.92	29.65
40	12.52	12.54	29.63
60	12.27	12.28	18.65

Table 8. PSNR after initialization with $x_0 = y + n$ with $n \sim \mathcal{N}(0, \sigma_{init}^2 \, \mathrm{Id})$.

G.5. Convergence comparison with DPIR

DPIR (Zhang et al., 2021) performs well in PSNR when run on very few iterations but we show Figure 8 below that it actually does not converge asymptotically, contrary to our PnP methods. Instead of being runned on 8 iterations, DPIR is here runned on 2000 iterations. We observe that the both the PSNR and the norm of the residual do not converge in practice. Such a divergence strongly limits DPIR outputs interpretation and numerical control.



(a) PSNR(x_k) (b) $||x_{k+1} - x_k||^2$

Figure 8. Evolution, for DPIR and our PnP algorithms, on 2000 iterations of the PSNR and of the squared norm of the residual when deblurring the blurred "starfish" image from Figure 1.

G.6. Inpainting experiment

We finally apply PnP-DRS to image inpainting, with the degradation model $y = Ax$ where A is a diagonal matrix with values in $\{0, 1\}$. In this context, the data-fidelity term is the indicator function of the set $\mathcal{A} = \{x \mid Ax = y\}$: $f(x) = \imath_{\mathcal{A}}(x)$ which, by definition, equals 0 on $\mathcal{A}$ and $+\infty$ elsewhere. The convergence of PnP-DRS for such a non-differentiable data-fidelity term is ensured by Theorem 4.4. In our experiments, the diagonal of A is filled with Bernoulli random variables with parameter $p = 0.5$. We run our algorithm with $\sigma = 10/255$. As can be observed in Figure 9, PnP-DRS restores the input images with high accuracy. Furthermore, convergence of the residual at rate $\mathcal{O}(\frac{1}{k})$ is empirically confirmed.

In our experiments, the diagonal of A is filled with Bernoulli random variables with parameter $p = 0.5$. We run our PnP algorithm with $\lambda = 2$ and $\sigma = 15/255$. The algorithm is initialized with 10 iterations of the algorithm at larger noise level $\sigma = 50/255$ andterminates when the number of iterations exceeds $K = 200$.

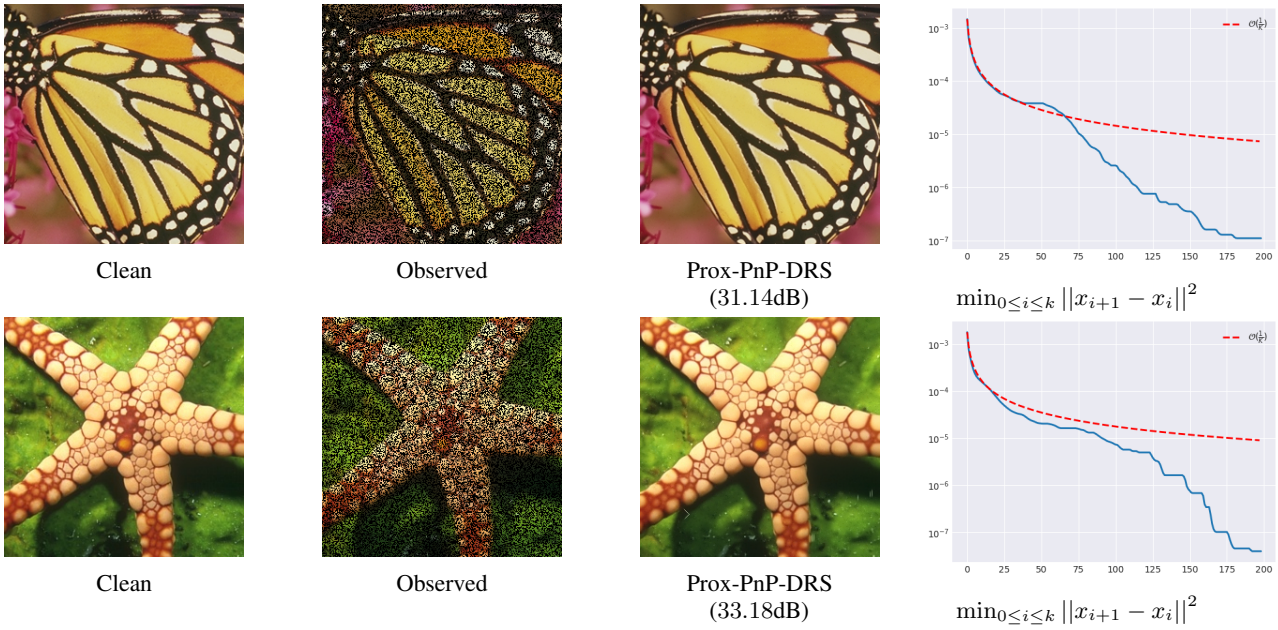


Figure 9. Inpainting results with pixels randomly masked with probability $p = 0.5$. In the last column, we show the evolution of along the iterations.