CONVERGENT PLUG-AND-PLAY IMAGE RESTORATION WITH ANNEALED NOISE LEVELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Plug-and-Play (PnP) methods solve imaging inverse problems by incorporating deep denoisers into iterative optimization algorithms. Although practical implementations often decrease the denoiser noise level σ along iterations, most existing convergence analyses assume a fixed denoiser. In this work, we establish convergence guarantees for a broad family of Plug-and-Play algorithms with annealed noise level, spanning deterministic methods (RED–GD and PnP–PGD) and stochastic methods (SNORE, equivariant RED, and a variant of PnP–Flow). For each method, we identify an explicit, nonconvex objective associated with the terminal denoising level and prove asymptotic stationarity of the iterates with respect to this objective. Our analysis does not prescribe any decay rate for the noise schedule, and our assumptions cover both learned gradient-step denoisers and exact MMSE denoisers. Overall, our theoretical results bridge the gap between existing PnP convergence theory and the decreasing-denoising practices used by state-of-the-art image restoration methods. We empirically demonstrate the benefits of such schedules and illustrate the predicted convergence behavior on several imaging inverse problems, including inpainting, super-resolution, demosaicing and tomography.

1 INTRODUCTION

Imaging inverse problems aim to recover an unknown image from incomplete or corrupted measurements. These tasks are typically ill-posed, motivating the use of prior information through regularization. A standard variational approach solves

$$\min_{x\in\mathbb{R}^d} f(x) + \lambda g(x), \tag{1}$$

where f measures consistency with the observations, g encodes an image prior, and $\lambda > 0$ balances the two terms. In a Bayesian formulation, choosing the negative log-likelihood $f = -\log p(y|x)$ and the negative log-prior $g = -\log p$ yields maximum a posteriori (MAP) estimation. The central challenge is to design a prior that captures the complexity of natural images while retaining an optimization problem solvable with a converging algorithm. Plug-and-Play (PnP) methods (Venkatakrishnan et al., 2013) replace the proximal operator of the regularizer in such optimization algorithms by a Gaussian denoiser D_σ, trained at noise level σ. Regularization by Denoising (RED) (Romano et al., 2017) instead uses the residual $x - D_\sigma(x)$ in place of the regularization gradient. Both PnP and RED algorithms alternate between a data-consistency update and a denoising operation, allowing the same pretrained model to serve as a prior across different inverse problems; see Hurault (2023) for a review. More recently, the use of deep denoisers as learned priors has been further popularized by diffusion models for sampling instead of optimization.

The parameter σ of the plugged denoiser acts as a regularization parameter and changes the fixed point of the plug-and-play algorithm. In practice, it is well established that progressively decreasing the denoiser noise level σ over the iterations can substantially improve restoration quality (Zhang et al., 2017; Wei et al., 2020; Zhang et al., 2022). A large σ helps recover coarse structures early in the iterations, while a smaller σ allows finer details to emerge later. By contrast, a fixed noise level imposes a trade-off, as illustrated in Figure 1: a small σ may fail to recover large structures, whereas a large σ may oversmooth fine details. Decreasing σ over the iterations combines these two regimes and enables a more accurate multiscale reconstruction. This continuation scheme can also ease optimization: at large σ, the associated objective is smoother and may be easier to

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



Figure 1: Inpainting of a 64×64 hole in a 128×128 face image using RED gradient descent with the same denoiser (GS–DiffUNet; see Section 4). With a constant noise level, a small σ leaves the initial flat patch, while larger values fill the hole at the cost of missing fine details. Decreasing geometrically σ_k along iterates, from a large initial noise level to a small terminal one, produces a sharp face.

optimize; decreasing σ then gradually restores the finer structure of the objective at the terminal noise level (Pesme et al., 2025). Noise annealing is also central to diffusion sampling models (Song et al., 2021; Dhariwal & Nichol, 2021), whose denoisers are trained over a wide range of noise levels. Recent PnP methods leverage these large denoisers with decreasing noise schedules to achieve state-of-the-art restoration performance (Mardani et al., 2024; Martin et al., 2025; Park et al., 2026).

Despite its practical success, most existing convergence analyses still assume a fixed denoiser throughout the iterations, even when decreasing noise schedules are used in experiments. These analyses typically provide fixed-point guarantees under suitable Lipschitz assumptions on the plugged denoiser (Chan et al., 2017; Ryu et al., 2019; Sun et al., 2019; 2021; Kamilov et al., 2023) or optimization guarantees for structured denoisers (Xu et al., 2020; Hurault et al., 2022a;b; 2023). Only a few existing convergence analyses handle varying the denoising level. Among them, analyses of annealed PnP–ADMM (Chan et al., 2017; Hou et al., 2022; Shrestha & Fu, 2026) provide fixed-point guarantees, but under restrictive assumptions on the plugged denoiser and without identifying an underlying variational objective. In a variational setting, Pesme et al. (2025) anneal MMSE denoisers in an inner loop to recover MAP denoising steps, but their strategy does not address single-loop PnP algorithms. Finally, Martin et al. (2025) establish convergence of PnP–Flow, a stochastic variant of PnP–PGD, under summable stepsizes, without characterizing the limit.

In this work, we analyze the convergence of a broad class of deterministic and stochastic plug-and-play schemes in which the denoiser noise level monotonically decreases. We prove, for all algorithms, asymptotic stationarity for an explicit variational objective. The deterministic methods are RED gradient descent (RED–GD) and PnP proximal gradient descent (PnP–PGD). The stochastic ones are SNORE (Renaud et al., 2024), Equivariant RED (Terris et al., 2024; Renaud et al., 2025), and a decoupled variant of SNOPnP (Renaud et al., 2026) and PnP–Flow (Martin et al., 2025).

To retain a variational interpretation, we consider denoisers that are exact gradient or proximal maps. This allows us to interpret PnP and RED methods as real optimization algorithms for minimizing explicit objectives of the form (1). The minimum mean square error (MMSE) Gaussian denoiser provides a canonical example: by Tweedie's formula (Efron, 2011) it is conservative and it is also a proximal operator (Gribonval, 2011). Although deep denoisers are trained to approximate the MMSE denoiser, they are not conservative in practice (Reehorst & Schniter, 2018). Gradient-step denoisers re-impose the conservative property by parametrizing the denoiser as $D_\sigma = \mathrm{Id} - \nabla g_\sigma$, where g_σ is a learned scalar potential (Hurault et al., 2022a). Under an additional contractivity condition on $\mathrm{Id} - D_\sigma$, the denoiser also admits a proximal representation (Hurault et al., 2022b).

Building on this variational structure, we establish convergence guarantees for all annealed algorithms when the noise level σ_k is nonincreasing and converges to a floor $\epsilon > 0$. Under algorithm-specific stepsize assumptions, the iterates become asymptotically stationary for an explicit objective associated with the terminal noise level ϵ. Our stationarity guarantees require neither convexity of the data term or denoiser potential nor a prescribed annealing rate: any monotone schedule $\sigma_k \downarrow \epsilon > 0$ suffices. The convergence assumptions are verified for both exact MMSE denoisers and state-of-the-art learned conservative denoisers. Experiments on inpainting, demosaicing, super-resolution, and tomography, using denoisers finetuned as gradient-step or proximal denoisers, support the predicted convergence behavior and show consistent gains from annealing over constant-noise baselines.

2 VARIOUS PNP AND RED ALGORITHMS

In this section, we review the different algorithms that will be theoretically analyzed in section 3. There are two families of Plug-and-Play methods, which we call RED (Romano et al., 2017) and PnP (Venkatakrishnan et al., 2013), that are both built by incorporating a Gaussian denoiser D_σ into iterative optimization algorithms. For each family, we study both deterministic and stochastic variants, when the noise level σ of the denoiser decreases along the iterations.

RED algorithms (Romano et al., 2017). They are built by replacing the gradient of the prior term $\nabla g(x)$ with a denoising residual $x - D_\sigma(x)$ in gradient-based optimization schemes for solving (1). Applying this substitution to gradient descent, with iteration-dependent noise levels σ_k yields:

$$\text{(RED–GD)} \qquad x_{k+1} = x_k - \gamma\Big[\nabla f(x_k) + \lambda\big(x_k - D_{\sigma_k}(x_k)\big)\Big]. \tag{2}$$

Renaud et al. (2024) proposed SNORE, a stochastic version of this algorithm that adds Gaussian noise at each iteration, so that the denoiser is applied to images carrying the level of noise it was trained for. The authors study its convergence for constant σ only. With a varying σ_k, the update is:

$$\text{(SNORE)} \qquad x_{k+1} = x_k - \gamma_k\Big[\nabla f(x_k) + \lambda\big(x_k - D_{\sigma_k}(x_k + \sigma_k Z_{k+1})\big)\Big] \qquad Z_{k+1} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathrm{Id}) \tag{3}$$

These Gaussian perturbations of the iterates have then been extended to other random transformations, such as rotations, reflections, and permutations, in the Equivariant RED (ERED) algorithm Terris et al. (2024); Renaud et al. (2025). This aims to make reconstructions robust to transformations under which the image prior is expected to be invariant.

We analyze in Section 3 the convergence of these RED algorithms, under the assumption that the denoiser D_σ is conservative i.e. its residual is the gradient of a smooth potential $g_\sigma : \mathbb{R}^d \to \mathbb{R}$:

$$D_\sigma(z) = z - \nabla g_\sigma(z). \tag{4}$$

The MMSE Gaussian denoiser $D_\sigma^{\mathrm{MMSE}}(z) = \mathbb{E}[X|X+\sigma Z = z]$, which is the optimal denoiser when training a neural network by minimizing the L^2 loss, satisfies this property with $g_\sigma = -\sigma^2 \log p_\sigma$ (Tweedie's formula (Efron, 2011)). Another example is the Gradient-Step (GS) denoiser (Hurault et al., 2022a) which is trained explicitly in this form, with g_σ parametrized by a neural network. In the following, we focus on these two classes of denoisers.

PnP algorithms (Venkatakrishnan et al., 2013). They are built by replacing the proximal operator of the regularization term $\mathrm{Prox}_{\tau g}(x)$ by the denoiser D_σ in proximal optimization schemes for solving (1). For instance, starting from Proximal Gradient Descent, we get the PnP–PGD algorithm:

$$\text{(PnP–PGD)} \qquad x_{k+1} = D_{\sigma_k}\Big(x_k - \gamma\nabla f(x_k)\Big). \tag{5}$$

Similarly to SNORE for GD, a stochastic version of this PGD algorithm has been proposed independently by Renaud et al. (2026) (SNOPnP) and by Martin et al. (2025) (PnP–Flow, see Appendix B.5).

For the convergence of these PnP algorithms, we will suppose that the plugged denoiser D_σ is a proximal map i.e. there is a potential $\phi_\sigma : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ such that:

$$D_\sigma(z) = \mathrm{Prox}_{\phi_\sigma}(z) \in \operatorname*{arg\,min}_{x\in\mathbb{R}^d} \left\{ \frac{1}{2}\|x - z\|^2 + \phi_\sigma(x) \right\}. \tag{6}$$

The exact MMSE Gaussian denoiser is proximal for the following nonconvex potential (see Lemma A.1 for the proof) (Gribonval, 2011):

$$\phi_\sigma(x) = \left(\frac{1}{2}\|\cdot\|^2 - g_\sigma\right)^*(x) - \frac{1}{2}\|x\|^2. \tag{7}$$

This is also the case of any conservative denoiser (4) with contractive ∇g_σ (Gribonval & Nikolova, 2020). In order to make a learned gradient-step denoiser (4) proximal, Hurault et al. (2022b) proposed a regularization of the training loss to force the contractivity of ∇g_σ.

Before turning to their convergence analysis, we clarify that each algorithm includes as parameters: the denoiser noise level σ_k, which decreases along the iterations, the stepsize $\gamma_k > 0$, which may vary along the iterations, and a fixed regularization weight λ that balances the regularization term against the data-fidelity term f.

3 CONVERGENCE ANALYSIS WITH ANNEALED NOISE LEVELS σ_k

In this section, we establish convergence guarantees for the algorithms introduced above, when the denoising level σ_k decreases along iterates. Because regularity of the denoiser can deteriorate as $\sigma \to 0$, noise annealing does not decrease below a small positive floor ϵ. Throughout the analysis, we write $I_\epsilon = [\epsilon, \sigma_0]$. For all methods, monotone convergence of σ_k suffices for stationarity, we thus keep the following assumption.

Assumption 3.1 (Noise-level schedule)**.** The denoiser levels satisfy $\sigma_{k+1} \leq \sigma_k$ and $\sigma_k \to \epsilon > 0$.

We assume that RED algorithms use the conservative denoiser (4) and PnP algorithms its proximal version (6). Each iterate is then associated with an explicit moving objective consisting of the data-fidelity term f and a regularization term that depends on the current denoising level. No convexity assumption is imposed on these objectives since the regularizers induced by gradient-step and MMSE denoisers are typically nonconvex.We take the following general assumption on the data-fidelity term.

Assumption 3.2 (Data-fidelity term)**.** The function $f \in \mathcal{C}^1(\mathbb{R}^d)$ is bounded from below, subanalytic and has globally L_f-Lipschitz gradient.

This assumption is verified for a large class of inverse problems, in particular when the observation noise is Gaussian $y \sim \mathcal{N}(Ax, \nu^2 \operatorname{Id})$, for which the data-fidelity term is the L^2 norm $f = \frac{1}{2}\|Ax - y\|^2$. Subanalyticity is a mild sufficient condition for verifying the Kurdyka–Łojasiewicz condition (Bolte et al., 2007), which is a standard assumption for nonconvex convergence.

In the following subsections, we prove convergence of the different algorithms introduced in Section 2. For all algorithms, we prove (almost sure) asymptotic stationarity of the iterates relative to an explicit nonconvex objective. Moreover, for the deterministic algorithms, we also prove point convergence of the iterates when $|\sigma_k - \epsilon| = O(k^{-\alpha})$ for some $\alpha > 1$. Table 1 summarizes the convergence guarantees established for the algorithms considered in this work.

Table 1: Convergence guarantees for the algorithms introduced in Section 2 with $\sigma_k \downarrow \epsilon$.

Method	**Objective**	**Stepsize conditions**	**Guarantee**
RED–GD (Section 3.1)	$f + \lambda g_\epsilon$	$\gamma(L_f + \lambda L_g) < 2$	Stationarity and point convergence
SNORE (Section 3.2)	$f + \lambda \mathbb{E}_Z[g_\epsilon(\cdot + \epsilon Z)]$	$\gamma_k(L_f + \lambda L_g) < 2$; $\sum_k \gamma_k = \infty, \sum_k \gamma_k^2 < \infty$	Stationarity a.s.
ERED (Appendix B.3)	$f + \lambda \mathbb{E}_Q[g_\epsilon(Q\cdot)]$	$\gamma_k(L_f + \lambda L_g) < 2$; $\sum_k \gamma_k = \infty, \sum_k \gamma_k^2 < \infty$	Stationarity a.s.
PnP–PGD (Section 3.3)	$f + \gamma^{-1}\phi_\epsilon$	$\gamma L_f < 1$	Stationarity and point convergence
SNOPnP / PnP–Flow (Appendix B.5)	$f + \gamma^{-1}\phi_\epsilon$	$\gamma L_f < 1$; $\sum_k \tau_k^2 < \infty$	Stationarity a.s.

3.1 ANNEALED RED–GD

We first study the convergence of the annealed RED–GD algorithm (2). We consider a conservative denoiser $D_\sigma = \operatorname{Id} - \nabla g_\sigma$. Each update is then a gradient-descent step for the moving objective $F_{\sigma_k} = f + \lambda g_{\sigma_k}$.

We require the potential g_σ to be, uniformly in σ, coercive, with Lipschitz gradient, and with controlled growth of its derivative w.r.t σ. More precisely:

Assumption 3.3 (Gradient-Step Denoiser)**.** The denoiser writes $D_\sigma = \operatorname{Id} - \nabla g_\sigma$ where $(x, \sigma) \mapsto g_\sigma(x)$ is jointly $\mathcal{C}^2$ and subanalytic on $\mathbb{R}^d \times (0, \infty)$. We also assume that ∇g_σ is L_g-Lipschitz for all $\sigma \in I_\epsilon$ and that there are positive constants c_0, C_0, M_g such that, for all $x \in \mathbb{R}^d, \sigma \in I_\epsilon$

$$g_\sigma(x) \geq c_0\|x\|^2 - C_0, \quad \text{and} \quad |\partial_\sigma g_\sigma(x)| \leq M_g(1 + \|x\|^2). \tag{8}$$

Appendix A.1 verifies these conditions for the exact MMSE denoiser under a compactly supported prior. This assumption is natural for digital images, whose clean pixel values lie in a bounded range, typically $[0, 1]$. For learned gradient-step denoisers, we show that the conditions hold up to a mild regularization that leaves the denoiser unchanged over the relevant pixel range.

Then, under standard stepsize conditions, we show that monotone convergence of σ_k to ϵ suffices for asymptotic stationarity of the iterates with respect to the terminal objective F_ϵ. For summable polynomial decay, we also prove convergence of the sequence to a critical point of this objective.

Theorem 3.4 (Convergence of RED–GD (2), proof in Appendix B.1)**.** *Suppose Assumptions 3.1, 3.2, 3.3 hold and constant stepsize $\gamma < \frac{2}{L_f + \lambda L_g}$. Then $\|\nabla F_\epsilon(x_k)\| \to 0$, $\|x_{k+1} - x_k\| \to 0$, and every cluster point is critical for F_ϵ. In particular,* $\operatorname{dist}(x_k, \operatorname{crit} F_\epsilon) \to 0$. *Moreover,*

$$\min_{0 \le k \le K} \|\nabla F_{\sigma_k}(x_k)\|^2 = \mathcal{O}\left(\frac{1}{K}\right). \tag{9}$$

If, in addition, $|\sigma_k - \epsilon| = \mathcal{O}(k^{-\alpha})$ for some $\alpha > 1$, then the iterates have finite length and converge to a critical point of F_ϵ.

Our proof extends to annealed denoisers the fixed-noise convergence analysis of Hurault et al. (2022a). Under the stepsize condition, Assumption 3.3 yields bounded iterates. On this bounded set, changing the noise level adds at most $|\sigma_{k+1} - \sigma_k|$ to the descent inequality. This error is summable because $\sigma_k \downarrow \epsilon$, giving asymptotic stationarity for F_ϵ without a rate condition. The inexact Kurdyka–Łojasiewicz theorem of Ochs (2019, Theorem 10) gives full-sequence convergence.

Remark 3.5 (Rate for the terminal objective)**.** The rate (9) concerns the gradient of the moving objective F_{σ_k}. The proof gives the same rate for the terminal gradient if $\sum_k |\sigma_k - \epsilon|^2 < \infty$.

3.2 Annealed SNORE

We now prove convergence of SNORE (3), the stochastic variant of RED–GD. As detailed in Renaud et al. (2024), the algorithm with fixed σ targets minimization of $F_\sigma = f + \lambda \tilde{g}_\sigma$ with regularizer

$$\tilde{g}_\sigma(x) = \mathbb{E}_Z[g_\sigma(x + \sigma Z)], \qquad Z \sim \mathcal{N}(0, I_d). \tag{10}$$

Using decreasing noise levels $\sigma_k \downarrow \epsilon$, we prove stationarity with respect to the limiting objective $F_\epsilon = f + \lambda \tilde{g}_\epsilon$. We obtain the same asymptotic stationarity conclusion as for RED–GD in Theorem 3.4, now almost surely. The main difference is that, as typical for stochastic optimization algorithms, convergence requires square summable vanishing stepsizes.

Theorem 3.6 (Convergence of SNORE (3). Proof in Appendix B.2)**.** *Suppose Assumptions 3.1, 3.2, 3.3 hold and that the stepsizes verify*

$$0 < \gamma_k \le \bar{\gamma}, \qquad (L_f + \lambda L_g)\bar{\gamma} < 2, \qquad \sum_k \gamma_k = \infty, \qquad \sum_k \gamma_k^2 < \infty. \tag{11}$$

Then, almost surely, $\|\nabla F_\epsilon(x_k)\| \to 0$, $\|x_{k+1} - x_k\| \to 0$, and every cluster point is critical for F_ϵ; in particular, $\operatorname{dist}(x_k, \operatorname{crit} F_\epsilon) \to 0$, *almost surely. Moreover*

$$\min_{0 \le k \le K} \mathbb{E}\|\nabla F_{\sigma_k}(x_k)\|^2 = \mathcal{O}\left(\left(\textstyle\sum_{j=0}^{K} \gamma_j\right)^{-1}\right). \tag{12}$$

The proof follows the RED–GD strategy in conditional expectation. The Gaussian noise adds a variance term of order γ_k^2, which is handled by the summability assumption. The Robbins–Siegmund theorem (Robbins & Siegmund, 1971) gives summability of $\gamma_k \|\nabla F_{\sigma_k}(x_k)\|^2$. We then control the remaining stochastic fluctuations to obtain almost-sure stationarity for F_ϵ.

Remark 3.7 (Normalization by σ^2)**.** Following the original SNORE version (Renaud et al., 2024), for the exact MMSE denoiser, due to the form of Tweedie's formula, one may divide the denoiser residual in (4) by σ_k^2. Theorem 3.6 then applies with $F_\sigma = f + \frac{\lambda}{\sigma^2}\tilde{g}_\sigma$. Since $\sigma_k \to \epsilon > 0$, the only difference is that the stepsize condition becomes $\left(L_f + \lambda L_g/\epsilon^2\right)\bar{\gamma} < 2$.

In Appendix B.3, we also prove the same stationarity result for the stochastic ERED algorithm (60).

3.3 Annealed PnP–PGD

We now study the convergence of PnP algorithms, in which the proximal operator of the regularization term is replaced by a denoiser. Our analysis relies on denoisers that can be written as proximal mappings $D_\sigma = \operatorname{Prox}_{\phi_\sigma}$ for a potential ϕ_σ. As explained in Section 2, this is verified for both the MMSE denoiser and a learned gradient-step denoiser $D_\sigma = \operatorname{Id} - \nabla g_\sigma$ when ∇g_σ is contractive.

Assumption 3.8 (Proximal denoiser)**.** Let $D_\sigma = \mathrm{Id} - \nabla g_\sigma$. Assume either that there exists $\rho < 1$ such that, for every $\sigma \in I_\epsilon$, ∇g_σ is ρ-Lipschitz, or that D_σ is the exact MMSE denoiser associated with a compactly supported prior whose support is not contained in an affine hyperplane.

As detailed in Appendix A.2, in the MMSE case, compact support ensures that $\mathrm{Im}(D_\sigma) \subseteq \mathrm{conv}(\mathrm{supp}\, P_X)$ uniformly in σ. This ensures that every PnP iterate remains bounded.

With such proximal denoiser, annealed PnP–PGD (5) is proximal gradient descent for the nonconvex moving objective $F_{\sigma_k} := f + \gamma^{-1}\phi_{\sigma_k}$. The stepsize appears in the objective because the proximal denoiser applies $\mathrm{prox}_{\phi_{\sigma_k}}$ instead of $\mathrm{prox}_{\gamma\phi_{\sigma_k}}$ as for standard PGD. The stepsize thus changes the effective regularization weight. Similar to RED–GD, we prove convergence of the algorithm and asymptotic stationarity with respect to the terminal objective $F_\epsilon := f + \gamma^{-1}\phi_\epsilon$.

Theorem 3.9 (Convergence of PnP–PGD (5), proof in Appendix B.4)**.** *Suppose Assumptions 3.1, 3.2, 3.3, 3.8 hold and constant stepsize $\gamma < \frac{1}{L_f}$. Then, $\|\nabla F_\epsilon(x_k)\| \to 0$, $\|x_{k+1} - x_k\| \to 0$, and every cluster point is critical for F_ϵ. Moreover,*

$$\min_{1\le k\le K} \|\nabla F_{\sigma_k}(x_k)\|^2 = \mathcal{O}\left(\frac{1}{K}\right). \tag{13}$$

If, in addition, $|\sigma_k - \epsilon| = \mathcal{O}(k^{-\alpha})$ for some $\alpha > 1$, then the iterates have finite length and converge to a critical point of F_ϵ.

Note that, because the stepsize enters the objective $F_\sigma = f + \gamma^{-1}\phi_\sigma$, the condition $\gamma L_f < 1$ imposes a lower bound on the effective regularization weight (Hurault et al., 2023). This can be restrictive for PnP–PGD with fixed σ: in our face-hole experiment (Section 4), σ must be large enough to fill the missing region, but this leads to oversmoothing unless the regularization weight is reduced, i.e., unless γ is increased. Annealing precisely mitigates this tradeoff: a large σ helps fill the hole early, while a smaller terminal σ limits overregularization near convergence.

Annealed PnP–Flow / SNOPnP. We also study in Appendix B.5 the convergence of a stochastic version of PnP–PGD with decreasing noise-levels. The algorithm appeared in both PnP–Flow (Martin et al., 2025) and SNOPnP (Renaud et al., 2026), and we prove that they are equivalent. In these algorithms, the Gaussian perturbation and the denoiser use the same level σ_k. When $\sigma_k \to \epsilon > 0$, the injected noise retains nonzero variance, so our descent argument cannot control its accumulated effect. We therefore let the noise amplitude τ_k vanish independently of the denoiser level σ_k

$$x_{k+1} = D_{\sigma_k}\Big(x_k - \gamma\nabla f(x_k) + \tau_k Z_{k+1}\Big), \qquad Z_{k+1} \sim \mathcal{N}(0, I_d). \tag{14}$$

Theorem B.2 in Appendix B.5 then proves, for $\sum_k \tau_k^2 < \infty$, almost-sure asymptotic stationarity of these iterates with respect to F_ϵ.

4 Experiments

In the experiments, we illustrate, on various image inverse problems, the convergence properties established in the previous section. We also confirm that decreasing the denoiser noise along the iterations improves restoration quality over running the same algorithm at constant σ. All experiments are implemented with the DeepInverse library (Tachella et al., 2025).

Inverse problems and data. We experiment on both natural images (68 images of CBSD68 (Martin et al., 2001) downscaled to 128×128) and face images (100 images from FFHQ (Karras et al., 2019) at 128×128). We consider different linear inverse problems with observations $y = Ax + w$, where $w \sim \mathcal{N}(0, \sigma_n^2 I)$ ($\sigma_n = 0.01$ across all experiments) (i) inpainting: a centered 64×64 hole on faces and a random mask hiding 50% of the pixels on natural images (ii) demosaicing of a Bayer color filter array (iii) $\times 4$ super-resolution with a bicubic anti-aliasing filter. We also provide complementary experiments on sparse-view parallel-beam tomography in Appendix C. The algorithms are initialized from the observed pixels completed by a normalized convolution for inpainting and demosaicing, and bicubic upsampling for super-resolution.

Algorithms. We compare RED–GD (2), PnP–PGD (5), SNORE (3) (Renaud et al., 2024), Equivariant RED (ERED) (60) (Renaud et al., 2025), and SNOPnP / PnP–Flow (14) (Renaud et al., 2026; Martin et al., 2025). For ERED, the random transformations are drawn from the dihedral group (rotations by multiples of 90° and flips) on natural images, and from $\{\mathrm{Id}, \text{horizontal flip}\}$ on face images. As an external reference we also report DPIR (Zhang et al., 2022), i.e. eight half-quadratic splitting iterations with its default decreasing noise schedule.

Denoisers. All experiments are run with gradient-step (GS) denoisers $D_\sigma = \mathrm{Id} - \nabla g_\sigma$ with, following Hurault et al. (2022a) $g_\sigma(x) = \frac{1}{2}\|x - N_\sigma(x)\|^2$ where N_σ is a UNet neural network. The denoiser D_σ is trained using the standard L^2 denoising loss. For experiments on FFHQ faces, we parameterize N_σ using the pretrained DDPM U-Net architecture from Dhariwal & Nichol (2021) (referred to as *DiffUNet* in the deepinv library) which we finetune with our GS denoiser loss on 128×128 images over noise levels $\sigma \in [0.002, 5]$. We call it GS–DiffUNet. For natural image experiments, we parameterize N_σ using the DRUNet architecture (Zhang et al., 2022), the resulting gradient-step denoiser is called *GS–DRUNet*. Since the original model of Hurault et al. (2022a) was trained only for noise levels $\sigma \leq 0.2$, we finetune the denoiser on the extended range $\sigma \in [0.002, 2]$. This wider noise-level range is particularly useful for annealed schemes. Moreover, we recorded every input passed to a GS denoiser in the reported experiments and verified that all such inputs remain in a fixed compact set. Consequently, the regularization introduced in Appendix A.1 is never activated, and Assumption 3.3 is satisfied along the reported trajectories.

For the PnP experiments, in order to obtain proximal denoisers, following Hurault et al. (2022b), we finetune (and relax) these GS denoisers with a soft penalty that encourages $\|\nabla^2 g_\sigma(x)\|_S < 1$. The resulting denoiser is then, at least locally, the proximal map of a potential ϕ_σ. We use Lanczos iterations for estimating the spectral norm during training, which is considerably faster to converge than the power iterations initially used by Hurault et al. (2022b). Unlike conservativity, which holds by construction, the contraction condition sufficient for proximality is promoted by a soft penalty, which is not guaranteed for every input, but is expected to hold near the image manifold. See Appendix C.1 for details.

Schedules and hyper-parameters. In all experiments, the noise level follows the geometric schedule

$$\sigma_k = \epsilon + (\sigma_0 - \epsilon) r^k \quad \text{with} \quad 0 < r < 1 \tag{15}$$

which decreases at every iteration and converges to ϵ faster than the $\mathcal{O}(k^{-\alpha})$, $\alpha > 1$, required for full sequence convergence of Theorems 3.4 and 3.9.

We use stepsizes which follow each algorithm's requirements for convergence. For RED–GD, SNORE and ERED, the stepsize condition involves the Lipschitz constant of ∇g_σ which is first estimated empirically on a validation set. The stochastic algorithms SNORE and ERED require vanishing stepsizes ($\sum \gamma_k = \infty$ and $\sum \gamma_k^2 < \infty$), we keep the step constant during the main noise annealing phase and then let it decay polynomially. We tune these method-specific parameters on validation images. Appendix C.2 gives the exact schedules, and parameter values.

Results. Table 2 reports final PSNR on FFHQ, while Table 4 in the Appendix reports results on natural images (CBSD68). Both tables compare constant and annealed noise levels versions.

Table 2: PSNR (dB) on 100 FFHQ face images after 500 iterations with constant or annealed (15) noise level. RED methods (RED–GD, SNORE, ERED) and DPIR use the GS–DiffUNet denoiser; PnP methods (PnP–PGD, SNOPnP) use Prox–DiffUNet denoiser. Best value per problem in bold. The corresponding CBSD68 results are in Table 4 of Appendix C.

	Inpainting		**Demosaicing**		**SR** $\times 4$	
Method	Constant	Annealed	Constant	Annealed	Constant	Annealed
RED–GD	24.22	25.36	33.54	37.89	27.17	27.51
PnP–PGD	20.03	25.09	35.35	37.76	26.49	27.25
SNORE	24.34	24.98	33.32	**38.03**	27.14	27.34
SNOPnP / PnP–Flow	21.93	24.42	36.22	37.78	27.18	27.29
ERED	24.08	**25.44**	33.55	38.00	27.24	**27.56**
DPIR	19.83		35.90		26.26	

Figures 2, 3, and 4 show representative reconstructions. Annealing improves PSNR in every case, with the largest gains in inpainting and demosaicing. As Figure 2 shows, a low constant σ can leave large holes unfilled (PnP–PGD), while a high constant σ can produce blurry results (RED–GD). Annealing σ fills the holes with more detail. In demosaicing, it removes the color artifacts that remain with a constant σ. Sparse-view tomography also benefits substantially from noise-level annealing (Table 4). It provides smaller gains for inpainting with randomly missing pixels (Table 4) and $\times 4$ super-resolution (Table 2). In these settings, no large missing region needs to be reconstructed, so a small, fixed σ can work well throughout the iterations.

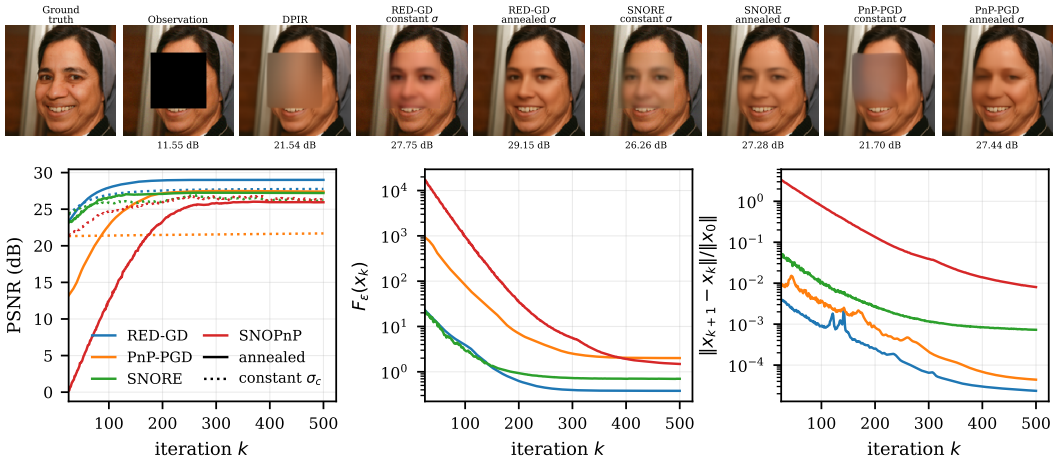

Figure 2: Inpainting of a 64×64 hole on FFHQ 128×128. Top: the constant-σ versions (and DPIR) leave a smooth patch, while the annealed versions fill the hole. Bottom: PSNR, target objective $F_\epsilon(x_k)$ and residual $\|x_{k+1} - x_k\|/\|x_0\|$ of the annealed runs, along the iterations, for the image shown above. Note that every algorithm has its own target objective F_ϵ, so only the decrease of each curve is meaningful, not the levels.

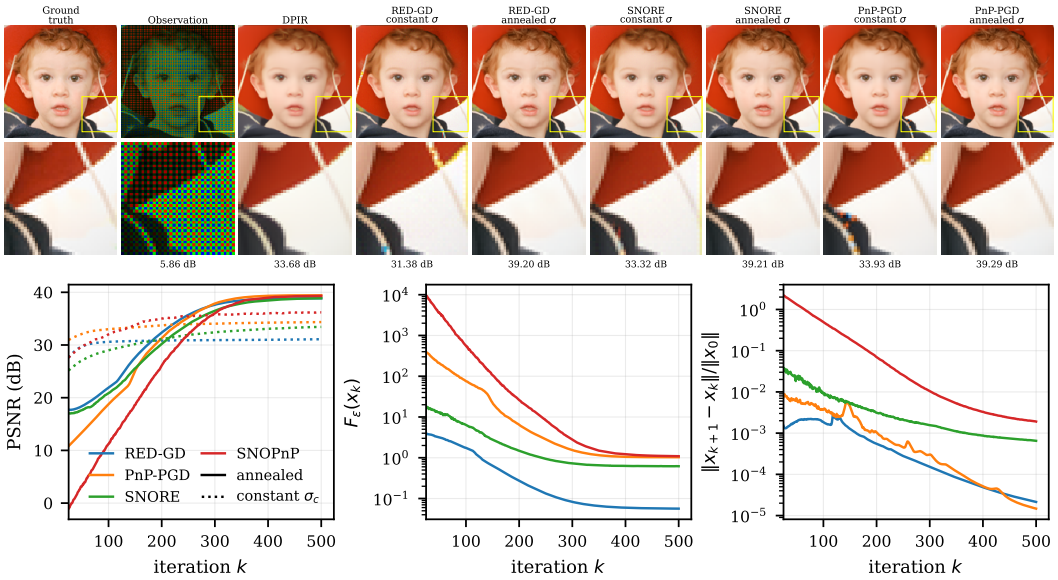

Figure 3: Demosaicing of a Bayer color filter array on FFHQ 128×128. The constant-σ versions (and DPIR) keep the color fringes of the initial interpolation along the edges, the annealed versions remove them. Bottom: PSNR, terminal objective and residual along the iterations.

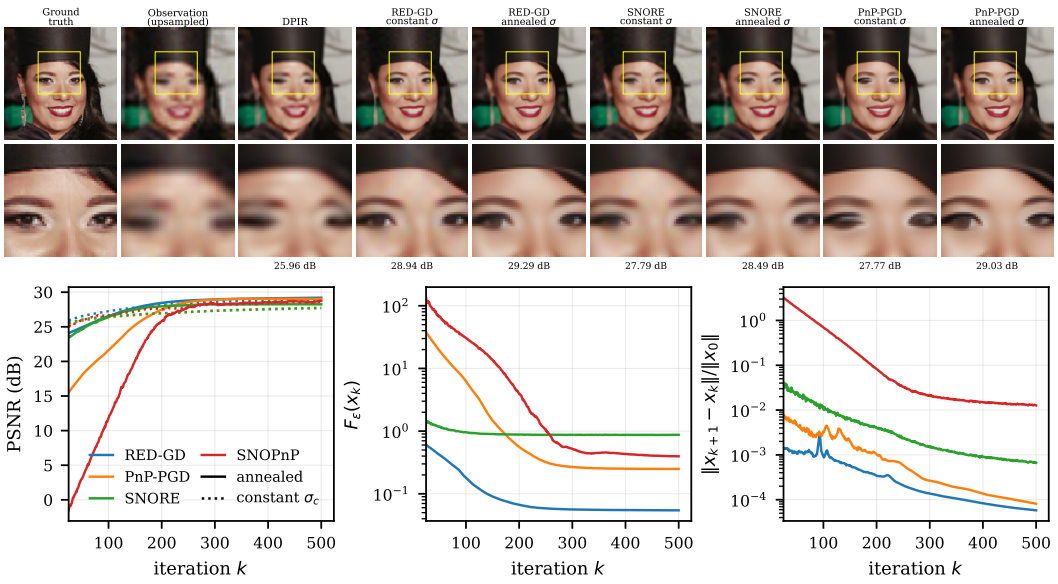

Figure 4: $\times 4$ super-resolution on FFHQ 128×128. PSNR differences are smaller than in the previous experiments, but the annealed methods yield visibly sharper and more detailed reconstructions. Bottom: PSNR, terminal objective and residual along the iterations.

Table 2 also shows that annealing is particularly useful for PnP algorithms (PnP–PGD and SNOPnP), in particular for large hole inpainting. As explained in Section 3.3, for these algorithms, the stepsize γ also determines the effective regularization weight. Using a large constant σ would require increasing γ to avoid over-regularization, but the convergence condition $\gamma L_f < 1$ limits this adjustment. RED methods do not face this constraint, as their regularization weight can be tuned independently. Noise annealing alleviates the limitation for PnP–PGD by using a large σ early to fill the hole and a small σ near convergence, thereby avoiding over-regularization in the final reconstruction.

Convergence. The bottom rows of Figures 2, 3, and 4 track the PSNR, the terminal objective $F_\epsilon(x_k)$ associated with each algorithm, and the normalized residual $\|x_{k+1} - x_k\|/\|x_0\|$ for the image shown above. Details on the computation of $F_\epsilon(x_k)$ are given in Appendix C.2. Across the three tasks, as predicted by our theoretical results, the objective values stabilize and the residuals decay toward zero. Note that although annealing can initially yield lower PSNR than a constant-noise run, its PSNR later overtakes the baseline as the noise level decreases.

5 Conclusion

We have shown theoretically that plug-and-play (RED and PnP) algorithms with decreasing denoising strength converge toward stationary points of an explicit terminal objective. Our analysis does not prescribe an annealing rate or require convexity assumptions. This closes the gap between previous fixed-denoiser plug-and-play convergence theories and the decreasing noise schedules used by modern restoration algorithms. We confirm experimentally on various inverse problems that noise-annealed plug-and-play converges and gives better reconstruction than its constant noise level counterpart.

Our results show that annealing preserves convergence to stationarity, but do not explain why it may yield better solutions: annealed and constant-level algorithms satisfy the same stationarity guarantee. Understanding theoretically how annealing influences stationary-point selection remains an open question. A second limitation is that our analysis requires conservative denoisers and/or contractive residuals. These properties hold for the ideal MMSE denoiser and are enforced or encouraged through regularization for our trained denoisers. However, standard learned denoisers, even if trained to approximate the MMSE denoiser, may satisfy these assumptions only approximately due to finite model capacity and imperfect training. Quantifying how such approximation errors affects convergence and performance is an important direction for future work.

REPRODUCIBILITY STATEMENT

Appendix C contains all the necessary information for reproducibility. It describes the experimental setup in detail: the datasets and denoiser architectures, the gradient-step and proximal fine-tuning procedures, and the noise schedules, stepsizes, and selected hyperparameters used by each optimization algorithm. The code, the trained denoiser checkpoints, and a notebook reproducing all figures of the paper are available at https://anonymous.4open.science/r/pnp-sigma.

AI USE STATEMENT

We used generative AI tools, including Codex and Claude to assist in drafting and revising mathematical proofs, reviewing related literature, and for running the experiments. The authors are responsible for the accuracy of all claims, citations, proofs, and experimental results in the final submission.

REFERENCES

Jérôme Bolte, Aris Daniilidis, and Adrian Lewis. The łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM Journal on Optimization*, 17(4):1205–1223, 2007.

Stanley H. Chan, Xiran Wang, and Omar A. Elgendy. Plug-and-play ADMM for image restoration: Fixed-point convergence and applications. *IEEE Transactions on Computational Imaging*, 3(1):84–98, 2017. doi: 10.1109/TCI.2016.2629286. URL https://arxiv.org/abs/1605.01710.

Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. ILVR: Conditioning method for denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14367–14376, 2021.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pp. 8780–8794, 2021.

Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011.

Gene H. Golub and Charles F. Van Loan. *Matrix computations*. Johns Hopkins University Press, 4 edition, 2013.

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

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

Ruizhi Hou, Fang Li, and Guixu Zhang. Truncated residual based plug-and-play ADMM algorithm for MRI reconstruction. *IEEE Transactions on Computational Imaging*, 8:96–108, 2022. doi: 10.1109/TCI.2022.3145187.

Samuel Hurault. *Convergent plug-and-play methods for image inverse problems with explicit and nonconvex deep regularization*. PhD thesis, Université de Bordeaux, 2023.

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

Samuel Hurault, Arthur Leclaire, and Nicolas Papadakis. Proximal denoiser for convergent plug-and-play optimization with nonconvex regularization. In *International Conference on Machine Learning*, 2022b.

Samuel Hurault, Antonin Chambolle, Arthur Leclaire, and Nicolas Papadakis. A relaxed proximal gradient descent algorithm for convergent plug-and-play with proximal denoiser. In *Scale Space and Variational Methods in Computer Vision*, volume 14009 of *Lecture Notes in Computer Science*, pp. 627–638. Springer, 2023.

Ulugbek S. Kamilov, Charles A. Bouman, Gregery T. Buzzard, and Brendt Wohlberg. Plug-and-play methods for integrating physical and learned models in computational imaging: Theory, algorithms, and applications. *IEEE Signal Processing Magazine*, 40(1):85–97, 2023. doi: 10.1109/MSP.2022.3199595.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

Morteza Mardani, Jiaming Song, Jan Kautz, and Arash Vahdat. A variational perspective on solving inverse problems with diffusion models. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2305.04391.

David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *IEEE International Conference on Computer Vision (ICCV)*, volume 2, pp. 416–423, 2001.

Ségolène Martin, Anne Gagneux, Paul Hagemann, and Gabriele Steidl. Pnp-flow: Plug-and-play image restoration with flow matching. In *International Conference on Learning Representations*, volume 2025, pp. 45466–45492, 2025.

Peter Ochs. Unifying abstract inexact convergence theorems and block coordinate variable metric iPiano. *SIAM Journal on Optimization*, 29(1):541–570, 2019. doi: 10.1137/17M1124085.

Chicago Y. Park, Edward P. Chandler, Yuyang Hu, Michael T. McCann, Cristina Garcia-Cardona, Brendt Wohlberg, and Ulugbek S. Kamilov. Stochastic generative plug-and-play priors. *arXiv preprint arXiv:2604.03603*, 2026.

Scott Pesme, Giacomo Meanti, Michael Arbel, and Julien Mairal. MAP estimation with denoisers: Convergence rates and guarantees. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2507.15397.

Edward T. Reehorst and Philip Schniter. Regularization by denoising: Clarifications and new interpretations. *arXiv preprint arXiv:1806.02296*, 2018. URL https://arxiv.org/abs/1806.02296.

Marien Renaud, Jean Prost, Arthur Leclaire, and Nicolas Papadakis. Plug-and-play image restoration with stochastic deNOising REgularization. In *International Conference on Machine Learning*, pp. 42484–42520, 2024.

Marien Renaud, Arthur Leclaire, and Nicolas Papadakis. Equivariant denoisers for image restoration. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pp. 227–240. Springer, 2025.

Marien Renaud, Eliot Guez, Arthur Leclaire, and Nicolas Papadakis. Equivariant denoisers for plug-and-play image restoration. *Journal of Mathematical Imaging and Vision*, 68(4):54, 2026.

Herbert Robbins and David Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing methods in statistics*, pp. 233–257. Elsevier, 1971.

Yaniv Romano, Michael Elad, and Peyman Milanfar. The little engine that could: Regularization by denoising (red). *SIAM journal on imaging sciences*, 10(4):1804–1844, 2017.

Ernest K. Ryu, Jialin Liu, Sicheng Wang, Xiaohan Chen, Zhangyang Wang, and Wotao Yin. Plug-and-play methods provably converge with properly trained denoisers. In *International Conference on Machine Learning*, 2019.

Rajesh Shrestha and Xiao Fu. Taming score-based denoisers in ADMM: A convergent plug-and-play framework. *arXiv preprint arXiv:2603.10281*, 2026. URL https://arxiv.org/abs/2603.10281.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2011.13456.

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

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

Julián Tachella, Matthieu Terris, Samuel Hurault, Andrew Wang, Dongdong Chen, Minh-Hai Nguyen, et al. DeepInverse: A Python package for solving imaging inverse problems with deep learning. *arXiv preprint arXiv:2505.20160*, 2025.

Matthieu Terris, Thomas Moreau, Nelly Pustelnik, and Julian Tachella. Equivariant plug-and-play image reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25255–25264, 2024.

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

Kaixuan Wei, Angelica Aviles-Rivero, Jingwei Liang, Ying Fu, Carola-Bibiane Schönlieb, and Hua Huang. Tuning-free plug-and-play proximal algorithm for inverse imaging problems. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 10158–10169, 2020. URL https://proceedings.mlr.press/v119/wei20b.html.

Xiaojian Xu, Yu Sun, Jiaming Liu, Brendt Wohlberg, and Ulugbek S. Kamilov. Provable convergence of plug-and-play priors with MMSE denoisers. *IEEE Signal Processing Letters*, 27:1280–1284, 2020.

Kai Zhang, Wangmeng Zuo, Shuhang Gu, and Lei Zhang. Learning deep CNN denoiser prior for image restoration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. URL https://arxiv.org/abs/1704.03264.

Kai Zhang, Yawei Li, Wangmeng Zuo, Lei Zhang, Luc Van Gool, and Radu Timofte. Plug-and-play image restoration with deep denoiser prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6360–6376, 2022.

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

A VERIFICATION OF ASSUMPTIONS

A.1 VERIFICATION OF ASSUMPTION 3.3

We check the three global bounds in Assumption 3.3, as well as joint regularity and subanalyticity.

Exact MMSE denoiser with compact prior. Assume that $\|X\| \le R$ almost surely, and let p_σ be the density of $Y = X + \sigma Z$, where $Z \sim \mathcal{N}(0, I_d)$. Set $g_\sigma(x) = -\sigma^2 \log p_\sigma(x)$. Gaussian convolution makes $p_\sigma > 0$ jointly real analytic in (x, σ) for $\sigma > 0$. Thus g_σ is jointly real analytic, hence jointly $\mathcal{C}^\infty$ and subanalytic. Tweedie's formula and differentiation of the posterior mean give

$$\nabla g_\sigma(x) = x - D_\sigma^{\mathrm{MMSE}}(x), \qquad \nabla^2 g_\sigma(x) = I - \sigma^{-2} \operatorname{Cov}(X \mid Y = x).$$

Since $\| \operatorname{Cov}(X \mid Y = x)\|_{\mathrm{op}} \le R^2$, the gradients ∇g_σ are globally L_g-Lipschitz for all $\sigma \in I_\epsilon$, with $L_g = 1 + R^2/\epsilon^2$.

To verify coercivity and growth in σ, write

$$g_\sigma(x) = \frac{d}{2}\sigma^2 \log(2\pi\sigma^2) - \sigma^2 \log W_\sigma(x). \tag{16}$$

with $W_\sigma(x) := \mathbb{E} \exp\Big(-\frac{\|x-X\|^2}{2\sigma^2}\Big)$. Using

$$\|x - X\|^2 \ge \tfrac{1}{2}\|x\|^2 - \|X\|^2 \ge \tfrac{1}{2}\|x\|^2 - R^2,$$

we have

$$\log W_\sigma(x) \le -\frac{\|x\|^2}{4\sigma^2} + \frac{R^2}{2\sigma^2}$$

Thus, for all $\sigma \in I_\epsilon$,

$$g_\sigma(x) \ge \tfrac{1}{4}\|x\|^2 - \tfrac{1}{2}R^2 + \tfrac{d}{2}\sigma^2 \log(2\pi\sigma^2).$$

Finally, differentiating (16) with respect to σ gives

$$\partial_\sigma g_\sigma(x) = d\sigma\big(\log(2\pi\sigma^2) + 1\big) - 2\sigma \log W_\sigma(x) - \frac{1}{\sigma}\mathbb{E}[\|x - X\|^2 \mid Y = x].$$

Both $|\log W_\sigma(x)|$ and the conditional squared distance are bounded by a constant times $1 + \|x\|^2$, uniformly for $\sigma \in I_\epsilon$. Hence $|\partial_\sigma g_\sigma(x)| \le M_g(1 + \|x\|^2)$, as required.

Gradient-step denoisers. Suppose $g_\sigma : \mathbb{R}^d \to \mathbb{R}$ is parametrized by a feedforward network with affine layers and $\mathcal{C}^2$ activations whose first and second derivatives are globally bounded. Assume these bounds are uniform for $\sigma \in I_\epsilon$. Then the ∇g_σ are uniformly globally Lipschitz. However, Hurault et al. (2022a) consider learned potential parametrized as

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

where $(x, \sigma) \mapsto N_\sigma(x)$ is a neural network, with activation chosen such that it is jointly $\mathcal{C}^2$ in (x, σ) and subanalytic. We propose here a mild regularization of this potential such that it is uniformly Lipschitz in x and uniformly coercive.

Fix a compact box $\mathcal{B}$ containing the prescribed image range, typically $\mathcal{B} = [-1, 2]^d$ for image pixels normalized in $[0, 1]$. Choose a $\mathcal{C}^2$ semialgebraic cutoff $\chi : \mathbb{R}^d \to [0, 1]$ that equals 1 on a neighbourhood of $\mathcal{B}$ and vanishes outside a ball of radius R. For any $\mu > 0$, define

$$\widetilde{g}_\sigma(x) = \chi(x)^2 g_\sigma(x) + \frac{\mu}{2}\big(1 - \chi(x)^2\big)\|x\|^2, \qquad \widetilde{D}_\sigma = \mathrm{Id} - \nabla\widetilde{g}_\sigma. \tag{17}$$

On $\mathcal{B}$, both the potential and the denoiser coincide with their original counterparts. Outside the ball of radius R, $\widetilde{g}_\sigma(x) = \frac{\mu}{2}\|x\|^2$.

This single modification gives the required global bounds. Since $g_\sigma \ge 0$, for every x and $\sigma \in I_\epsilon$,

$$\widetilde{g}_\sigma(x) \ge \frac{\mu}{2}\|x\|^2 - \frac{\mu R^2}{2}.$$

Moreover, the x-Hessian of $\widetilde{g}_\sigma$ is uniformly bounded on the compact support of χ, by joint $\mathcal{C}^2$ regularity and compactness of I_ϵ; outside that support it equals μI. Hence $\nabla \widetilde{g}_\sigma$ is globally Lipschitz, with a constant uniform in $\sigma \in I_\epsilon$. Finally,

$$\partial_\sigma \widetilde{g}_\sigma(x) = \chi(x)^2 \partial_\sigma g_\sigma(x)$$

is uniformly bounded on $\mathbb{R}^d \times I_\epsilon$, and therefore satisfies the required bound $|\partial_\sigma \widetilde{g}_\sigma(x)| \le M_g(1 + \|x\|^2)$. Joint $\mathcal{C}^2$ regularity and subanalyticity are preserved. Thus $\widetilde{g}_\sigma$ verifies Assumption 3.3.

This penalty provides a coercive safeguard against unbounded iterates. It remains inactive whenever the inputs to the denoiser stay inside $\mathcal{B}$, in which case the regularized and original iterations coincide.

A.2 Regularity properties of the proximal denoisers

MMSE denoiser. We first prove in Lemma A.1 the form (7) of the MMSE proximal potential. Then, in Proposition A.2, we prove its regularity properties that will be useful in the PnP proofs.

Lemma A.1 (MMSE proximal potential)**.** *Assume $\mathbb{E}\|X\|^2 < \infty$, let $Z \sim \mathcal{N}(0, I_d)$ be independent of X, and fix $\sigma > 0$. Let p_σ be the density of $X + \sigma Z$, define $g_\sigma = -\sigma^2 \log p_\sigma$, and write $D_\sigma^{\mathrm{MMSE}}(z) = \mathbb{E}[X \mid X + \sigma Z = z]$. Then the potential*

$$\phi_\sigma(x) = \left(\frac{1}{2}\|\cdot\|^2 - g_\sigma\right)^*(x) - \frac{1}{2}\|x\|^2$$

is proper and lower semicontinuous, and $D_\sigma^{\mathrm{MMSE}}(z) = \mathrm{prox}_{\phi_\sigma}(z)$ for every $z \in \mathbb{R}^d$.

Proof. By Gribonval & Nikolova (2020, Theorem 1), it suffices for proximality to show that D_σ^{MMSE} is the gradient of a convex function. We then identify the potential explicitly. Expanding the Gaussian density gives

$$p_\sigma(z) = (2\pi\sigma^2)^{-d/2} e^{-\|z\|^2/(2\sigma^2)} \mathbb{E} \exp\left(\frac{\langle z, X\rangle - \|X\|^2/2}{\sigma^2}\right),$$

and hence

$$\frac{1}{2}\|z\|^2 - g_\sigma(z) = \sigma^2 \log \mathbb{E} \exp\left(\frac{\langle z, X\rangle - \|X\|^2/2}{\sigma^2}\right) - \frac{d}{2}\sigma^2 \log(2\pi\sigma^2). \tag{18}$$

The right-hand side is a log-partition function of affine functions of z, up to a constant, so it is convex. Using Tweedie's formula, we have

$$D_\sigma^{\mathrm{MMSE}}(z) = \nabla\left(\frac{1}{2}\|z\|^2 - g_\sigma(z)\right).$$

For the ϕ_σ defined in the Lemma, we have

$$\frac{1}{2}\|x - z\|^2 + \phi_\sigma(x) = \left(\frac{1}{2}\|\cdot\|^2 - g_\sigma\right)^*(x) - \langle z, x\rangle + \frac{1}{2}\|z\|^2,$$

By Fenchel-Young inequality, the right-hand side is greater than

$$\langle z, x\rangle - \frac{1}{2}\|z\|^2 + g_\sigma(z) - \langle z, x\rangle + \frac{1}{2}\|z\|^2 = g_\sigma(z)$$

with equality if and only if $x = \nabla(\frac{1}{2}\|\cdot\|^2 - g_\sigma)(z) = D_\sigma^{\mathrm{MMSE}}(z)$ □

Proposition A.2 (Properties of the MMSE denoiser)**.** *Suppose P_X has compact support not contained in an affine hyperplane, and set*

$$C = \mathrm{conv}(\mathrm{supp}\, P_X), \qquad U = \mathrm{int}\, C.$$

Then, for every $\sigma > 0$, D_σ^{MMSE} is a smooth diffeomorphism from $\mathbb{R}^d$ onto U, and $\mathrm{int}(\mathrm{dom}\, \phi_\sigma) = U$. Moreover, the maps $(z, \sigma) \mapsto D_\sigma^{\mathrm{MMSE}}(z)$ and $(x, \sigma) \mapsto \phi_\sigma(x)$ are jointly smooth on $\mathbb{R}^d \times (0, \infty)$ and $U \times (0, \infty)$, respectively, ϕ_σ is real analytic on U, and

$$\nabla \phi_\sigma(x) = (D_\sigma^{\mathrm{MMSE}})^{-1}(x) - x, \qquad x \in U.$$

Proof. Fix $\sigma > 0$, and let

$$\nu_\sigma(du) = e^{-\|u\|^2/(2\sigma^2)} P_X(du), \qquad A_\sigma(\theta) = \log \int e^{\langle \theta, u \rangle} \nu_\sigma(du).$$

The posterior at $Y = z$ is its exponential tilt with parameter z/σ^2; thus

$$D_\sigma^{\mathrm{MMSE}}(z) = \nabla A_\sigma(z/\sigma^2), \qquad J_z D_\sigma^{\mathrm{MMSE}}(z) = \sigma^{-2} \operatorname{Cov}(X \mid Y = z) \succ 0.$$

Each posterior has the same full-dimensional support as P_X; hence the covariance is positive definite and the mean map is injective and locally invertible. Its values lie in U: a posterior mean on the boundary of C would force the posterior to be supported on a proper supporting face.

Conversely, if $x \in U$, finitely many support points have a convex hull containing a ball around x. Small neighborhoods of these points have positive ν_σ-mass and show that, for some $c > 0$ and $C_x < \infty$,

$$A_\sigma(\theta) - \langle \theta, x \rangle \geq c\|\theta\| - C_x.$$

This strictly convex function attains its minimum, where $\nabla A_\sigma(\theta) = x$. Hence $\operatorname{Im}(D_\sigma^{\mathrm{MMSE}}) = U$, independently of σ.

Let $h_\sigma(z) = \frac{1}{2}\|z\|^2 - g_\sigma(z)$. For $x \in U$, let $z = (D_\sigma^{\mathrm{MMSE}})^{-1}(x)$. Since $\nabla h_\sigma(z) = x$, Fenchel equality gives

$$\phi_\sigma(x) = \langle z, x \rangle - \frac{1}{2}\|z\|^2 + g_\sigma(z) - \frac{1}{2}\|x\|^2. \tag{19}$$

Thus $U \subset \operatorname{int}(\operatorname{dom} \phi_\sigma)$. Compact support also gives $h_\sigma(z) \leq \sup_{u \in C} \langle z, u \rangle + C_\sigma$. Separation makes $h_\sigma^*(x) = +\infty$ for $x \notin C$; therefore $\operatorname{dom} \phi_\sigma \subset C$ and its interior equals U.

The Jacobian of $(z, \sigma) \mapsto (D_\sigma^{\mathrm{MMSE}}(z), \sigma)$ is invertible. The inverse is jointly smooth on $U \times (0, \infty)$, and differentiation of (19) gives $\nabla \phi_\sigma(x) = z - x$. Gaussian convolution makes g_σ and D_σ^{MMSE} jointly real analytic for $\sigma > 0$. The analytic inverse function theorem and (19) give the same regularity for ϕ_σ on $U \times (0, \infty)$. □

Learned Gradient-Step denoiser Assume a gradient-step denoiser $D_\sigma = \mathrm{Id} - \nabla g_\sigma$ with contractive ∇g_σ. Then D_σ is the proximal operator of a weakly convex potential ϕ_σ (Hurault et al., 2022b). In the following Proposition, we prove the different regularity properties of this potential. Note that parts of these proofs are already written in (Hurault et al., 2022b; Gribonval & Nikolova, 2020). We keep them here for completeness and clarity.

Proposition A.3 (Regularity of proximal gradient-step denoisers)**.** *Suppose Assumption 3.3 holds and, for some $\rho < 1$, $\mathrm{Lip}(\nabla g_\sigma) \leq \rho$ uniformly for $\sigma \in I_\epsilon$. Define $h_\sigma = \frac{1}{2}\|\cdot\|^2 - g_\sigma$ and $\phi_\sigma = h_\sigma^* - \frac{1}{2}\|\cdot\|^2$, as in (7). Then $D_\sigma = \mathrm{Id} - \nabla g_\sigma$ is a global $\mathcal{C}^1$ diffeomorphism of $\mathbb{R}^d$ and $D_\sigma = \mathrm{prox}_{\phi_\sigma}$. The potential ϕ_σ is finite, lower semicontinuous, jointly $\mathcal{C}^2$ and subanalytic in (x, σ) on $\mathbb{R}^d \times I_\epsilon$. With $z = D_\sigma^{-1}(x)$,*

$$\phi_\sigma(x) = g_\sigma(z) - \frac{1}{2}\|z - x\|^2, \qquad \nabla \phi_\sigma(x) = z - x, \qquad \partial_\sigma \phi_\sigma(x) = \partial_\sigma g_\sigma(z). \tag{20}$$

Moreover, $\nabla \phi_\sigma$ is globally Lipschitz with

$$\mathrm{Lip}(\nabla \phi_\sigma) \leq \frac{\rho}{1 - \rho}.$$

Finally, uniformly for $x \in \mathbb{R}^d$ and $\sigma, \sigma' \in I_\epsilon$,

$$\begin{aligned} \phi_\sigma(x) \geq g_\sigma(x) \geq c_0\|x\|^2 - C_0, \\ \|D_\sigma^{-1}(x)\| \leq \frac{\|x\| + B}{1 - \rho}, \qquad B := \sup_{\sigma \in I_\epsilon} \|D_\sigma(0)\| < \infty, \\ |\phi_{\sigma'}(x) - \phi_\sigma(x)| \leq C|\sigma' - \sigma|(1 + \|x\|^2). \end{aligned} \tag{21}$$

Moreover, on every compact $\mathcal{K} \subset \mathbb{R}^d$, the quantities $\nabla \phi_\sigma$, $\partial_\sigma \phi_\sigma$, and $\partial_\sigma \nabla \phi_\sigma$ are uniformly bounded on $\mathcal{K} \times I_\epsilon$.

Proof. For every x, the equation $D_\sigma(z) = x$ is $z = x + \nabla g_\sigma(z)$, whose right-hand side is a contraction. Hence D_σ is bijective, and the inverse function theorem gives a jointly $\mathcal{C}^1$ inverse in (x, σ). The bound on $\nabla^2 g_\sigma$ makes h_σ $(1-\rho)$-strongly convex, thus h_σ^* is finite and ϕ_σ is also finite and lower semi-continuous. For fixed z and any $u \in \mathbb{R}^d$,

$$\begin{aligned}\frac{1}{2}\|u - z\|^2 + \phi_\sigma(u) &= h_\sigma^*(u) - \langle z, u\rangle + \frac{1}{2}\|z\|^2 \\ &\geq -h_\sigma(z) + \frac{1}{2}\|z\|^2 = g_\sigma(z),\end{aligned}$$

where the inequality is Fenchel–Young: $h_\sigma(z) + h_\sigma^*(u) \geq \langle z, u\rangle$. Equality holds exactly when $u = \nabla h_\sigma(z) = D_\sigma(z)$. Thus $D_\sigma(z)$ attains the lower bound and is the unique minimizer. Consequently, $D_\sigma(z) = \operatorname{prox}_{\phi_\sigma}(z)$.

Let $z = D_\sigma^{-1}(x)$. Since $x = D_\sigma(z) = z - \nabla g_\sigma(z) = \nabla h_\sigma(z)$, Fenchel equality gives

$$h_\sigma^*(x) = \langle x, z\rangle - h_\sigma(z).$$

Substituting the definitions of h_σ and ϕ_σ,

$$\phi_\sigma(x) = \langle x, z\rangle - \frac{1}{2}\|z\|^2 + g_\sigma(z) - \frac{1}{2}\|x\|^2 = g_\sigma(z) - \frac{1}{2}\|z - x\|^2.$$

The chain rule yields

$$\nabla_x \phi_\sigma(x) = (J_x z)^\top \big(\nabla g_\sigma(z) - (z - x)\big) + (z - x) = z - x,$$

because $\nabla g_\sigma(z) = z - x$. Likewise, differentiating with respect to σ while holding x fixed gives

$$\partial_\sigma \phi_\sigma(x) = \partial_\sigma g_\sigma(z) + \big\langle \nabla g_\sigma(z) - (z - x), \partial_\sigma z\big\rangle = \partial_\sigma g_\sigma(z).$$

As z is jointly $\mathcal{C}^1$ and g_σ is jointly $\mathcal{C}^2$, these derivative identities also show that ϕ_σ is jointly $\mathcal{C}^2$.

Now, for $x, y \in \mathbb{R}^d$, write $u = D_\sigma^{-1}(x)$ and $v = D_\sigma^{-1}(y)$. Since $D_\sigma = \mathrm{Id} - \nabla g_\sigma$,

$$\|x - y\| \geq \|u - v\| - \|\nabla g_\sigma(u) - \nabla g_\sigma(v)\| \geq (1 - \rho)\|u - v\|.$$

Using $\nabla \phi_\sigma(x) = u - x = \nabla g_\sigma(u)$, we obtain

$$\|\nabla \phi_\sigma(x) - \nabla \phi_\sigma(y)\| \leq \rho\|u - v\| \leq \frac{\rho}{1 - \rho}\|x - y\|.$$

Since g_σ is jointly $\mathcal{C}^2$ and subanalytic, the map $(z, \sigma) \mapsto D_\sigma(z)$, is subanalytic. The graph of the jointly continuous inverse $(x, \sigma) \mapsto D_\sigma^{-1}(x)$ is obtained by exchanging the input and output coordinates in the graph of D_σ; it is therefore subanalytic. Finally, the first identity in (20) expresses $\phi_\sigma(x)$ in terms of g_σ and this inverse. Since the inverse is locally bounded, this composition is subanalytic.

Evaluating the conjugate defining $\phi_\sigma(x)$ at $z = x$ gives $\phi_\sigma(x) \geq g_\sigma(x)$, hence coercivity.

For $z = D_\sigma^{-1}(x)$,

$$\|x - D_\sigma(0)\| = \|z - (\nabla g_\sigma(z) - \nabla g_\sigma(0))\| \geq (1 - \rho)\|z\|,$$

which proves the inverse bound.

Assumption 3.3, (20) and (21) give $|\partial_\sigma \phi_\sigma(x)| \leq C(1 + \|x\|^2)$. Integrating this bound between σ and σ' proves the variation estimate.

Finally, $\partial_\sigma \nabla \phi_\sigma(x) = (I - \nabla^2 g_\sigma(z))^{-1} \partial_\sigma \nabla g_\sigma(z)$. The inverse bound sends $\mathcal{K} \times I_\epsilon$ into a compact set of z's, so this formula and joint continuity give all the asserted compact bounds. □

Common compact-set consequences. In the gradient-step case take $U = \mathbb{R}^d$, and in the MMSE case take $U = \operatorname{int} C$. On every compact convex $\mathcal{K} \subset U$, joint $\mathcal{C}^2$ regularity bounds $\partial_\sigma \phi_\sigma$, $\partial_\sigma \nabla \phi_\sigma$, and $\nabla \phi_\sigma$ uniformly on $\mathcal{K} \times I_\epsilon$. For MMSE, joint continuity also maps every compact set of denoiser inputs, uniformly over $\sigma \in I_\epsilon$, into a compact subset of U.

B CONVERGENCE PROOFS

B.1 PROOF OF THEOREM 3.4

Proof. Subtracting a lower bound from f, which does not change the iterates, we may assume $f \geq 0$.

We want to control $F_{\sigma_{k+1}}(x_{k+1}) - F_{\sigma_k}(x_k)$. Applying the descent lemma to the objective F_{σ_k} used at iteration k gives, with $L = L_f + \lambda L_g$,

$$\begin{aligned} F_{\sigma_k}(x_{k+1}) &\leq F_{\sigma_k}(x_k) + \langle \nabla F_{\sigma_k}(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2 \\ &= F_{\sigma_k}(x_k) - \gamma\left(1 - \frac{L\gamma}{2}\right)\|\nabla F_{\sigma_k}(x_k)\|^2 \\ &\leq F_{\sigma_k}(x_k) - a\gamma\|\nabla F_{\sigma_k}(x_k)\|^2, \qquad a := 1 - \frac{L\gamma}{2} > 0. \end{aligned} \tag{22}$$

We then prove that the iterates remain bounded. Recall the uniform coercivity from Assumption 3.3:

$$g_\sigma(x) \geq c_0\|x\|^2 - C_0 \tag{23}$$

Set

$$V_k(x) = 1 + \lambda C_0 + F_{\sigma_k}(x).$$

Then (23) gives

$$V_k(x) \geq 1 + \lambda c_0\|x\|^2. \tag{24}$$

Moreover, (23) also implies

$$1 + \|x\|^2 \leq C\big(1 + C_0 + g_\sigma(x)\big). \tag{25}$$

The mean-value theorem and the bound $|\partial_\sigma g_\sigma(x)| \leq M_g(1 + \|x\|^2)$ from Assumption 3.3 give

$$|g_{\sigma'}(x) - g_\sigma(x)| \leq M_g|\sigma' - \sigma|(1 + \|x\|^2) \leq C|\sigma' - \sigma|\big(1 + C_0 + g_\sigma(x)\big). \tag{26}$$

By (26),

$$V_{k+1}(x) \leq (1 + a_k)V_k(x), \qquad a_k = C|\sigma_{k+1} - \sigma_k|. \tag{27}$$

Consequently,

$$V_{k+1}(x_{k+1}) \leq (1 + a_k)V_k(x_{k+1}) \leq (1 + a_k)V_k(x_k). \tag{28}$$

The monotone noise schedule has finite total variation, hence $\prod_k(1 + a_k) < \infty$ and $\sup_k V_k(x_k) < \infty$. The uniform coercivity (24) then yields $\sup_k \|x_k\| < \infty$.

We can now control the change of objective between two successive noise levels. Let $\mathcal{K}$ be a compact set containing all the iterates. Since $(x, \sigma) \mapsto g_\sigma(x)$ is jointly $\mathcal{C}^2$, continuity of $\partial_\sigma \nabla g_\sigma(x)$ and compactness of $\mathcal{K} \times I_\epsilon$ imply that $\sigma \mapsto \nabla g_\sigma(x)$ is Lipschitz continuous uniformly for $x \in \mathcal{K}$, with constant $L^\sigma_{g,\mathcal{K}}$.

Since f does not depend on σ, the fundamental theorem of calculus and Assumption 3.3 give, for every $x \in \mathcal{K}$ and $\sigma, \sigma' \in I_\epsilon$,

$$\begin{aligned} |F_\sigma(x) - F_{\sigma'}(x)| &= \lambda|g_\sigma(x) - g_{\sigma'}(x)| \\ &= \lambda\left|\int_{\sigma'}^{\sigma} \partial_s g_s(x)\, ds\right| \\ &\leq \lambda M_g(1 + \|x\|^2)|\sigma - \sigma'| \\ &\leq C_{\mathcal{K}}|\sigma - \sigma'|, \end{aligned} \tag{29}$$

where

$$C_{\mathcal{K}} := \lambda M_g\left(1 + \sup_{x \in \mathcal{K}} \|x\|^2\right) < \infty.$$

Combining (22) and (29) at the point x_{k+1} yields

$$F_{\sigma_{k+1}}(x_{k+1}) \leq F_{\sigma_k}(x_k) - a\gamma\|\nabla F_{\sigma_k}(x_k)\|^2 + C_{\mathcal{K}}|\sigma_{k+1} - \sigma_k|. \tag{30}$$

Define

$$r_k := C_{\mathcal{K}} \sum_{j=k}^{\infty} |\sigma_{j+1} - \sigma_j|. \tag{31}$$

It is finite and tends to zero by the monotone convergence of (σ_k) in Assumption 3.1. Since $r_k - r_{k+1} = C_{\mathcal{K}}|\sigma_{k+1} - \sigma_k|$, (30) implies

$$F_{\sigma_{k+1}}(x_{k+1}) + r_{k+1} \le F_{\sigma_k}(x_k) + r_k - a\gamma\|\nabla F_{\sigma_k}(x_k)\|^2. \tag{32}$$

Thus $\big(F_{\sigma_k}(x_k) + r_k\big)$ is a decreasing sequence, and bounded below, it is therefore convergent. Because $r_k \to 0$, the sequence $\big(F_{\sigma_k}(x_k)\big)$ also converges, and summing (32) gives

$$\sum_{k=0}^{\infty} \|\nabla F_{\sigma_k}(x_k)\|^2 < \infty. \tag{33}$$

Let $F_\star := \inf_{\sigma \in I_\epsilon,\, x \in \mathcal{K}} F_\sigma(x) > -\infty$. More precisely, for every $K \ge 0$,

$$\min_{0 \le k \le K} \|\nabla F_{\sigma_k}(x_k)\|^2 \le \frac{F_{\sigma_0}(x_0) - F_\star + C_{\mathcal{K}} \sum_{k=0}^{K} |\sigma_{k+1} - \sigma_k|}{a\gamma(K+1)}. \tag{34}$$

The numerator is uniformly bounded in K, proving the rate stated in (9).

Since the stepsize is a fixed positive constant, (33) directly implies

$$\|\nabla F_{\sigma_k}(x_k)\| \longrightarrow 0. \tag{35}$$

Lipschitz continuity with respect to σ on $\mathcal{K}$ also gives

$$\|\nabla F_{\sigma_k}(x_k) - \nabla F_\epsilon(x_k)\| \le \lambda L^{\sigma}_{g,\mathcal{K}} |\sigma_k - \epsilon| \longrightarrow 0. \tag{36}$$

Combining this with (35) proves $\|\nabla F_\epsilon(x_k)\| \to 0$. If $x_{k_j} \to \bar{x}$ is any convergent subsequence, continuity of ∇F_ϵ gives $\nabla F_\epsilon(\bar{x}) = 0$, so every cluster point is critical.

Taking $\sigma' = \epsilon$ in (29) gives

$$|F_{\sigma_k}(x_k) - F_\epsilon(x_k)| \le C_{\mathcal{K}} |\sigma_k - \epsilon| \longrightarrow 0.$$

We already proved convergence of the first term; hence $F_\epsilon(x_k)$ converges.

Moreover, since

$$x_{k+1} - x_k = -\gamma \nabla F_{\sigma_k}(x_k)$$

and (35) yields $\|x_{k+1} - x_k\| \to 0$. Since (x_k) is bounded and every cluster point is critical for F_ϵ, we also obtain

$$\operatorname{dist}(x_k, \operatorname{crit} F_\epsilon) \to 0.$$

Rate for the terminal gradient. Suppose that $\sum_k |\sigma_k - \epsilon|^2 < \infty$, as in the rate claim following Theorem 3.4. Let

$$b_k := \nabla F_{\sigma_k}(x_k) - \nabla F_\epsilon(x_k) = \lambda\big(\nabla g_{\sigma_k}(x_k) - \nabla g_\epsilon(x_k)\big), \tag{37}$$

so that $x_{k+1} = x_k - \gamma(\nabla F_\epsilon(x_k) + b_k)$ and

$$\|b_k\| \le \lambda L^{\sigma}_{g,\mathcal{K}} |\sigma_k - \epsilon|. \tag{38}$$

Let $\delta := 2 - L\gamma > 0$, $c := \delta/4$, and $C := 1 + 1/\delta$. Using the descent lemma on F_ϵ gives

$$\begin{aligned} F_\epsilon(x_{k+1}) \le F_\epsilon(x_k) &- \gamma\left(1 - \frac{L\gamma}{2}\right)\|\nabla F_\epsilon(x_k)\|^2 \\ &- \gamma(1 - L\gamma)\langle \nabla F_\epsilon(x_k), b_k\rangle + \frac{L\gamma^2}{2}\|b_k\|^2. \end{aligned} \tag{39}$$

Since $1 - L\gamma/2 = \delta/2$, $|1 - L\gamma| \le 1$, and

$$\|\nabla F_\epsilon(x_k)\|\,\|b_k\| \le \frac{\delta}{4}\|\nabla F_\epsilon(x_k)\|^2 + \frac{1}{\delta}\|b_k\|^2,$$

we obtain

$$F_\epsilon(x_{k+1}) \le F_\epsilon(x_k) - c\gamma\|\nabla F_\epsilon(x_k)\|^2 + C\gamma\|b_k\|^2. \tag{40}$$

The last terms are summable by (38). Since F_ϵ is bounded below, summation gives

$$\min_{0 \le k \le K} \|\nabla F_\epsilon(x_k)\|^2 \le \frac{F_\epsilon(x_0) - \inf F_\epsilon + C\gamma\lambda^2 (L^{\sigma}_{g,\mathcal{K}})^2 \sum_{k=0}^{K} |\sigma_k - \epsilon|^2}{c\gamma(K+1)}. \tag{41}$$

Point convergence. Finally, for convergence of the sequence, we apply the inexact Kurdyka–Łojasiewicz theorem of Ochs (2019, Theorem 10).

We need to verify Ochs (2019, Assumption H). Under the additional decay assumption of Theorem 3.4, (37) gives

$$e_k := \|\nabla F_{\sigma_k}(x_k) - \nabla F_\epsilon(x_k)\| = \mathcal{O}(k^{-\alpha}), \qquad \sum_k e_k < \infty, \qquad \sum_k e_k^2 < \infty. \tag{42}$$

Let $m := 1/\gamma - L/2 > 0$. Using the identity

$$\nabla F_\epsilon(x_k) = -\frac{x_{k+1} - x_k}{\gamma} - (\nabla F_{\sigma_k}(x_k) - \nabla F_\epsilon(x_k)) \tag{43}$$

in the descent lemma on F_ϵ gives

$$F_\epsilon(x_{k+1}) \le F_\epsilon(x_k) - \left(\frac{1}{\gamma} - \frac{L}{2}\right) \|x_{k+1} - x_k\|^2 + e_k \|x_{k+1} - x_k\|$$

Young's inequality

$$e_k \|x_{k+1} - x_k\| \le \frac{m}{2}\|x_{k+1} - x_k\|^2 + \frac{1}{2m} e_k^2 \tag{44}$$

gives

$$F_\epsilon(x_{k+1}) \le F_\epsilon(x_k) - \frac{m}{2}\|x_{k+1} - x_k\|^2 + \frac{1}{2m} e_k^2. \tag{45}$$

The last term is an additional error term compared to the sufficient decrease condition from Assumption H1 of Ochs (2019). We can however absorb it in a different Lyapunov function. Choose an even integer

$$\theta > \frac{2\alpha - 1}{2\alpha - 2},$$

and define

$$t_k := \left(\frac{\theta}{2m} \sum_{j=k}^{\infty} e_j^2\right)^{1/\theta}, \qquad \Phi(x, t) := F_\epsilon(x) + \frac{t^\theta}{\theta}.$$

Then (45) rewrites on $\Phi(x, t)$ as

$$\Phi(x_{k+1}, t_{k+1}) + \frac{m}{2}\|x_{k+1} - x_k\|^2 \le \Phi(x_k, t_k),$$

which is the sufficient-decrease condition in Assumption H1 of Ochs (2019). Now, Lipschitz continuity of ∇F_ϵ gives

$$\begin{aligned} \|\nabla F_\epsilon(x_{k+1})\| &\le \|\nabla F_\epsilon(x_{k+1}) - \nabla F_\epsilon(x_k)\| + \|\nabla F_\epsilon(x_k)\| \\ &\le L\|x_{k+1} - x_k\| + \|\nabla F_\epsilon(x_k)\| \end{aligned}$$

From (43),

$$\begin{aligned} \|\nabla F_\epsilon(x_k)\| &= \left\| \frac{x_{k+1} - x_k}{\gamma} + (\nabla F_{\sigma_k}(x_k) - \nabla F_\epsilon(x_k)) \right\| \\ &\le \frac{1}{\gamma}\|x_{k+1} - x_k\| + e_k \end{aligned}$$

and thus

$$\|\nabla F_\epsilon(x_{k+1})\| \le \left(L + \frac{1}{\gamma}\right) \|x_{k+1} - x_k\| + e_k.$$

And in terms of $\Phi(x, t)$

$$\|\nabla \Phi(x_{k+1}, t_{k+1})\| \le \left(L + \frac{1}{\gamma}\right) \|x_{k+1} - x_k\| + e_k + t_{k+1}^{\theta - 1}.$$

Since $e_k = \mathcal{O}(k^{-\alpha})$,

$$t_k^{\theta-1} = \mathcal{O}\Big(k^{-(2\alpha-1)(\theta-1)/\theta}\Big).$$

The choice of θ therefore makes both (e_k) and $(t_k^{\theta-1})$ summable, and we thus verify the relative-error condition from Ochs (2019, Assumption H2). Moreover, since (x_k) is bounded and $t_k \to 0$, the sequence (x_k, t_k) is bounded and admits a convergent subsequence. Continuity of Φ makes this subsequence Φ-attentively convergent, which verifies (H3). The distance and parameter conditions follow by taking the distance $\|x_{k+1} - x_k\|$ and constant positive coefficients.

Finally, since F_ϵ is subanalytic (as f and g_σ are by assumption) and $t \mapsto t^\theta$ is polynomial, Φ is subanalytic and hence KL (Bolte et al., 2007). Since θ is even and F_ϵ is bounded below, Φ is also bounded below. Therefore, Ochs (2019, Theorem 10) yields

$$\sum_k \|x_{k+1} - x_k\| < \infty, \qquad x_k \longrightarrow x_\star, \qquad \nabla F_\epsilon(x_\star) = 0.$$

□

B.2 Proof of Theorem 3.6

Proof. In this proof, $F_\sigma = f + \lambda \tilde{g}_\sigma$ denotes the SNORE objective, with $\tilde{g}_\sigma$ defined in (10). We first check that its Gaussian average is well defined. Assumption 3.3 and continuity on I_ϵ give a constant C such that, uniformly in $\sigma \in I_\epsilon$,

$$\|\nabla g_\sigma(y)\| \le C + L_g\|y\|, \qquad |g_\sigma(y)| \le C(1 + \|y\|^2).$$

Thus Gaussian integration and differentiation in x give

$$\nabla \tilde{g}_\sigma(x) = \mathbb{E}_Z \nabla g_\sigma(x + \sigma Z) = \mathbb{E}_Z\big[x - D_\sigma(x + \sigma Z)\big]. \tag{46}$$

The uniform L_g-Lipschitz continuity of ∇g_σ implies that $\nabla \tilde{g}_\sigma$ is also globally L_g-Lipschitz. Hence ∇F_σ is globally $L = L_f + \lambda L_g$-Lipschitz for every $\sigma \in I_\epsilon$.

Let $\mathcal{F}_k$ contain the information available before drawing Z_{k+1}, and set

$$\xi_{k+1} := \lambda\big(\mathbb{E}_Z D_{\sigma_k}(x_k + \sigma_k Z) - D_{\sigma_k}(x_k + \sigma_k Z_{k+1})\big).$$

By (46), the SNORE update is

$$x_{k+1} = x_k - \gamma_k(\nabla F_{\sigma_k}(x_k) + \xi_{k+1}), \qquad \mathbb{E}[\xi_{k+1} \mid \mathcal{F}_k] = 0. \tag{47}$$

Moreover, $D_\sigma = \mathrm{Id} - \nabla g_\sigma$ is $(1 + L_g)$-Lipschitz. If Z' and Z are independent, we get, conditionally on $\mathcal{F}_k$,

$$\begin{aligned}
\mathbb{E}[\|\xi_{k+1}\|^2 \mid \mathcal{F}_k] &= \frac{\lambda^2}{2}\mathbb{E}_{Z,Z'}\|D_{\sigma_k}(x_k + \sigma_k Z) - D_{\sigma_k}(x_k + \sigma_k Z')\|^2 \\
&\le \frac{\lambda^2}{2}\sigma_k^2(1 + L_g)^2\mathbb{E}_{Z,Z'}\|Z - Z'\|^2 \\
&\le \lambda^2\sigma_0^2(1 + L_g)^2 d := C_\xi.
\end{aligned} \tag{48}$$

We next control the objective used at iteration k, as in the RED–GD proof. The descent lemma, (47), and conditional expectation yield

$$\mathbb{E}[F_{\sigma_k}(x_{k+1}) \mid \mathcal{F}_k] \le F_{\sigma_k}(x_k) - a\gamma_k\|\nabla F_{\sigma_k}(x_k)\|^2 + \frac{L}{2}C_\xi\gamma_k^2, \qquad a := 1 - \frac{L\bar{\gamma}}{2} > 0. \tag{49}$$

We now prove that the iterates remain bounded. After subtracting a lower bound from f, we may assume $f \ge 0$. Uniform coercivity of g_σ (Assumption 3.3) passes to its Gaussian average:

$$\tilde{g}_\sigma(x) \ge c_0\mathbb{E}\|x + \sigma Z\|^2 - C_0 \ge c_0\|x\|^2 - C_0. \tag{50}$$

Let

$$V_k(x) := 1 + \lambda C_0 + F_{\sigma_k}(x).$$

Then $V_k(x) \geq 1 + \lambda c_0 \|x\|^2$, and in particular $1 + \|x\|^2 \leq C V_k(x)$ for a constant independent of k. Using the chain rule,

$$|\partial_\sigma \tilde{g}_\sigma(x)| = |\mathbb{E}_Z \left[\partial_\sigma g_\sigma(x + \sigma Z) + \langle \nabla g_\sigma(x + \sigma Z), Z\rangle\right]| \tag{51}$$

Assumption 3.3 then gives:

$$|\partial_\sigma \tilde{g}_\sigma(x)| \leq C(1 + \|x\|^2). \tag{52}$$

Consequently, with the mean-value theorem, we get, for all x and $\sigma, \sigma' \in I_\epsilon$,

$$|F_{\sigma'}(x) - F_\sigma(x)| \leq C|\sigma' - \sigma|(1 + \|x\|^2). \tag{53}$$

Writing $a_k = C|\sigma_{k+1} - \sigma_k|$ with a larger constant if necessary, we obtain $V_{k+1}(x) \leq (1 + a_k)V_k(x)$. Combining this inequality with (49) gives

$$\mathbb{E}[V_{k+1}(x_{k+1}) \mid \mathcal{F}_k] \leq (1+a_k)V_k(x_k) - a\gamma_k \|\nabla F_{\sigma_k}(x_k)\|^2 + b_k, \qquad b_k := \frac{L}{2} C_\xi (1+a_k)\gamma_k^2. \tag{54}$$

The monotone schedule has finite total variation, so $\sum_k a_k < \infty$, and the stepsize condition gives $\sum_k b_k < \infty$. The Robbins–Siegmund almost-supermartingale theorem (Robbins & Siegmund, 1971, Theorem 1), applied to (54), shows that $V_k(x_k)$ converges almost surely to a finite random variable and, since $a > 0$,

$$\sum_k \gamma_k \|\nabla F_{\sigma_k}(x_k)\|^2 < \infty \quad \text{almost surely.} \tag{55}$$

The coercivity bound $1 + \|x\|^2 \leq C V_k(x)$ also gives $\sup_k \|x_k\| < \infty$ almost surely.

Taking expectations in (54) also gives $\sup_k \mathbb{E}[V_k(x_k)] < \infty$. Summing the same inequality through K yields

$$a \sum_{k=0}^{K} \gamma_k \mathbb{E}[\|\nabla F_{\sigma_k}(x_k)\|^2] \leq \mathbb{E}[V_0(x_0)] + \sup_k \mathbb{E}[V_k(x_k)] \sum_{k=0}^{K} a_k + \sum_{k=0}^{K} b_k. \tag{56}$$

The right-hand side is bounded independently of K; division by $\sum_{k=0}^{K} \gamma_k$ proves (12).

Now that we have bounded iterates, we can use local regularity in the noise level. Fix a sample path on which the above conclusions hold, and let $\mathcal{K}$ be a compact convex ball containing all its iterates.

Recall the parameter-derivative identity

$$\partial_\sigma \tilde{g}_\sigma(x) = \mathbb{E}[\partial_\sigma g_\sigma(x + \sigma Z) + \langle \nabla g_\sigma(x + \sigma Z), Z\rangle]\,.$$

Differentiating with respect to x and applying the Gaussian integration by parts $\nabla_x \mathbb{E}_Z[\phi(x+\sigma Z)] = \sigma^{-1}\mathbb{E}_Z[Z\phi(x + \sigma Z)]$, we get

$$\partial_\sigma \nabla \tilde{g}_\sigma(x) = \frac{1}{\sigma}\mathbb{E}_Z[Z\, \partial_\sigma g_\sigma(x + \sigma Z)] + \mathbb{E}_Z\left[\nabla^2 g_\sigma(x + \sigma Z) Z\right]. \tag{57}$$

The first expectation is bounded on $\mathcal{K} \times I_\epsilon$ by the quadratic growth of $\partial_\sigma g_\sigma$ (Assumption 3.3), and the second by $L_g \mathbb{E}\|Z\|$. Thus there is a finite constant $L^\sigma_{\tilde{g},\mathcal{K}}$ such that

$$\sup_{x \in \mathcal{K}} \|\nabla F_\sigma(x) - \nabla F_{\sigma'}(x)\| \leq \lambda L^\sigma_{\tilde{g},\mathcal{K}} |\sigma - \sigma'|, \qquad \sigma, \sigma' \in I_\epsilon. \tag{58}$$

In particular, $\|\nabla F_{\sigma_k}(x_k) - \nabla F_\epsilon(x_k)\| \to 0$.

Consider the martingale $M_n := \sum_{j=0}^{n-1} \gamma_j \xi_{j+1}$. Since $\sum \gamma_k^2 < \infty$ and (48), it converges almost surely. On a sample path where this holds, let

$$A_n := \sum_{j=n}^{\infty} \gamma_j \|\nabla F_{\sigma_j}(x_j)\|^2, \qquad R_n := \sup_{m \geq n} \|M_m - M_n\|.$$

Both tend to zero. For any $T > 0$, let $m(n)$ be the first index such that $\sum_{j=n}^{m(n)-1} \gamma_j \geq T$. For $n \leq l \leq m(n)$, summing the recursion (47) gives

$$x_l - x_n = -\sum_{j=n}^{l-1} \gamma_j \nabla F_{\sigma_j}(x_j) - (M_l - M_n).$$

Hence, by Cauchy–Schwarz,

$$\|x_l - x_n\| \le \sum_{j=n}^{l-1} \gamma_j \|\nabla F_{\sigma_j}(x_j)\| + R_n$$
$$\le \left(\sum_{j=n}^{l-1} \gamma_j\right)^{1/2} \left(\sum_{j=n}^{l-1} \gamma_j \|\nabla F_{\sigma_j}(x_j)\|^2\right)^{1/2} + R_n.$$

By the minimality of $m(n)$ and $\gamma_j \le \bar{\gamma}$, the first sum is at most $T + \bar{\gamma}$; the second is at most A_n. The bound is uniform in l, so

$$\max_{n \le l \le m(n)} \|x_l - x_n\| \le \sqrt{(T + \bar{\gamma}) A_n} + R_n \longrightarrow 0. \tag{59}$$

By the global L-Lipschitz continuity in x and (58),

$$\max_{n \le l \le m(n)} \|\nabla F_{\sigma_l}(x_l) - \nabla F_{\sigma_n}(x_n)\| \le L \max_{n \le l \le m(n)} \|x_l - x_n\| + 2\lambda L_{\bar{g},\mathcal{K}}^{\sigma} \sup_{j \ge n} |\sigma_j - \epsilon| \longrightarrow 0.$$

Suppose that $\|\nabla F_{\sigma_n}(x_n)\|$ does not tend to zero. Then there are $\eta > 0$ and arbitrarily large indices n such that $\|\nabla F_{\sigma_n}(x_n)\| \ge \eta$. The preceding uniform gradient estimate shows that, for every sufficiently large such n,

$$\max_{n \le l \le m(n)} \|\nabla F_{\sigma_l}(x_l) - \nabla F_{\sigma_n}(x_n)\| \le \frac{\eta}{2}.$$

The triangle inequality therefore gives $\|\nabla F_{\sigma_l}(x_l)\| \ge \eta/2$ for every $n \le l \le m(n)$. By the definition of $m(n)$,

$$A_n \ge \sum_{j=n}^{m(n)-1} \gamma_j \|\nabla F_{\sigma_j}(x_j)\|^2 \ge \frac{\eta^2}{4} \sum_{j=n}^{m(n)-1} \gamma_j \ge \frac{T\eta^2}{4},$$

contradicting $A_n \to 0$. Thus $\|\nabla F_{\sigma_k}(x_k)\| \to 0$, and (58) gives $\|\nabla F_{\epsilon}(x_k)\| \to 0$.

Finally, we get from (53) $|F_{\sigma_k}(x_k) - F_{\epsilon}(x_k)| \to 0$. Since $V_k(x_k)$ converges, so does $F_{\epsilon}(x_k)$. Moreover, $\gamma_k \to 0$, $\nabla F_{\sigma_k}(x_k)$ is bounded, and convergence of M_n implies $\gamma_k \xi_{k+1} \to 0$. Hence $\|x_{k+1} - x_k\| \to 0$. Continuity of ∇F_{ϵ} makes every cluster point critical; boundedness and compactness of $\mathcal{K}$ yield $\operatorname{dist}(x_k, \operatorname{crit} F_{\epsilon}) \to 0$. □

B.3 Annealed ERED

We consider a probability distribution π supported on a subgroup of orthogonal transformations. The equivariant RED iterates (ERED) (Terris et al., 2024) write:

$$\text{(ERED)} \qquad x_{k+1} = x_k - \gamma_k \Big[\nabla f(x_k) + \lambda\big(x_k - Q_{k+1}^{\top} D_{\sigma_k}(Q_{k+1} x_k)\big)\Big], \qquad Q_{k+1} \overset{\text{i.i.d.}}{\sim} \pi. \tag{60}$$

For the conservative denoiser (4), set

$$\bar{g}_{\sigma}(x) = \mathbb{E}_Q[g_{\sigma}(Qx)], \qquad F_{\sigma}(x) = f(x) + \lambda \bar{g}_{\sigma}(x). \tag{61}$$

The compactness of $O(\mathbb{R}^d)$ and Assumption 3.3 justify differentiation under the expectation and give

$$\nabla \bar{g}_{\sigma}(x) = \mathbb{E}_Q\big[Q^{\top} \nabla g_{\sigma}(Qx)\big]. \tag{62}$$

The following theorem gives for ERED (60) the same convergence guarantees as Theorem 3.6 gives for SNORE.

Theorem B.1 (Convergence of ERED (60))**.** *Suppose Assumptions 3.1, 3.2, and 3.3 hold, and the transformations are sampled as above. For a deterministic starting point x_0, suppose the stepsizes satisfy*

$$0 < \gamma_k \le \bar{\gamma}, \qquad (L_f + \lambda L_g)\bar{\gamma} < 2, \qquad \sum_k \gamma_k = \infty, \qquad \sum_k \gamma_k^2 < \infty. \tag{63}$$

Then the iterates are almost surely bounded, $F_\epsilon(x_k)$ converges almost surely, and

$$\sum_k \gamma_k \|\nabla F_{\sigma_k}(x_k)\|^2 < \infty, \qquad \|\nabla F_\epsilon(x_k)\| \to 0, \qquad \operatorname{dist}(x_k, \operatorname{crit} F_\epsilon) \to 0 \quad \textit{almost surely}. \tag{64}$$

Moreover, $\|x_{k+1} - x_k\| \to 0$ almost surely and

$$\min_{0\le k\le K} \mathbb{E}\|\nabla F_{\sigma_k}(x_k)\|^2 = \mathcal{O}\Big((\textstyle\sum_{k=0}^K \gamma_k)^{-1}\Big). \tag{65}$$

Proof. The proof follows the same arguments as the proof of SNORE. Let $L := L_f + \lambda L_g$. Orthogonality and (62) give, for every $x, y \in \mathbb{R}^d$,

$$\|\nabla \bar g_\sigma(x) - \nabla \bar g_\sigma(y)\| \le \mathbb{E}_Q \|\nabla g_\sigma(Qx) - \nabla g_\sigma(Qy)\| \le L_g \|x - y\|.$$

Thus ∇F_σ is globally L-Lipschitz, uniformly in σ. Let $\mathcal{F}_k$ contain the information available before drawing Q_{k+1}, and set

$$\xi_{k+1} := \lambda\big(Q_{k+1}^\top \nabla g_{\sigma_k}(Q_{k+1}x_k) - \nabla \bar g_{\sigma_k}(x_k)\big).$$

Then ERED (60) becomes

$$x_{k+1} = x_k - \gamma_k(\nabla F_{\sigma_k}(x_k) + \xi_{k+1}), \qquad \mathbb{E}[\xi_{k+1} \mid \mathcal{F}_k] = 0. \tag{66}$$

The variance of this stochastic gradient depends on x_k. Continuity on I_ϵ gives $C_g := \sup_{\sigma\in I_\epsilon} \|\nabla g_\sigma(0)\| < \infty$. Since Q is orthogonal,

$$\|Q^\top \nabla g_\sigma(Qx)\| \le C_g + L_g\|x\|,$$

and conditional centering yields

$$\mathbb{E}[\|\xi_{k+1}\|^2 \mid \mathcal{F}_k] \le 2\lambda^2(C_g^2 + L_g^2\|x_k\|^2). \tag{67}$$

The descent lemma and $\mathbb{E}[\xi_{k+1} \mid \mathcal{F}_k] = 0$ now give

$$\mathbb{E}[F_{\sigma_k}(x_{k+1}) \mid \mathcal{F}_k] \le F_{\sigma_k}(x_k) - a\gamma_k\|\nabla F_{\sigma_k}(x_k)\|^2 + \frac{L}{2}\gamma_k^2 \mathbb{E}[\|\xi_{k+1}\|^2 \mid \mathcal{F}_k], \qquad a := 1 - \frac{L\bar\gamma}{2} > 0. \tag{68}$$

We prove boundedness of the iterates. Subtracting a lower bound from f, we may assume $f \ge 0$. Orthogonality and Assumption 3.3 imply

$$\bar g_\sigma(x) \ge c_0\|x\|^2 - C_0, \qquad |\partial_\sigma \bar g_\sigma(x)| = |\mathbb{E}_Q \partial_\sigma g_\sigma(Qx)| \le M_g(1 + \|x\|^2). \tag{69}$$

Set $V_k(x) := 1 + \lambda C_0 + F_{\sigma_k}(x)$. Then $V_k(x) \ge 1 + \lambda c_0\|x\|^2$, and hence $1 + \|x\|^2 \le CV_k(x)$ uniformly in k. The mean-value theorem therefore gives

$$|F_{\sigma'}(x) - F_\sigma(x)| \le \lambda M_g|\sigma' - \sigma|(1 + \|x\|^2), \qquad V_{k+1}(x) \le (1 + a_k)V_k(x), \tag{70}$$

where $a_k = C|\sigma_{k+1} - \sigma_k|$. By (67) and the coercivity bound, there is a deterministic C_ξ such that $\mathbb{E}[\|\xi_{k+1}\|^2 \mid \mathcal{F}_k] \le C_\xi V_k(x_k)$. Write $b_k = (L/2)C_\xi\gamma_k^2$ and $\alpha_k = (1 + a_k)(1 + b_k) - 1$. Combining (68) with (70) yields

$$\mathbb{E}[V_{k+1}(x_{k+1}) \mid \mathcal{F}_k] \le (1 + \alpha_k)V_k(x_k) - a\gamma_k\|\nabla F_{\sigma_k}(x_k)\|^2. \tag{71}$$

The monotone schedule and the stepsize assumptions give $\sum_k \alpha_k < \infty$. The Robbins–Siegmund almost-supermartingale theorem applied to (71) shows that $V_k(x_k)$ converges almost surely to a finite random variable and

$$\sum_k \gamma_k\|\nabla F_{\sigma_k}(x_k)\|^2 < \infty \quad \text{almost surely}. \tag{72}$$

Because $1 + \|x_k\|^2 \le CV_k(x_k)$, the iterates are almost surely bounded.

Taking expectations in (71) gives $\sup_k \mathbb{E}V_k(x_k) < \infty$. Summing the same inequality through K then yields

$$a\sum_{k=0}^K \gamma_k \mathbb{E}\big[\|\nabla F_{\sigma_k}(x_k)\|^2\big] \le \mathbb{E}\big[V_0(x_0)\big] + \sup_j \mathbb{E}\big[V_j(x_j)\big] \sum_{k=0}^K \alpha_k.$$

The right-hand side is bounded independently of K, so division by $\sum_{k=0}^{K}\gamma_k$ proves (65).

Fix a sample path on which the iterates are bounded and let $\mathcal{K}$ be a compact ball centered at the origin and containing them. Since $Q\mathcal{K}=\mathcal{K}$ for $Q\in O(d)$ and g_σ is jointly C^2,

$$L^{\sigma}_{g,\mathcal{K}} := \sup_{\substack{y\in\mathcal{K}\\ \sigma\in I_\epsilon}} \|\partial_\sigma \nabla g_\sigma(y)\| < \infty.$$

Using (62) and the mean-value theorem, we get

$$\sup_{x\in\mathcal{K}} \|\nabla F_\sigma(x) - \nabla F_{\sigma'}(x)\| \le \lambda L^{\sigma}_{g,\mathcal{K}} |\sigma-\sigma'|. \tag{73}$$

In particular, $\|\nabla F_{\sigma_k}(x_k) - \nabla F_\epsilon(x_k)\| \to 0$.

Consider the martingale $M_n := \sum_{j=0}^{n-1}\gamma_j\xi_{j+1}$. Its increments are orthogonal in L^2, and the expectation bound above gives

$$\sup_n \mathbb{E}\|M_n\|^2 \le C_\xi \sup_j \mathbb{E}V_j(x_j)\sum_j \gamma_j^2 < \infty.$$

Thus M_n converges almost surely. On a sample path where this holds, set

$$A_n := \sum_{j=n}^{\infty}\gamma_j\|\nabla F_{\sigma_j}(x_j)\|^2, \qquad R_n := \sup_{m\ge n}\|M_m - M_n\|.$$

Both tend to zero. Fix $T>0$, and let $m(n)$ be the first index such that $\sum_{j=n}^{m(n)-1}\gamma_j \ge T$. For $n\le l\le m(n)$, summing (66) and applying Cauchy–Schwarz gives

$$\begin{aligned}\|x_l - x_n\| &\le \sum_{j=n}^{l-1}\gamma_j\|\nabla F_{\sigma_j}(x_j)\| + R_n \\ &\le \left(\sum_{j=n}^{l-1}\gamma_j\right)^{1/2}\left(\sum_{j=n}^{l-1}\gamma_j\|\nabla F_{\sigma_j}(x_j)\|^2\right)^{1/2} + R_n.\end{aligned}$$

The first sum is at most $T+\bar\gamma$ by the minimality of $m(n)$, and the second is at most A_n. Therefore

$$\max_{n\le l\le m(n)}\|x_l - x_n\| \le \sqrt{(T+\bar\gamma)A_n} + R_n \longrightarrow 0. \tag{74}$$

The global L-Lipschitz bound and (73) then imply

$$\max_{n\le l\le m(n)}\|h_l - h_n\| \le L\max_{n\le l\le m(n)}\|x_l - x_n\| + 2\lambda L^{\sigma}_{g,\mathcal{K}}\sup_{j\ge n}|\sigma_j - \epsilon| \longrightarrow 0.$$

If $\|h_n\|\not\to 0$, there are $\eta>0$ and arbitrarily large n with $\|h_n\|\ge\eta$. For every sufficiently large such n, the last estimate gives $\|h_l\|\ge\eta/2$ throughout the window, so

$$A_n \ge \sum_{j=n}^{m(n)-1}\gamma_j\|\nabla F_{\sigma_j}(x_j)\|^2 \ge \frac{\eta^2}{4}\sum_{j=n}^{m(n)-1}\gamma_j \ge \frac{T\eta^2}{4},$$

contradicting $A_n\to 0$. Hence $\|\nabla F_{\sigma_k}(x_k)\|\to 0$, and (73) gives $\|\nabla F_\epsilon(x_k)\|\to 0$.

Finally, (70) and boundedness imply $|F_{\sigma_k}(x_k) - F_\epsilon(x_k)|\to 0$. Since $V_k(x_k)$ converges, so does $F_\epsilon(x_k)$. Also $\gamma_k\to 0$, $(\nabla F_{\sigma_k}(x_k))$ is bounded, and convergence of M_n implies $\gamma_k\xi_{k+1}\to 0$. The recursion therefore gives $\|x_{k+1}-x_k\|\to 0$. Continuity of ∇F_ϵ makes every cluster point critical, and compactness of $\mathcal{K}$ gives $\operatorname{dist}(x_k, \operatorname{crit} F_\epsilon)\to 0$. □

B.4 Proof of Theorem 3.9

Proof. We first spell out the consequences of Assumption 3.8. In both cases, $D_\sigma = \operatorname{prox}_{\phi_\sigma}$ with the potential in (7). For a gradient-step denoiser, Proposition A.3 shows that D_σ is a global diffeomorphism and ϕ_σ is jointly $\mathcal{C}^2$ on $U\times I_\epsilon$, where $U=\mathbb{R}^d$. For an exact MMSE denoiser, Proposition A.2

gives the same joint regularity on the common open convex set $U = \mathrm{int}(\mathrm{conv}(\mathrm{supp}\, P_X))$, and D_σ maps $\mathbb{R}^d$ into U. In this case U is bounded because the prior has compact support. In either case, ϕ_ϵ is subanalytic on U by Appendix A. Set $\Psi_\sigma := \gamma F_\sigma = \gamma f + \phi_\sigma$; in particular, $\Psi_\epsilon = \gamma F_\epsilon$.

We begin with descent for the objective used at iteration k. Write $z_k = x_k - \gamma \nabla f(x_k)$. Since $x_{k+1} = \mathrm{prox}_{\phi_{\sigma_k}}(z_k)$, comparison with $x_k \in U$ gives

$$\phi_{\sigma_k}(x_{k+1}) \le \phi_{\sigma_k}(x_k) - \gamma\langle \nabla f(x_k), x_{k+1} - x_k\rangle - \frac{1}{2}\|x_{k+1} - x_k\|^2. \tag{75}$$

The descent lemma for f gives

$$\gamma f(x_{k+1}) \le \gamma f(x_k) + \gamma\langle \nabla f(x_k), x_{k+1} - x_k\rangle + \frac{\gamma L_f}{2}\|x_{k+1} - x_k\|^2. \tag{76}$$

Adding the two inequalities, we obtain

$$\Psi_{\sigma_k}(x_{k+1}) \le \Psi_{\sigma_k}(x_k) - m\|x_{k+1} - x_k\|^2, \qquad m := \frac{1 - \gamma L_f}{2} > 0. \tag{77}$$

We next prove boundedness, before using constants on a compact set. In the MMSE case, $x_{k+1} = D_{\sigma_k}(z_k) \in U$ for every k. We may start the descent argument at $k = 1$, since discarding the first update does not affect the asymptotic conclusions or the rate from indices 1 to K. For notational simplicity, we relabel this tail from $k = 0$. Thus the iterates used below lie in the bounded set U. In the gradient-step case, Proposition A.3 gives

$$\phi_\sigma(x) \ge g_\sigma(x) \ge c_0\|x\|^2 - C_0. \tag{78}$$

The same proposition gives the global variation estimate, uniformly for $x \in \mathbb{R}^d$ and $\sigma, \sigma' \in I_\epsilon$,

$$|\phi_{\sigma'}(x) - \phi_\sigma(x)| \le C|\sigma' - \sigma|(1 + \|x\|^2). \tag{79}$$

Subtracting a lower bound from f, we may assume $f \ge 0$. Set $V_k(x) := 1 + C_0 + \Psi_{\sigma_k}(x)$. From (78), $V_k(x) \ge 1 + \gamma f(x) + c_0\|x\|^2$, and hence $1 + \|x\|^2 \le (1 + c_0^{-1})V_k(x)$. Therefore,

$$V_{k+1}(x) - V_k(x) = \phi_{\sigma_{k+1}}(x) - \phi_{\sigma_k}(x) \le C(1 + c_0^{-1})|\sigma_{k+1} - \sigma_k|V_k(x).$$

Thus

$$V_{k+1}(x) \le (1 + a_k)V_k(x), \qquad a_k = C|\sigma_{k+1} - \sigma_k|,$$

after enlarging C. The sequence (a_k) is summable by monotonicity of (σ_k). Combining this bound with (77) and iterating yields $\sup_k V_k(x_k) < \infty$. The coercivity in (78) then gives $\sup_k \|x_k\| < \infty$.

We can now choose a compact convex set $\mathcal{K} \subset U$ containing all iterates. In the gradient-step case, take a sufficiently large closed ball. For MMSE, the inputs z_k are bounded because the iterates, stepsizes, and $\nabla f(x_k)$ are bounded. Let B_R be a closed ball containing them. Joint continuity of D_σ makes $S := \{D_\sigma(z) : z \in B_R,\ \sigma \in I_\epsilon\}$ a compact subset of U. Since U is convex, the convex hull of $S \cup \{x_0\}$ is a compact subset of U and contains every iterate; take it as $\mathcal{K}$.

Joint regularity of ϕ_σ on $\mathcal{K} \times I_\epsilon$ gives finite constants

$$L_{\phi,\mathcal{K}} := \sup_{\substack{x\in\mathcal{K}\\ \sigma\in I_\epsilon}} |\partial_\sigma \phi_\sigma(x)|, \qquad L_{\nabla\phi,\mathcal{K}} := \sup_{\substack{x\in\mathcal{K}\\ \sigma\in I_\epsilon}} \|\partial_\sigma \nabla\phi_\sigma(x)\|.$$

Continuity on the compact parameter set also gives the uniform lower bound

$$\Psi_\star := \inf_{x\in\mathcal{K},\, \sigma\in I_\epsilon} \Psi_\sigma(x) > -\infty.$$

For every $x \in \mathcal{K}$, the mean-value theorem thus gives

$$|\Psi_{\sigma_{k+1}}(x) - \Psi_{\sigma_k}(x)| \le L_{\phi,\mathcal{K}}|\sigma_{k+1} - \sigma_k| =: \delta_k. \tag{80}$$

Since $\sum_k \delta_k < \infty$ by monotonicity of (σ_k), combining this bound at x_{k+1} with (77) yields

$$\Psi_{\sigma_{k+1}}(x_{k+1}) \le \Psi_{\sigma_k}(x_k) - m\|x_{k+1} - x_k\|^2 + \delta_k. \tag{81}$$

Define $r_k := \sum_{j=k}^{\infty} \delta_j$. Then $\Psi_{\sigma_k}(x_k) + r_k$ is nonincreasing and bounded below, so it converges. As $r_k \to 0$, $\Psi_{\sigma_k}(x_k)$ converges, and summing (81) gives

$$m \sum_{k=0}^{K} \|x_{k+1} - x_k\|^2 \le \Psi_{\sigma_0}(x_0) - \Psi_\star + \sum_{k=0}^{K} \delta_k. \tag{82}$$

Hence $\sum_k \|x_{k+1} - x_k\|^2 < \infty$ and $\|x_{k+1} - x_k\| \to 0$. Moreover,

$$|\Psi_{\sigma_k}(x_k) - \Psi_\epsilon(x_k)| \le L_{\phi,\mathcal{K}} |\sigma_k - \epsilon| \longrightarrow 0.$$

Thus $F_\epsilon(x_k) = \Psi_\epsilon(x_k)/\gamma$ converges.

We next prove asymptotic stationarity, as in the RED–GD proof (Theorem 3.4). First-order optimality of the proximal step gives

$$x_{k+1} - x_k + \gamma \nabla f(x_k) + \nabla \phi_{\sigma_k}(x_{k+1}) = 0.$$

Consequently,

$$\nabla \Psi_{\sigma_k}(x_{k+1}) = \gamma\big(\nabla f(x_{k+1}) - \nabla f(x_k)\big) - (x_{k+1} - x_k), \quad \|\nabla \Psi_{\sigma_k}(x_{k+1})\| \le (1 + \gamma L_f)\|x_{k+1} - x_k\|. \tag{83}$$

Moreover, the change of objective from iteration $k-1$ to k satisfies

$$\|\nabla \Psi_{\sigma_k}(x_k) - \nabla \Psi_{\sigma_{k-1}}(x_k)\| \le L_{\nabla\phi,\mathcal{K}} |\sigma_k - \sigma_{k-1}| =: b_k.$$

Since (σ_k) has finite total variation, $\sum_k b_k^2 < \infty$. Therefore, for $k \ge 1$,

$$\|\nabla F_{\sigma_k}(x_k)\| = \frac{1}{\gamma} \|\nabla \Psi_{\sigma_k}(x_k)\| \le \frac{(1 + \gamma L_f)\|x_k - x_{k-1}\| + b_k}{\gamma}. \tag{84}$$

Together with (82), this proves $\sum_{k \ge 1} \|\nabla F_{\sigma_k}(x_k)\|^2 < \infty$. In particular $\|\nabla F_{\sigma_k}(x_k)\| \to 0$, and

$$\min_{1 \le k \le K} \|\nabla F_{\sigma_k}(x_k)\|^2 \le \frac{1}{K} \sum_{k=1}^{\infty} \|\nabla F_{\sigma_k}(x_k)\|^2 = \mathcal{O}(K^{-1}). \tag{85}$$

Finally, on $\mathcal{K}$, the terminal and moving gradients satisfy

$$\|\nabla F_{\sigma_k}(x) - \nabla F_\epsilon(x)\| \le \frac{L_{\nabla\phi,\mathcal{K}}}{\gamma} |\sigma_k - \epsilon| \longrightarrow 0 \quad \text{uniformly in } x \in \mathcal{K}. \tag{86}$$

Thus $\|\nabla F_\epsilon(x_k)\| \to 0$. If $x_{k_j} \to \bar{x}$, continuity of ∇F_ϵ gives $\nabla F_\epsilon(\bar{x}) = 0$, so every cluster point is critical.

Point convergence. Assume now the additional decay condition $|\sigma_k - \epsilon| = \mathcal{O}(k^{-\alpha})$ with $\alpha > 1$. Set

$$e_k := L_{\nabla\phi,\mathcal{K}} |\sigma_k - \epsilon|.$$

Then $\sup_{x \in \mathcal{K}} \|\nabla(\Psi_\epsilon - \Psi_{\sigma_k})(x)\| \le e_k = \mathcal{O}(k^{-\alpha})$, and $\sum_k e_k < \infty$, $\sum_k e_k^2 < \infty$. Since $\mathcal{K}$ is convex, integration along $[x_k, x_{k+1}]$ and (77) give

$$\Psi_\epsilon(x_{k+1}) \le \Psi_\epsilon(x_k) - m\|x_{k+1} - x_k\|^2 + e_k \|x_{k+1} - x_k\| \le \Psi_\epsilon(x_k) - \frac{m}{2}\|x_{k+1} - x_k\|^2 + \frac{e_k^2}{2m}. \tag{87}$$

As in the RED–GD proof, choose an even integer $\theta > (2\alpha - 1)/(2\alpha - 2)$ and define

$$t_k := \left(\frac{\theta}{2m} \sum_{j=k}^{\infty} e_j^2 \right)^{1/\theta}, \qquad \Phi(x, t) := \Psi_\epsilon(x) + \frac{t^\theta}{\theta}.$$

Then (87) becomes

$$\Phi(x_{k+1}, t_{k+1}) + \frac{m}{2}\|x_{k+1} - x_k\|^2 \le \Phi(x_k, t_k),$$

which verifies the sufficient-decrease condition (H1) of Ochs (2019, Assumption H). By (83),

$$\|\nabla \Phi(x_{k+1}, t_{k+1})\| \le (1 + \gamma L_f)\|x_{k+1} - x_k\| + e_k + t_{k+1}^{\theta - 1}.$$

The choice of θ ensures $t_k^{\theta-1} = \mathcal{O}(k^{-(2\alpha-1)(\theta-1)/\theta})$ is summable. Thus the relative-error condition (H2) holds. The bounded sequence (x_k, t_k) has a convergent subsequence; continuity of Φ on $U \times \mathbb{R}$ makes it Φ-attentively convergent, as required in (H3). The remaining distance and parameter conditions follow by taking $\|x_{k+1} - x_k\|$ and constant positive coefficients.

Finally, Ψ_ϵ is subanalytic on U, hence Φ is KL there. To use the global formulation of the KL theorem, choose a compact polytope P such that $\mathcal{K} \subset \operatorname{int} P \subset P \subset U$ and extend $\Phi|_{P\times\mathbb{R}}$ by $+\infty$ outside $P \times \mathbb{R}$. This extension is proper, lower semicontinuous, subanalytic and bounded below; it agrees with Φ near all iterates and cluster points. Therefore Ochs (2019, Theorem 10) gives

$$\sum_k \|x_{k+1} - x_k\| < \infty, \qquad x_k \to x_\star, \qquad \nabla\Psi_\epsilon(x_\star) = \gamma\nabla F_\epsilon(x_\star) = 0.$$

This proves finite length and convergence to a critical point of F_ϵ. □

B.5 SNOPnP and PnP–Flow

PnP–Flow (Martin et al., 2025) interpolates a gradient step with fresh Gaussian noise before applying the flow denoiser:

$$x_{k+1} = \widetilde{D}_{t_k}\big(t_k\big(x_k - \gamma\nabla f(x_k)\big) + (1-t_k)Z_{k+1}\big), \qquad t_k \in (0,1), \quad Z_{k+1} \sim \mathcal{N}(0, I_d). \tag{88}$$

Here $\widetilde{D}_t$ is a denoiser that takes as input the flow-matching time interpolation between image and noise. For $\sigma > 0$, set $s(\sigma) = 1/(1+\sigma)$ and $D_\sigma(u) = \widetilde{D}_{s(\sigma)}(s(\sigma)u)$. With this change of variable, D_σ is a proper image-space Gaussian denoiser that takes the rescaled image-space input u. Writing $\sigma_k = (1-t_k)/t_k$ gives $t_k = s(\sigma_k)$ and $t_k z + (1-t_k)Z = t_k(z + \sigma_k Z)$. Thus (88) becomes

$$x_{k+1} = D_{\sigma_k}\big(x_k - \gamma\nabla f(x_k) + \sigma_k Z_{k+1}\big).$$

which is exactly the SNOPnP algorithm proposed in Renaud et al. (2026).

Assuming that the rescaled denoiser is proximal, the PnP–Flow / SNOPnP update reads

$$x_{k+1} = \operatorname{prox}_{\phi_{\sigma_k}}\big(x_k - \gamma\nabla f(x_k) + \sigma_k Z_{k+1}\big), \qquad Z_{k+1} \sim \mathcal{N}(0, I_d).$$

Decoupling the two noise levels. In the above update, the injected noise has amplitude σ_k, the same level used by D_{σ_k}. In the proximal comparison below, this perturbation produces a term proportional to $\sigma_k^2\|Z_{k+1}\|^2$. Our descent argument in the proof applies when the sum of these stochastic terms is finite. If $\sigma_k \to \epsilon > 0$, this condition fails. We thus decouple the stochastic part and the noise level of the denoiser: we use a different schedule τ_k for each:

$$x_{k+1} = \operatorname{prox}_{\phi_{\sigma_k}}\big(x_k - \gamma\nabla f(x_k) + \tau_k Z_{k+1}\big), \qquad Z_{k+1} \sim \mathcal{N}(0, I_d). \tag{89}$$

The condition $\sum_k \tau_k^2 < \infty$ then controls the perturbations while $\sigma_k \to \epsilon > 0$ retains the terminal objective $F_\epsilon = f + \gamma^{-1}\phi_\epsilon$.

Theorem B.2 (Convergence of decoupled SNOPnP / PnP–Flow)**.** *Suppose Assumptions 3.1, 3.2, 3.3, and 3.8 hold. Use a constant stepsize $\gamma > 0$ with $\gamma L_f < 1$, and suppose $\sum_k \tau_k^2 < \infty$. Let $F_\sigma = f + \gamma^{-1}\phi_\sigma$. Almost surely, $\|\nabla F_\epsilon(x_k)\| \to 0$, $\|x_{k+1} - x_k\| \to 0$, and every cluster point is critical for F_ϵ. In particular,* $\operatorname{dist}(x_k, \operatorname{crit} F_\epsilon) \to 0$. *Moreover, almost surely,*

$$\min_{1\le k\le K} \|\nabla F_{\sigma_k}(x_k)\|^2 = \mathcal{O}\big(K^{-1}\big), \qquad K \ge 1. \tag{90}$$

Proof. The consequences of Assumption 3.8 follow from Propositions A.3 and A.2, as in the proof of Theorem 3.9. In particular, $D_\sigma = \operatorname{prox}_{\phi_\sigma}$, and ϕ_σ is jointly $\mathcal{C}^2$ on $U \times I_\epsilon$, where $U = \mathbb{R}^d$ for a contractive gradient-step denoiser and $U = \operatorname{int}(\operatorname{conv}(\operatorname{supp} P_X))$ for the exact MMSE denoiser. Set

$$\Psi_\sigma := \gamma F_\sigma = \gamma f + \phi_\sigma, \qquad e_{k+1} := \tau_k Z_{k+1}.$$

Since

$$\mathbb{E}\sum_k \|e_{k+1}\|^2 = d\sum_k \tau_k^2 < \infty,$$

we have $\sum_k \|e_{k+1}\|^2 < \infty$ almost surely. We work on a realization with this property; all subsequent bounds are pathwise.

We first establish descent for the moving objective. For a gradient-step denoiser the following comparison holds from $k = 0$; for MMSE it holds from $k = 1$, when $x_k \in U$. By definition of the proximal operator at input $z_k = x_k - \gamma \nabla f(x_k) + e_{k+1}$, we have

$$\phi_{\sigma_k}(x_{k+1}) \le \phi_{\sigma_k}(x_k) - \gamma\langle \nabla f(x_k), x_{k+1} - x_k\rangle - \frac{1}{2}\|x_{k+1} - x_k\|^2 + \langle e_{k+1}, x_{k+1} - x_k\rangle.$$

Adding the descent lemma for f yields

$$\Psi_{\sigma_k}(x_{k+1}) \le \Psi_{\sigma_k}(x_k) - m\|x_{k+1} - x_k\|^2 + \langle e_{k+1}, x_{k+1} - x_k\rangle, \qquad m := \frac{1 - \gamma L_f}{2} > 0. \tag{91}$$

Young's inequality then gives

$$\Psi_{\sigma_k}(x_{k+1}) \le \Psi_{\sigma_k}(x_k) - \frac{m}{2}\|x_{k+1} - x_k\|^2 + \frac{1}{2m}\|e_{k+1}\|^2. \tag{92}$$

We now prove boundedness of the iterates. In the MMSE case, every iterate from x_1 onward belongs to the bounded set $D_{\sigma_k}(\mathbb{R}^d) = U$ by Proposition A.2. Up to changing the start at $k = 1$, we can assume bounded iterates. In the gradient-step case, the estimates established under the same assumptions in the proof of Theorem 3.9 give

$$\phi_\sigma(x) \ge c_0\|x\|^2 - C_0, \qquad |\phi_{\sigma'}(x) - \phi_\sigma(x)| \le C|\sigma' - \sigma|(1 + \|x\|^2).$$

After subtracting a lower bound from f, assume $f \ge 0$. As in the PnP–PGD proof, set $V_k(x) := 1 + C_0 + \Psi_{\sigma_k}(x)$, so that $V_k(x) \ge 1 + \gamma f(x) + c_0\|x\|^2$ and

$$V_{k+1}(x) \le (1 + a_k)V_k(x), \qquad a_k = C|\sigma_{k+1} - \sigma_k|, \qquad \sum_k a_k < \infty.$$

Combining this bound with (92) gives

$$V_{k+1}(x_{k+1}) \le (1 + a_k)\left(V_k(x_k) + \frac{\|e_{k+1}\|^2}{2m}\right).$$

Since $\prod_k (1 + a_k) < \infty$ and $\sum_k \|e_{k+1}\|^2 < \infty$, iteration shows that $\sup_k V_k(x_k) < \infty$. Uniform coercivity then gives $\sup_k \|x_k\| < \infty$.

As in the PnP–PGD proof, choose a compact convex set $\mathcal{K} \subset U$ containing x_k for every $k \ge 1$. For gradient-step denoisers, a sufficiently large closed ball suffices. For MMSE, the proximal inputs z_k, the iterates and $\nabla f(x_k)$ are bounded. Joint continuity of D_σ makes the image of a closed ball containing all z_k, over $\sigma \in I_\epsilon$, a compact subset S of the open convex set U. Thus the convex hull of S is a compact subset of U and contains all iterates; we take it as $\mathcal{K}$.

Joint regularity on $\mathcal{K} \times I_\epsilon$ gives finite constants

$$L_{\phi,\mathcal{K}} := \sup_{\substack{x\in\mathcal{K}\\ \sigma\in I_\epsilon}} |\partial_\sigma \phi_\sigma(x)|, \quad L_{\nabla\phi,\mathcal{K}} := \sup_{\substack{x\in\mathcal{K}\\ \sigma\in I_\epsilon}} \|\partial_\sigma \nabla\phi_\sigma(x)\|.$$

Set $\Psi_\star := \inf_{x\in\mathcal{K},\, \sigma\in I_\epsilon} \Psi_\sigma(x) > -\infty$. For $x \in \mathcal{K}$,

$$|\Psi_{\sigma_{k+1}}(x) - \Psi_{\sigma_k}(x)| \le L_{\phi,\mathcal{K}}|\sigma_{k+1} - \sigma_k|.$$

For $k \ge 1$, define the summable error

$$\delta_k := L_{\phi,\mathcal{K}}|\sigma_{k+1} - \sigma_k| + \frac{\|e_{k+1}\|^2}{2m}.$$

Combining the above bound at x_{k+1} with (92) gives

$$\Psi_{\sigma_{k+1}}(x_{k+1}) \le \Psi_{\sigma_k}(x_k) - \frac{m}{2}\|x_{k+1} - x_k\|^2 + \delta_k. \tag{93}$$

With $r_k := \sum_{j=k}^\infty \delta_j$, the sequence $\Psi_{\sigma_k}(x_k) + r_k$ is nonincreasing for $k \ge 1$ and bounded below by $\Psi_\star$. Hence $\Psi_{\sigma_k}(x_k)$ converges and

$$\sum_{k=1}^\infty \|x_{k+1} - x_k\|^2 < \infty. \tag{94}$$

In particular, $\|x_{k+1} - x_k\| \to 0$. Moreover,

$$|\Psi_{\sigma_k}(x_k) - \Psi_\epsilon(x_k)| \le L_{\phi,\mathcal{K}}|\sigma_k - \epsilon| \longrightarrow 0,$$

so $F_\epsilon(x_k) = \Psi_\epsilon(x_k)/\gamma$ converges.

It remains to prove stationarity. First-order optimality of the noisy proximal step gives

$$x_{k+1} - x_k + \gamma\nabla f(x_k) + \nabla\phi_{\sigma_k}(x_{k+1}) - e_{k+1} = 0.$$

Consequently,

$$\|\nabla\Psi_{\sigma_k}(x_{k+1})\| \le (1+\gamma L_f)\|x_{k+1} - x_k\| + \|e_{k+1}\|. \tag{95}$$

At x_k for $k \ge 2$, the change of moving objective satisfies

$$\|\nabla\Psi_{\sigma_k}(x_k) - \nabla\Psi_{\sigma_{k-1}}(x_k)\| \le L_{\nabla\phi,\mathcal{K}}|\sigma_k - \sigma_{k-1}| =: b_k.$$

Thus, for $k \ge 2$,

$$\|\nabla F_{\sigma_k}(x_k)\| \le \frac{(1+\gamma L_f)\|x_k - x_{k-1}\| + \|e_k\| + b_k}{\gamma}.$$

The finite total variation of (σ_k) makes $\sum_k b_k^2 < \infty$. Together with (94) and $\sum_k \|e_{k+1}\|^2 < \infty$, this proves $\sum_{k\ge1}\|\nabla F_{\sigma_k}(x_k)\|^2 < \infty$. It follows that $\|\nabla F_{\sigma_k}(x_k)\| \to 0$ and (90). Moreover,

$$\sup_{x\in\mathcal{K}}\|\nabla F_{\sigma_k}(x) - \nabla F_\epsilon(x)\| \le \frac{L_{\nabla\phi,\mathcal{K}}}{\gamma}|\sigma_k - \epsilon| \longrightarrow 0.$$

Hence $\|\nabla F_\epsilon(x_k)\| \to 0$. Continuity of ∇F_ϵ makes every cluster point critical, and compactness of $\mathcal{K}$ gives $\operatorname{dist}(x_k, \operatorname{crit} F_\epsilon) \to 0$. □

C Experimental details

C.1 Denoisers: architectures, training and results

The four denoisers of Section 4 are gradient-step denoisers $D_\sigma = \mathrm{Id} - \nabla g_\sigma$ with

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

There are two training phases that we detail below: a *gradient-step phase*, which trains D_σ as a gradient-step denoiser and a *proximal phase* started from its output, which adds the penalty on $\|\nabla^2 g_\sigma\|_2$ to get a proximal denoiser.

Architectures and initialization. On natural images, N_σ is the DRUNet of Zhang et al. (2022) in the configuration of GS–DRUNet (Hurault et al., 2022a): a U-Net with four scales of 64, 128, 256 and 512 channels, two residual blocks per scale, the noise level given as a constant input channel, 17.0M parameters, and softplus activations instead of ELU, so that N_σ is $\mathcal{C}^\infty$ and the Hessian $\nabla^2 g_\sigma$ on which the proximal analysis rests is continuous (ELU is only $\mathcal{C}^1$). It is initialized with the weights of the published GS–DRUNet, trained for $\sigma \in [0, 50/255]$.

On FFHQ face images, N_σ is the U-Net of Dhariwal & Nichol (2021) trained by Choi et al. (2021) as an ϵ-prediction diffusion model on FFHQ at 256×256 and used by DiffPIR (Zhu et al., 2023): 128 base channels, one residual block per level, self-attention at one resolution, 93.6M parameters.

We normalize clean images in $[0, 1]$ intensity units. For a noisy input $y = x + \sigma z$, with $z \sim \mathcal{N}(0, I)$, conversion to the diffusion model's $[-1, 1]$ coordinates gives $2y - 1 = (2x - 1) + 2\sigma z$. To put this input in the model's variance-preserving form, we set $\bar\alpha = (1 + 4\sigma^2)^{-1}$ and the network receives $\sqrt{\bar\alpha}(2y - 1)$ at time $t(\sigma)$, and we use its noise prediction to define

$$N_\sigma(y) = y - \sigma\,\epsilon_\theta\big(\sqrt{\bar\alpha}(2y - 1),\, t(\sigma)\big).$$

We obtain $t(\sigma)$ from the continuous variance-preserving schedule, rather than using the discrete timestep lookup of the original DDPM implementation, to preserve regularity in σ.

Training data. Our DRUNet models are trained on the original training set of the published DRUNet from Zhang et al. (2022) (BSD400, the Waterloo Exploration Database, DIV2K and Flickr2K, 8 666 images), on random 128×128 crops taken at the native scale of the images with random horizontal and vertical flips; the last 16 images are held out for the validation of the training. DiffUNet models are trained on FFHQ at 128×128 with random horizontal flips.

Gradient-step training phase. The network minimizes the weighted L^2 denoising loss

$$\mathcal{L}_{\mathrm{GS}}(\theta) = \mathbb{E}_{x,\sigma,z}\Big[w(\sigma)\,\big\|D_\sigma(x+\sigma z) - x\big\|^2\Big], \qquad D_\sigma = \mathrm{Id} - \nabla g_\sigma, \quad w(\sigma) = \frac{\sigma^2+\sigma_{\mathrm{d}}^2}{\sigma^2\sigma_{\mathrm{d}}^2}, \tag{96}$$

where $\|\cdot\|^2$ is the mean squared error over the pixels, θ are the parameters of N_σ and w is the weighting proposed by Karras et al. (2022) with $\sigma_{\mathrm{d}} = 0.25$ (the standard deviation of the data for images in $[0,1]$), which equalizes the contribution of the noise levels: the loss is of order 1 at every σ. On natural images σ is log-uniform on $[0.002, 2]$, so that every decade of noise receives the same number of samples; on faces it follows the log-normal distribution of EDM, $\ln(2\sigma) \sim \mathcal{N}(-1.2, 1.2^2)$, clipped to $[0.002, 5]$.

The gradient $\nabla g_\sigma(y)$ is computed exactly by automatic differentiation, so a training step differentiates twice through the network (double backpropagation), as in Hurault et al. (2022a). Both phases use AdamW at learning rate 10^{-4} after a linear warm-up of 1000 steps. GS–DRUNet is trained for 10k steps at global batch 64. GS–DiffUNet is trained for 4k steps at global batch 64 (about four epochs of FFHQ).

Table 3 reports the PSNR performance of the original and finetuned models at different noise levels. It shows that the conservative gradient-step parametrization does not reduce denoising performance: initialized from the pretrained DiffUNet, GS–DiffUNet matches or exceeds its PSNR on FFHQ after only 4k training steps. Likewise, starting from the published GS–DRUNet trained only for $\sigma \le 50/255$, fine-tuning successively extends its performance to larger noise levels.

Proximal phase. Following Hurault et al. (2022b), the training of the proximal denoisers is done by starting from the gradient-step checkpoints, with the additional term in the loss:

$$\mathcal{L}_{\mathrm{Prox}}(\theta) = \mathcal{L}_{\mathrm{GS}}(\theta) + \mu\,\mathbb{E}\Big[\sum_{i=1}^{n_p} \omega_i\, \max\big(0, \|\nabla^2 g_{\sigma_i}(y_i)\|_2 - (1-\delta)\big)^2\Big], \tag{97}$$

where $0 < \delta < 1$ is the margin, μ the weight of the penalty and the weights ω_i chosen as a softmax over the spectral norms of the current batch: we find it advantageous to penalize stronger the higher spectral norms instead of the mean like Hurault et al. (2022b). Due to time and memory constraint, the penalty is evaluated on the first 8 images of each batch.

The spectral norm is estimated by the Lanczos method (Golub & Van Loan, 2013) with 20 iterations. The Lanczos iteration is warm-started from the dominant eigenvector obtained at the previous step (perturbed with Gaussian noise), and stopped early when the estimate moves by less than 10^{-3}. We find Lanczos iterations to converge faster than the power iteration of Hurault et al. (2022b). We set $\mu = 1$, $\delta = 0.2$, Prox–DRUNet runs 4k steps and Prox–DiffUNet runs 10k steps.

Table 3 reports the maximum spectral norm of the Hessian of g_σ and the PSNR denoising performance of the original and finetuned models, at different noise levels. On both datasets, proximal training sharply reduces the maximum Hessian spectral norm, bringing it below 1 for small noise levels, while keeping it close to 1 at larger noise levels. Handling the constraint on small noise levels is more important because the annealed schedules approach them in the late iterations, where the denoiser governs the limiting behavior of the algorithm.

Relaxation of the proximal denoisers. The spectral penalty in (97) encourages $\|\nabla^2 g_\sigma(x)\|_S < 1$ on training samples but does not certify this bound globally. Moreover, the validation measurements in Table 3 also slightly exceed 1 at some noise levels. In order to bring an extra safety margin to $\|\nabla^2 g_\sigma(x)\|_S < 1$, following Hurault et al. (2023, Section 2.3), we use at inference the relaxed denoiser

$$D_\sigma^\rho = (1-\rho)\,\mathrm{Id} + \rho D_\sigma = \mathrm{Id} - \rho \nabla g_\sigma$$

Table 3: Denoising PSNR and maximum spectral norm $\|\nabla^2 g_\sigma(x)\|_S$. GS and Prox denote the gradient-step and proximal denoisers.

FFHQ

σ	0.01	0.05	0.2	0.5	1	5
PSNR (dB)						
Pretrained DiffUNet (Choi et al., 2021)	44.61	35.54	28.60	24.45	21.49	15.48
GS–DiffUNet (fine-tuned to $\sigma \leq 5$)	44.54	35.69	28.87	24.87	22.08	15.75
Prox–DiffUNet	44.42	35.33	28.28	24.36	21.72	16.12
$\max_x \|\nabla^2 g_\sigma(x)\|_2$						
GS–DiffUNet	1.090	2.117	4.155	3.062	3.096	1.207
Prox–DiffUNet	0.902	0.969	0.982	1.008	1.025	1.035

CBSD68

σ	0.01	0.05	0.2	0.5	1	2
PSNR (dB)						
Published GS–DRUNet (Hurault et al., 2022a)	43.69	34.03	27.07	20.82	10.46	−0.66
GS–DRUNet (fine-tuned to $\sigma \leq 2$)	43.88	33.97	27.00	23.24	20.76	18.61
Prox–DRUNet	43.66	33.48	26.24	22.46	20.08	17.84
$\max_x \|\nabla^2 g_\sigma(x)\|_2$						
GS–DRUNet	1.202	1.287	1.628	1.524	1.800	1.574
Prox–DRUNet	0.882	0.919	0.951	0.977	0.989	1.040

The spectral norm of ρ If $\sup_x \|\nabla^2 g_\sigma(x)\|_2 < 1/\rho$ uniformly over the noise levels used, then D_σ^ρ is a proximal map. In practice, we use $\rho = 0.7$ for all our PnP experiments. This raises the sufficient threshold on $\|\nabla^2 g_\sigma\|_2$ from 1 to $1/\rho \simeq 1.43$, above every Prox denoisers validation norm reported in Table 3.

C.2 Algorithms, schedules and energy curves

Noise schedule. Every algorithm runs for $K = 500$ iterations with $\sigma_k = \epsilon + (\sigma_0 - \epsilon)r^k$, except on sparse-view tomography where the runs need $K = 1000$ to reach their floor. We set $\sigma_0 = 5$ on faces, $\sigma_0 = 1$ on natural images, and $r = 0.98$ on both; these choices are shared by all algorithms.

Stepsizes. The RED algorithms use $\gamma_0 = \gamma_{\text{rel}}/(L_f + \lambda L_g)$ with $\gamma_{\text{rel}} < 2$, where L_g bounds the Lipschitz constant of $\text{Id} - D_\sigma = \nabla g_\sigma$ over $\sigma \in [\epsilon, \sigma_0]$. Following the estimated Lipschitz constant given in Table 3, we use $L_g = 2.2$ for GS–DRUNet and $L_g = 4.2$ for GS–DiffUNet. PnP algorithms use $\gamma = c/L_f$ with $c < 1$. For SNORE and ERED, the stepsize is held constant at γ_0 until iteration step k_w and then follows the decrease

$$\gamma_k = \gamma_0 \big(1 + (k - k_w)_+/k_0\big)^{-q}, \qquad k_0 = K/4, \quad 1/2 < q \leq 1.$$

k_w is set either at 0 or at the first index at which $\sigma_k < 2\epsilon$, depending on the problem. Thus $\sum_k \gamma_k = \infty$ and $\sum_k \gamma_k^2 < \infty$. Delaying the decay can prevent the step from becoming small before a high initial noise level reaches its floor. This delayed decay is selected for SNORE on face demosaicing and super-resolution and for ERED on face super-resolution; everywhere else, the decay from $k = 0$ is kept. Table 5 shows the selected q parameter for each concerned problem. For decoupled SNOPnP (Theorem B.2), the injected noise follows

$$\tau_k = (\sigma_k - \epsilon) + \epsilon\big(1 + (k - k_w)_+/k_0\big)^{-q}, \qquad q > 1/2.$$

It initially tracks the denoiser level, then vanishes independently; $\sum_k \tau_k^2 < \infty$ because $\sigma_k - \epsilon$ decays geometrically. The same k_w convention is used in its search grid.

Plotted energies. In the different figures, we plot in the middle panel the evolution of the terminal objective along iterates $F_\epsilon(x_k)$. For SNORE, we estimate the potential $\tilde{g}_\sigma(x) = \mathbb{E}_Z[g_\sigma(x + \sigma Z)]$ using an average with four fresh samples at each iteration; these samples are not reused across iterations. For PnP, we need to invert the terminal denoiser at each iterate to evaluate $\phi_\epsilon(x_k)$ using

equation (20). We use damped fixed-point iteration warm-started at the denoiser input u_{k-1} of the algorithm's own step.

C.3 Further results on natural images

Natural images. Table 4 compares constant and annealed noise levels on CBSD68 for inpainting, tomography, and $\times 4$ super-resolution. RED methods (RED–GD, SNORE, ERED) use GS–DRUNet and PnP methods (PnP–PGD, SNOPnP) use Prox–DRUNet. For tomography, we use a parallel-beam Radon transform with 60 angles over $180°$, $\sigma_n = 0.01$, and filtered back-projection initialization. Annealing improves PSNR for every algorithm on all three problems. Annealing substantially improves performance for tomography. Gains are smaller for super-resolution: the measurements already constrain much of the coarse image structure that a large initial σ helps recover. Figures 5 and 6 show representative reconstructions and convergence curves for tomography and inpainting. Across these examples, the terminal objectives level off and the residuals decay, as predicted by the convergence analysis.

Table 4: PSNR (dB) on CBSD68 (128×128) for inpainting, tomography with 60 views, and super-resolution, same layout as Table 2: constant noise level σ_c versus annealed schedule $\sigma_k \to \epsilon$. RED algorithms (RED–GD, SNORE, ERED) and DPIR use GS–DRUNet, PnP algorithms (PnP–PGD, SNOPnP) use Prox–DRUNet, both finetuned to $\sigma \leq 2$. Best value per problem in bold.

CBSD68 (68 images), denoiser GS-DRUNet (RED algorithms) / Prox-DRUNet (PnP algorithms)						
	Inpainting		Tomography (60 views)		SR $\times 4$	
	constant σ	annealed	constant σ	annealed	constant σ	annealed
RED–GD	29.71	30.99	27.93	**29.02**	24.64	24.70
PnP–PGD	30.30	30.70	23.58	27.42	24.68	24.70
SNORE	29.93	30.93	27.89	28.98	24.64	**24.73**
SNOPnP / PnP–Flow	30.26	30.64	22.40	27.33	24.65	24.69
ERED	29.40	**31.07**	27.88	28.96	24.62	24.70
DPIR (GS-DRUNet)	29.19		28.79		24.38	

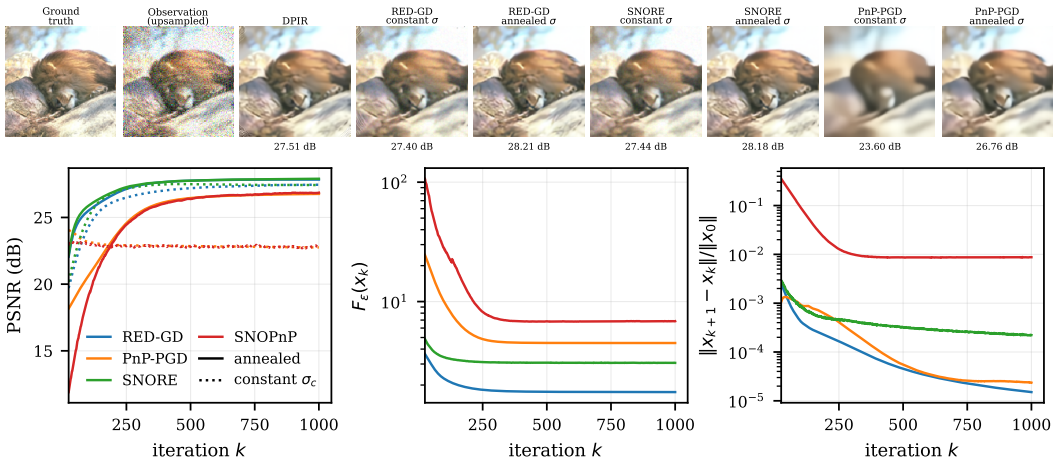

Figure 5: Tomography with 60 views on CBSD68, same layout as Figure 2; the observation is shown through its filtered back-projection.

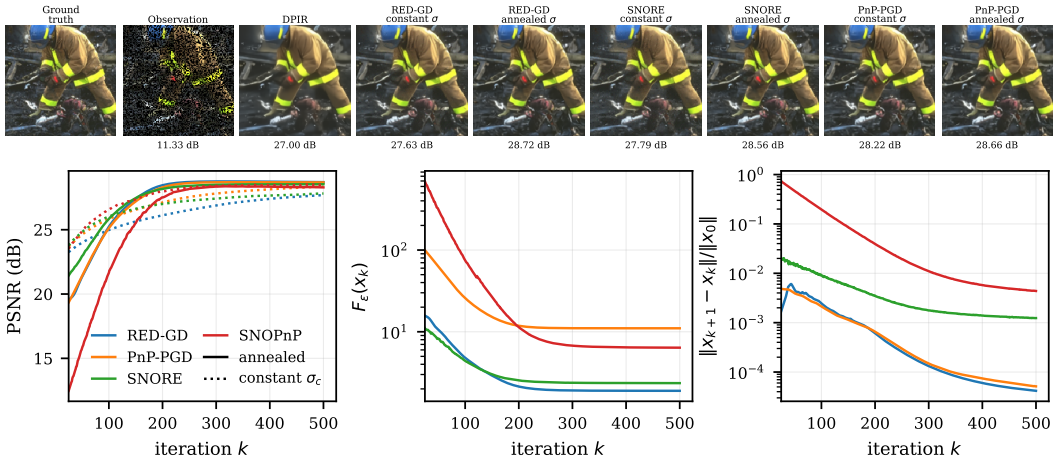

Figure 6: Inpainting of 50% random missing pixels on CBSD68, same layout as Figure 2.

Table 5: Selected hyper-parameters of the different algorithms for each problem and dataset (validation on 16 images)

dataset	problem	algorithm	ε (σ_c)	λ	step	q
FFHQ	Inpainting	RED–GD	0.014 (2)	0.3	$\gamma_{\text{rel}} = 1$	–
FFHQ	Inpainting	PnP–PGD	0.01 (0.04)	–	$c = 0.3$	–
FFHQ	Inpainting	SNORE	0.01 (2)	0.3	$\gamma_{\text{rel}} = 1$	0.55
FFHQ	Inpainting	SNOPnP	0.01 (0.16)	–	$c = 0.95$	1
FFHQ	Inpainting	ERED	0.01 (2)	0.5	$\gamma_{\text{rel}} = 1$	1
FFHQ	Demosaicing	RED–GD	0.01 (0.08)	0.1	$\gamma_{\text{rel}} = 1$	–
FFHQ	Demosaicing	PnP–PGD	0.01 (0.04)	–	$c = 0.5$	–
FFHQ	Demosaicing	SNORE	0.01 (0.04)	0.3	$\gamma_{\text{rel}} = 1$	0.55
FFHQ	Demosaicing	SNOPnP	0.01 (0.028)	–	$c = 0.5$	1
FFHQ	Demosaicing	ERED	0.01 (0.08)	1	$\gamma_{\text{rel}} = 1$	1
FFHQ	SR $\times 4$	RED–GD	0.08 (0.2)	0.01	$\gamma_{\text{rel}} = 1$	–
FFHQ	SR $\times 4$	PnP–PGD	0.04 (0.057)	–	$c = 0.3$	–
FFHQ	SR $\times 4$	SNORE	0.056 (0.11)	0.01	$\gamma_{\text{rel}} = 1$	1
FFHQ	SR $\times 4$	SNOPnP	0.08 (0.08)	–	$c = 0.95$	0.7
FFHQ	SR $\times 4$	ERED	0.08 (0.2)	0.01	$\gamma_{\text{rel}} = 1$	1
CBSD68	Inpainting	RED–GD	0.01 (0.03)	0.3	$\gamma_{\text{rel}} = 1$	–
CBSD68	Inpainting	PnP–PGD	0.02 (0.02)	–	$c = 0.95$	–
CBSD68	Inpainting	SNORE	0.01 (0.06)	0.3	$\gamma_{\text{rel}} = 1$	0.55
CBSD68	Inpainting	SNOPnP	0.01 (0.03)	–	$c = 0.7$	1
CBSD68	Inpainting	ERED	0.01 (0.03)	0.3	$\gamma_{\text{rel}} = 1$	1
CBSD68	Tomography	RED–GD	0.028 (0.057)	0.03	$\gamma_{\text{rel}} = 1$	–
CBSD68	Tomography	PnP–PGD	0.01 (0.04)	–	$c = 0.95$	–
CBSD68	Tomography	SNORE	0.04 (0.08)	0.03	$\gamma_{\text{rel}} = 1$	0.55
CBSD68	Tomography	SNOPnP	0.01 (0.057)	–	$c = 0.7$	0.7
CBSD68	Tomography	ERED	0.028 (0.057)	0.03	$\gamma_{\text{rel}} = 1$	0.55
CBSD68	SR $\times 4$	RED–GD	0.02 (0.04)	0.1	$\gamma_{\text{rel}} = 1$	–
CBSD68	SR $\times 4$	PnP–PGD	0.028 (0.04)	–	$c = 0.7$	–
CBSD68	SR $\times 4$	SNORE	0.04 (0.056)	0.03	$\gamma_{\text{rel}} = 1$	1
CBSD68	SR $\times 4$	SNOPnP	0.02 (0.04)	–	$c = 0.5$	1
CBSD68	SR $\times 4$	ERED	0.02 (0.04)	0.1	$\gamma_{\text{rel}} = 1$	0.55