MAP-CONVERGENT DENOISING IN NONCONVEX PNP FOR INVERSE PROBLEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Plug-and-Play (PnP) has become an increasingly popular framework for inverse problems, enabling optimization formulations with learning-based priors in proximal splitting methods. Score-based generative models have recently emerged as strong learned priors, but it remains challenging to efficiently utilize them in splitting-based PnP frameworks. Specifically, PnP frameworks often use denoisers as implicit priors in algorithms motivated by Maximum a Posteriori (MAP) estimation, but the annealed nature of score-based denoisers makes the theoretical connection between MAP and PnP challenging. Another difficulty comes from the geometry of the optimization iterates, which can deviate significantly from the noisy data that the score-based model is trained on. We consider MAP estimation with score-based PnP through primal-dual splitting and show that under suitable conditions, the primal-dual iterates converge to the stationary set of the corresponding nonconvex and nonsmooth MAP objective function. Additionally, we quantify the geometry mismatch introduced by the splitting method, while also revealing a tradeoff between convergence and geometrical alignment of the iterates. Our primal-dual approach yields updates which admit efficient closed-form solutions for structured nonlinear problems, including phase retrieval and sparse-view transmission computed tomography. Experiments on linear and nonlinear inverse problems show competitive reconstruction quality with fewer score evaluations and shorter runtimes compared to the state-of-the-art.

1 INTRODUCTION

Inverse problems arise when a latent signal must be recovered from indirect, incomplete, or corrupted measurements. Because many measurement operators discard information or amplify noise, reconstruction uses both the observations and prior knowledge about the class of plausible signals. Standard approaches consider reconstruction as a Maximum A Posteriori (MAP) problem that balances a data-fidelity term, which models the observation process, with a regularization term, which encodes prior knowledge. Since such MAP-problems often have exploitable structure, a standard algorithmic route is to use splitting methods, which separate the measurement-consistency part of the problem from the prior-correction part. Here, the way the structure is exploited depends on the splitting method. Alternating direction method of multipliers (ADMM)-type schemes are natural when auxiliary variables decouple the data and prior terms (Boyd et al., 2011), whereas primal-dual schemes such as primal-dual hybrid gradient (PDHG) exploit operator-induced structure through dual-space proximal updates (Chambolle & Pock, 2011).

Deep learning makes it possible to replace hand-crafted regularizers with priors learned from examples of the target signal class. Some approaches keep the variational structure explicit by learning the regularization functional itself (Lunz et al., 2018; Li et al., 2020), while generative models such as VAEs, GANs, and normalizing flows provide alternative ways of representing the distribution of plausible signals (Kingma & Welling, 2014; Goodfellow et al., 2014; Rezende & Mohamed, 2015). Recently, score-based models have been used to create even more powerful learned priors and instead of requiring a tractable density, they learn noise-conditioned scores of Gaussian-smoothed data distributions, giving local directions toward the data distribution across noise levels (Song & Ermon, 2019; Ho et al., 2020; Song et al., 2021). This makes them highly expressive priors for inverse problems, but it also exposes the main algorithmic bottleneck in classical splitting methods.

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.

The prior-correction step is a proximal update, which is often available for simple analytical priors but is generally unavailable for learned priors.

Plug-and-Play (PnP) methods address this bottleneck by replacing the intractable proximal update with a denoiser (Venkatakrishnan et al., 2013; Chan et al., 2017; Ryu et al., 2019). This idea makes pretrained denoisers usable inside iterative reconstruction algorithms without retraining them for each measurement model. With score-based diffusion models, the connection is especially appealing because, by Tweedie's formula (Efron, 2011), a score model defines a family of denoisers indexed by the noise level. Diffusion priors can therefore be coupled with a task-specific forward model at test time, either through plug-and-play restoration schemes, variable splitting, or posterior-sampling methods (Graikos et al., 2022; Zhu et al., 2023; Chung et al., 2023; Zhang et al., 2025).

Replacing the proximal prior update with a Tweedie denoiser introduces a gap between the MMSE denoising step and the MAP objective. Relating the resulting limit points to stationary points of that objective therefore requires further analysis (Pesme et al., 2025; Shumaylov et al., 2024). A further complication arises because score-based models are trained to denoise inputs produced by adding Gaussian noise to clean data. However, in optimization algorithms, the quantity passed to the denoiser accumulates effects from previous iterations through repeated inner solves or dual-variable corrections (Shrestha & Fu, 2026). The denoiser can therefore be queried away from the noisy-data geometry on which its score estimates are reliable.

In further reducing computational cost, the use of splitting methods such as PDHG can be beneficial for efficient iterations in structured nonlinear inverse problems. However, the theory behind PDHG-based methods often relies on convexity to control the dual proximal step (Chambolle & Pock, 2011). Extending the theory to the nonconvex domain and introducing PnP thus becomes a nontrivial task (Shumaylov et al., 2024; Sun et al., 2018).

1.1 Contributions

Motivated by the discussion above, we study MAP stationarity and denoiser-input geometry in annealed score-based PnP. We present our contributions below.

- **Nonconvex primal-dual PnP with MAP-stationary limits.** We propose a primal-dual score-PnP framework, compatible with a general class of nonconvex inverse problems. Under suitable conditions, we prove that limit points of our proposed algorithm are stationary points of the associated MAP objective function
- **Geometry mismatch in score-based PnP.** We quantify how the injected noise changes the conditional distribution of the denoiser input. Comparing this distribution with Gaussian kernels at the denoising level reveals a tradeoff between noise-scale alignment and convergence to the MAP-stationary set.
- **Efficient reconstruction for structured inverse problems.** We evaluate our approach on linear and nonlinear inverse problems. Our experiments show competitive reconstruction quality with fewer score evaluations and shorter runtimes compared to state-of-the-art methods.

2 Background and Related Work

2.1 Inverse Problems

We consider inverse problems in which the observations are generated as

$$y = f(Ax^\star) + \xi, \tag{1}$$

where $x^\star \in \mathbb{R}^d$ is the unknown signal, $A : \mathbb{R}^d \to \mathbb{R}^m$ is a known linear operator, $f : \mathbb{R}^m \to \mathcal{Y}$ is a known measurement map and ξ is measurement noise. This model covers standard linear imaging models like blurring and inpainting, as well as nonlinear measurement models such as phase retrieval (Candès et al., 2015) or nonlinear CT-imaging (Li et al., 2024; Fessler, 2000).

Once the observation y is fixed, we define a measurement-space data term $h(y; f(z)) := h(z)$ with $z \in \mathbb{R}^m$. The reconstruction problem is then written as

$$\min_x \lambda g(x) + h(Ax), \tag{2}$$

where g is the prior term. This is the problem class studied throughout the paper. For the discussions in Sections 2 and 3, we let $\lambda = 1$.

2.2 Related Work

Optimization-based diffusion priors. In variable-splitting methods, the prior often enters through a proximal step on g in (2),

$$z^{k+1} = \arg\min_z \left\{ g(z) + \frac{1}{2\tau}\|z - v^k\|^2 \right\} = \text{prox}_{\tau g}(v^k). \tag{3}$$

Classical plug-and-play methods replace the proximal step in (3) by a denoiser so that $z^{k+1} = D_{\sigma_k}(v)$ (Chan et al., 2017; Ryu et al., 2019; Venkatakrishnan et al., 2013). DiffPIR (Zhu et al., 2023) combines variable splitting with diffusion sampling and considers a score-based denoiser D_{σ_k}, namely

$$D_{\sigma_k}(v) := v + \sigma_k^2 \nabla \log p_{\sigma_k}(v) \approx v + \sigma_k^2 s_\theta(v, \sigma_k), \tag{4}$$

where $s_\theta(v, \sigma_k)$ is a learned score function and σ_k is an annealed noise schedule. However, the learned score-function is trained on Gaussian perturbations of the data, whereas the denoiser inputs in PnP frameworks generally are solutions of optimization problems related to the forward model in (1). This geometry mismatch can severely impact the quality of the denoiser outputs, particularly in the small-noise regime, where score fields increasingly encode manifold geometry over distributional properties (Stanczuk et al., 2024; Li et al., 2026b). This issue is particularly prevalent in primal-dual methods, where the denoiser input may be distorted by the introduction of dual variables. AC-DC ADMM (Shrestha & Fu, 2026) addresses the geometry mismatch problem by a two step procedure. First, the auto correction (AC) step adds Gaussian noise to the denoiser input to bring it closer to the noisy data manifolds. Next, the direct correction (DC) applies sampling correction towards the noisy data manifolds through Langevin dynamics. Such correction procedures can incur many expensive score evaluations. In addition, AC may significantly change the nature of the optimization iterates, but quantifying this change is nontrivial. This is further made clear in the primal method SNORE (Renaud et al., 2024), where AC is interpreted as a stochastic regularization. This regularization corresponds to a relaxation of the original MAP problem, showcasing that an empirically motivated heuristic may end up significantly changing the problem space of the algorithm.

MAP convergence in score-based PnP. To interpret the proximal step, consider the case in which the prior is represented by a density p_0 and $g(x) := -\log p_0(x)$. Then

$$\text{prox}_{\tau g}(v) = \arg\max_x p_0(x) \exp\left(-\frac{1}{2\tau}\|x - v\|^2 \right), \tag{5}$$

meaning the proximal operator is a local MAP update under a Gaussian centered at v. The score defines a minimum mean square error (MMSE) denoiser through Tweedie's identity. Indeed, let

$$x_0 \sim p_0, \qquad x_\sigma = x_0 + \sigma n, \qquad n \sim \mathcal{N}(0, I_d), \tag{6}$$

and let $p_\sigma = p_0 * \mathcal{N}(0, \sigma^2 I_d)$ denote the density of x_σ. Then Tweedie's identity gives

$$\mathbb{E}[x_0 \mid x_\sigma = v] = v + \sigma^2 \nabla \log p_\sigma(v), \tag{7}$$

which denotes the MMSE estimator under a noisy observation x_σ (Robbins, 1956; Efron, 2011). Hence, PnP with score denoisers replaces MAP estimation associated with the proximal operator in (5) with the more tractable MMSE estimation in (7).

In relating the resulting score-based PnP method back to the MAP problem, SNORE (Renaud et al., 2024) proposes to instead solve a regularized MAP problem, which is shown to be equivalent under a Gaussian prior. Recent work considers the connection between MAP and MMSE denoising in the convex setting (Pesme et al., 2025), showing that MMSE averaging with the proximal gradient methods allows (5) to be approximated at each iteration through an inner iterative process.

AC-DC-ADMM (Shrestha & Fu, 2026) establishes high-probability convergence to a fixed-point neighborhood with a constant penalty under a strongly convex likelihood, extending Ryu et al. (2019). Its Theorem 3 gives high-probability fixed-point convergence for differentiable, possibly nonconvex likelihoods with bounded gradients, using the adaptive penalty rule of Chan et al. (2017).

Table 1: Selected convergence guarantees. Checkmarks indicate permitted properties of the prior energy and negative log-likelihood. SNORE uses fixed denoiser noise. The AC-DC-ADMM row refers to Theorem 3 of (Shrestha & Fu, 2026), which uses an adaptive ADMM penalty and assumes a bounded likelihood gradient.

Work	Nonconvex prior	Nonconvex likelihood	Nonsmooth likelihood	Target
(Pesme et al., 2025)	×	×	×	MAP
(Renaud et al., 2024)	✓	✓	×	Relaxed MAP
(Shrestha & Fu, 2026)	✓	✓	×	Fixed point
This work	✓	✓	✓	MAP

Table 1 summarizes the convergence targets and assumptions of closely related denoiser-based methods.

PDHG in PnP. PDHG decouples the linear operator from the data-term h in (2) through the dual proximal step, which can yield computationally efficient iterations. However, the dual proximal step is with respect to the conjugate function h^*, which is an inherently convex object. This creates a difficulty with respect to structured nonconvex inverse problems, where phase retrieval is one prominent example. In (Sun et al., 2018), the PDHG framework is extended to the nonconvex setting under the Kurdyka-Lojasiewicz (KL) assumption. For PnP in PDHG, convergence results have been established in the convex setting (Suzuki et al., 2026) as well as extended to the case of weakly convex regularizers with convex losses (Shumaylov et al., 2024). It remains a challenge to establish the theoretical basis for PnP with PDHG in the nonconvex setting with score-based denoisers.

3 Proposed approach

In this section, we derive the primal-dual splitting from a convex saddle formulation of the MAP objective (2). We then show that the resulting splitting relates to the MAP objective even in the nonconvex setting. The prior proximal update is then replaced with a score-denoiser, where we quantify the resulting error between denoiser and proximal update under suitable conditions.

3.1 Nonconvex Primal-Dual Iterations

Consider the saddle formulation of the MAP objective (2),

$$\min_x \max_w \; g(x) + \langle Ax, w\rangle - h^*(w), \tag{8}$$

and a PDHG (Chambolle & Pock, 2011) iteration[1] is of the form

$$w^{k+1} = \operatorname{prox}_{\gamma h^*}\left(w^k + \gamma A x^k\right), \tag{9a}$$

$$v^{k+1} = x^k - \tau A^\top w^{k+1} \tag{9b}$$

$$x^{k+1} = \operatorname{prox}_{\tau g}\left(v^{k+1}\right). \tag{9c}$$

Using the Moreau identity for proper, lower semicontinuous, convex h (Bauschke & Combettes, 2017),

$$\operatorname{prox}_{\gamma h^*}(u) = u - \gamma \operatorname{prox}_{h/\gamma}(u/\gamma), \tag{10}$$

we can rewrite the dual step (9a) in terms of the proximal operator of h by introducing a variable z as follows

$$z^{k+1} := \operatorname{prox}_{h/\gamma}\left(Ax^k + \frac{1}{\gamma} w^k\right), \tag{11a}$$

$$w^{k+1} = w^k + \gamma\left(Ax^k - z^{k+1}\right). \tag{11b}$$

$$v^{k+1} := x^k - \tau A^\top w^{k+1}, \tag{11c}$$

$$x^{k+1} = \operatorname{prox}_{\tau g}(v^{k+1}). \tag{11d}$$

[1]Here we use the momentum parameter $\theta = 0$.

Remark 1. *The splitting procedure* (11a) - (11d) *has several useful properties in our setting. By decoupling z from the measurement operator A, the update* (11a) *admits closed form solutions in many practical inverse problems. Examples are CT image reconstruction and phase retrieval. Furthermore, since the input to* (11d) *directly depends on the sum $x^k - \tau A^\top w^{k+1}$, then suitable control of the dual variable w^{k+1} allows us to analyse denoiser geometry mismatch under this variable splitting. We present the detailed analysis in Section 4.*

Note that when h is nonconvex, the Moreau identity (10) does not hold in general. However, under some regularity conditions, we give a sufficient condition for the limit points of (11a) - (11d) to be stationary points of (2).

Proposition 1. *Let g and h be proper and lower semicontinuous, and let $(x^k, z^{k+1}, w^{k+1})_k$ be generated by* (11a)-(11d)*. Assume that w^k is bounded for all k and*

$$\|x^{k+1} - x^k\| \to 0, \quad \|w^{k+1} - w^k\| \to 0. \tag{12}$$

If at $x^k \to x^*$ it holds that

$$\partial g(x^*) + A^\top \partial h(Ax^*) \subseteq \partial(g + h \circ A)(x^*), \tag{13}$$

then x^* satisfies

$$0 \in \partial(g + h \circ A)(x^*). \tag{14}$$

Here ∂ denotes the limiting subdifferential (Rockafellar & Wets, 1998).

Proof. See Appendix B. □

The inclusion (13) is satisfied for a general class of nonconvex inverse problems. We may take, e.g., h as L_h-smooth. For nonsmooth h, the condition is satisfied when h is weakly convex and g is L_g-smooth.

3.2 Denoising and Stepsize Coupling

We replace the update in (11d) with a denoiser following the standard PnP framework (Chan et al., 2017; Ryu et al., 2019). The denoiser $D_{\sigma_k}(v) \approx v + \sigma_k^2 s_\theta(v, \sigma_k)$ uses a score function trained on Gaussian perturbations of data. Since the input v^{k+1} in (11c) may not be close to the training data, a standard trick is to add Gaussian noise to the denoiser input (Renaud et al., 2024; Li et al., 2026a; Shrestha & Fu, 2026), so that the update in (11d) is replaced by

$$x^{k+1} = D_{\sigma_k}(v^{k+1} + \rho_k n^k), \qquad n^k \sim \mathcal{N}(0, I_d). \tag{15}$$

The procedure is summarized in Algorithm 1. The Gaussian vectors are drawn independently at each iteration. The denoiser changes with the noise level σ_k, so its relation to the fixed MAP objective depends on the annealing schedule. The following proposition quantifies the discrepancy between the MMSE denoiser and the prior proximal update.

Proposition 2. *Suppose $p_0 > 0$ on $\mathbb{R}^d$ and $g_0 = -\log p_0$ is L_g-smooth. For every bounded set $K \subset \mathbb{R}^d$ and sufficiently small $\sigma > 0$,*

$$\sup_{v \in K} \|D_\sigma(v) - \mathrm{prox}_{\sigma^2 g_0}(v)\| = O(\sigma^3). \tag{16}$$

Proof. See Appendix C. □

Remark 2. *Note that we may consider $\tau = \frac{\sigma_k^2}{\lambda}$ for any $\lambda > 0$ which corresponds to the prior $\lambda g(x)$.*

Proposition 2 illustrates that the denoiser is a vanishing approximation of the local MAP problem in (11d) under the stepsize rule $\tau_k = \sigma_k^2$. In the next section, we use the insights of Propositions 1 and 2 to prove that the limit points of Algorithm 1 are stationary points of the MAP objective (2).

4 Theoretical Results

We combine the propositions from section 3 to establish MAP stationarity of Algorithm 1. We further extend the analysis to a scenario where the data is supported on a low-dimensional manifold as well as the case of a learned score function. Lastly, we consider how the injected noise into the denoiser affects the distribution of the denoiser input relative to the data on the support of p_0.

Algorithm 1 Primal-dual score-PnP

Require: observation y, dual parameter $\gamma > 0$, schedule $\{(\tau_k, \sigma_k, \rho_k)\}_{k=0}^{K-1}$

1: Initialize $x^0 \in \mathbb{R}^d$, $w^0 \in \mathbb{R}^m$
2: **for** $k = 0, 1, \ldots, K-1$ **do**
3: $\quad z^{k+1} \leftarrow \operatorname{prox}_{h/\gamma}\left(Ax^k + \frac{1}{\gamma} w^k\right)$
4: $\quad w^{k+1} \leftarrow w^k + \gamma(Ax^k - z^{k+1})$
5: $\quad v^{k+1} \leftarrow x^k - \tau_k A^\top w^{k+1}$
6: $\quad x^{k+1} \leftarrow D_{\sigma_k}(v^{k+1} + \rho_k n^k), \quad n^k \sim \mathcal{N}(0, I_d)$
7: **end for**
8: **return** x^K

4.1 MAP Stationarity

Assumption 1. *The prior distribution has a positive density p_0 on $\mathbb{R}^d$, and $g_0 = -\log p_0$ has an L_g-Lipschitz gradient. The primal iterates in Algorithm 1 are almost surely bounded.*

Assumption 2. *The function $h : \mathbb{R}^m \to \mathbb{R}$ is finite-valued and β-weakly convex.*

Assumptions 1 and 2 are standard in the literature (Shrestha & Fu, 2026; Renaud et al., 2024; Pesme et al., 2025). Note in particular that we do not require h to have a Lipschitz gradient. The iterate boundedness assumption is trivially satisfied when p_0 is supported on a compact set, which is natural in image applications where pixels are in bounded domains. In the more general case where p_0 has an unbounded support, as in Assumption 1, we show in Appendix E that the inverse problems considered in this work satisfy the boundedness assumption for Algorithm 1.

Let $F_\lambda = h \circ A + \lambda g_0$ and $\mathcal{C} = \{x \in \mathbb{R}^d \mid 0 \in A^\top \partial h(Ax) + \lambda \nabla g_0(x)\}$.

Theorem 1. *Under Assumptions 1–2, choose $\lambda > 0$ and $\gamma > 2\beta$. Let (σ_k) be positive, deterministic, and eventually nonincreasing, with $\sum_k \sigma_k^2 = \infty$ and $\sum_k \sigma_k^3 < \infty$. Let (ρ_k) be nonnegative and deterministic, with $\sum_k \rho_k^2 < \infty$, and set $\tau_k = \sigma_k^2/\lambda$. Then Algorithm 1 satisfies, almost surely,*

$$\operatorname{dist}(x^k, \mathcal{C}) \to 0, \qquad F_\lambda(x^k) \text{ converges}. \tag{17}$$

Proof. See Appendix D. □

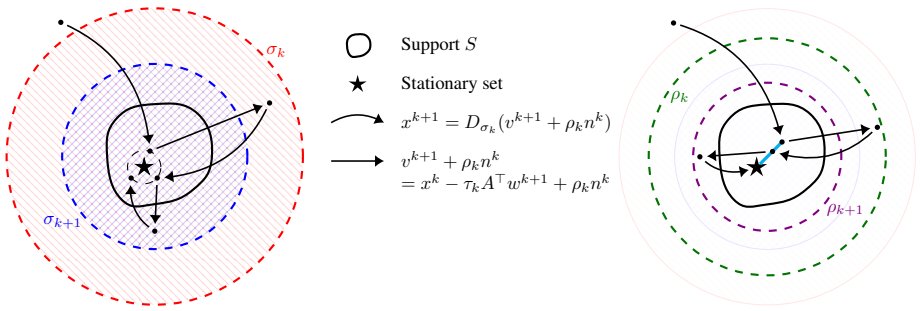
Theorem 1 identifies limit points of Algorithm 1 as stationary points of the MAP problem. The proof combines the estimate of Proposition 2 with the measurement optimality conditions and the stationary-point characterization in Proposition 1, using Moreau-envelope descent to control the annealing and injected noise. This gives stationarity for the associated MAP objective, complementing fixed-point convergence results for PnP methods in (Ryu et al., 2019; Shrestha & Fu, 2026).

Manifold hypothesis. The manifold hypothesis describes high-dimensional data as concentrated on or near a manifold of much lower dimension (Bengio et al., 2013). For images, correlations between pixels and shared factors of variation motivate such a lower-dimensional description. This perspective also underlies analyses of diffusion models with manifold-supported data (De Bortoli, 2022; Stanczuk et al., 2024). Appendix F considers a prior supported on an affine subspace. Under the remaining assumptions of Theorem 1, the ambient iteration has the same stationary-set and objective-value convergence conclusions for MAP estimation on that support.

Learned score. In practice, Algorithm 1 uses a trained approximation of the population score. Under the parameter and schedule conditions of Theorem 1, suppose the score error at the queried inputs is ε-bounded and the learned-score iterates are bounded almost surely. Then the sequence has, almost surely, at least one accumulation point $x^\star$ satisfying $\operatorname{dist}(0, \partial F_\lambda(x^\star)) \le \lambda\varepsilon$. We refer to Appendix G for more details.

4.2 Denoiser-input Geometry

At noise level σ_k, the learned score is trained on Gaussian corruptions of data with support $S = \operatorname{supp}(p_0)$. It is therefore of interest to quantify how the denoiser input $v^{k+1} + \rho_k n^k$ in Algorithm 1



(a) Noise scale $\rho_k = \sigma_k$. Iterates drift and need not converge to the stationary set.

(b) Noise scale ρ_k according to Theorem 1. Iterates converge to the stationary set.

Figure 1: Illustration of the geometry-convergence tradeoff of Theorem 2. When k is large, the noise scale $\rho_k = \sigma_k$ in (a) keeps $v^{k+1} + \rho_k n^k$ in the correct geometry with respect to the elements in S, while the iterates drift. In contrast, ρ_k is set according to Theorem 1 in (b) and the iterates converge to a stationary set, while $v^{k+1} + \rho_k n^k$ is closer to S than a typical Gaussian corrupted datapoint from S at noise level σ_k.

differs from Gaussian perturbed elements in S. The following result shows that adding noise can align the denoiser input with these elements.

Theorem 2. *Suppose Assumption 2 holds and the primal iterates are bounded with probability one. Choose $\lambda > 0$, $\gamma > 2\beta$, and $\tau_k = \sigma_k^2/\lambda$. Let (σ_k) and (ρ_k) be deterministic, with $\sigma_k > 0$, $\rho_k \geq 0$, and $\sigma_k \to 0$. We consider the cases where either (i) $x^k \in S$ eventually or (ii) S is a C^2 manifold with $p_0 > 0$ continuous and $\sum_k \sigma_k^3 < \infty$ with $\sigma_{k-1} = O(\sigma_k)$. Then there exist points $s^k \in S$ for which*

$$J_k := D_{\mathrm{KL}}\big(\mathcal{N}(s^k, \sigma_k^2 I_d) \,\big\|\, \mathcal{N}(v^{k+1}, \rho_k^2 I_d)\big)$$

satisfies, with probability one,

$$J_k \longrightarrow 0 \quad \text{if } \rho_k = \sigma_k \text{ for every } k,$$
$$J_k \not\longrightarrow 0 \quad \text{under the schedule conditions of Theorem 1.}$$

Proof. See Appendix H. □

Remark 3. *The two cases in Theorem 2 include priors on all of $\mathbb{R}^d$, as well as priors compatible with the manifold hypothesis, such as priors on low-dimensional or compact non-convex manifolds.*

Theorem 2 gives a distributional interpretation of the denoiser input. Although v^{k+1} is generated by the optimization loop, adding Gaussian noise can bring the conditional input distribution closer to a Gaussian perturbed element in $S = \mathrm{supp}(p_0)$.

Geometry-Convergence Tradeoff. Theorem 1 establishes convergence to the MAP-stationary set under its schedule conditions. Under the same conditions as Theorem 1, Theorem 2 precludes asymptotic noise-scale alignment. Choosing $\rho_k = \sigma_k$ instead gives $J_k \to 0$, but is incompatible with $\sum_k \sigma_k^2 = \infty$ and $\sum_k \rho_k^2 < \infty$. Indeed, we show in Appendix I that with added noise $\rho_k = \sigma_k$, Algorithm 1 can be asymptotically identified with subgradient Langevin dynamics at temperature $\lambda/2$. As such, we do not expect Algorithm 1 with noise scale $\rho_k = \sigma_k$ to converge to the stationary set. We illustrate this tradeoff in Figure 1.

5 Simulation Results

We evaluate our approach on two datasets, namely FFHQ (Karras et al., 2019) and the AAPM Low Dose CT Grand Challenge (Moen et al., 2021). For FFHQ, we use the pre-trained model in (Chung et al., 2023) and for the CT dataset we use the pre-trained model in (Shi et al., 2026). For the state-of-the-art comparisons, we consider representative diffusion-based inverse-problem solvers that cover both posterior-sampling and optimization-based approaches. For FFHQ, we compare against DPS (Chung et al., 2023), DAPS (Zhang et al., 2025), RED-DIFF (Mardani et al., 2024), DiffPIR (Zhu

Table 2: Inverse-problem reconstruction metrics. Panel (a) reports FFHQ reconstruction metrics over 100 images. Panel (b) reports sparse-view transmission CT results on the AAPM Low Dose CT Grand Challenge dataset. Best results are in **bold** while second-best results are in blue.

(a) FFHQ reconstruction metrics

Method	NFE↓	PSNR↑	SSIM↑	LPIPS↓
Superresolution (4×)				
Ours	**200**	**30.58**	**0.87**	0.16
AC-DC	1000	30.42	0.86	0.17
DAPS	1000	29.36	0.78	0.19
DPS	1000	26.83	0.75	**0.14**
RED-DIFF	1000	27.11	0.72	**0.14**
DiffPIR	200	27.86	0.79	0.24
DCDP	200	26.67	0.64	0.41
Random inpainting				
Ours	**100**	**33.42**	**0.90**	**0.09**
AC-DC	1000	32.05	**0.90**	0.13
DAPS	1000	30.73	0.80	0.16
DPS	1000	31.56	0.88	**0.09**
RED-DIFF	1000	26.51	0.72	0.14
DiffPIR	**100**	31.24	0.84	0.27
DCDP	400	31.89	0.84	0.21
Motion deblurring				
Ours	**200**	**32.74**	**0.90**	**0.13**
AC-DC	1000	32.27	0.89	**0.13**
DAPS	1000	31.65	0.84	0.14
DPS	1000	23.92	0.69	0.22
DiffPIR	200	24.35	0.72	0.34
Gaussian deblurring				
Ours	**200**	**30.34**	**0.86**	**0.17**
AC-DC	1000	30.14	0.85	0.18
DAPS	1000	29.61	0.79	0.18
DPS	1000	24.42	0.71	0.33
DiffPIR	200	28.73	0.81	0.25
DCDP	200	16.73	0.17	0.70
Box inpainting				
Ours	**200**	**25.43**	**0.87**	**0.11**
AC-DC	1000	24.03	0.86	0.14
DAPS	1000	24.69	0.75	0.18
DPS	1000	23.68	0.82	0.13
RED-DIFF	1000	20.94	0.67	0.18
DiffPIR	200	22.92	0.73	0.25
DCDP	400	22.05	0.71	0.26
Phase retrieval				
Ours	**500**	**28.38**	**0.80**	**0.20**
AC-DC	4000	27.20	0.78	0.23
DAPS	4000	26.81	0.73	0.24
DPS	1000	12.85	0.32	0.54
RED-DIFF	1000	21.66	0.47	0.40
DCDP	4000	20.40	0.57	0.42

(b) Sparse-view CT

Method	NFE↓	PSNR↑	SSIM↑
Ours	**200**	**33.50**	0.92
RED-DIFF	**200**	32.11	0.71
DPS	1000	32.35	0.78
DMPlug	3000	25.18	0.56
DDS	**200**	21.99	0.35

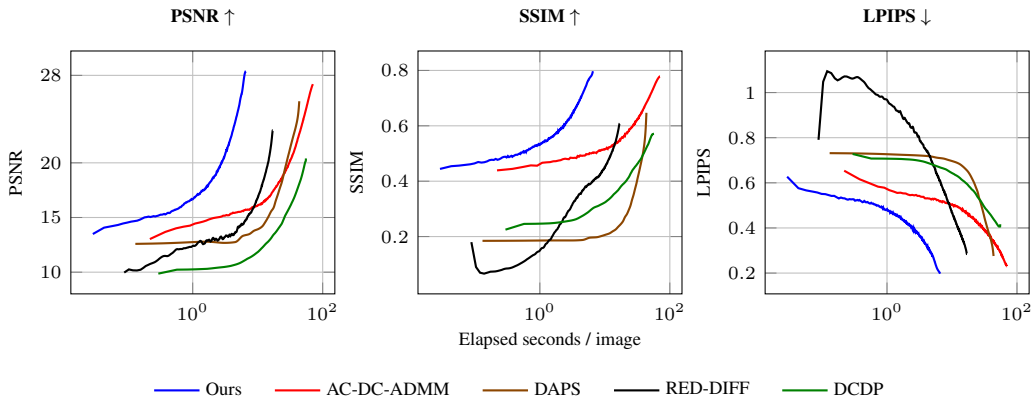


Figure 2: PSNR, SSIM and LPIPS for phase retrieval against elapsed time per image. The simulation corresponds to 100 FFHQ images.

et al., 2023), AC-DC-ADMM (Shrestha & Fu, 2026), and DCDP (Li et al., 2026a). For sparse CT-reconstruction, we compare against the DDS, DMPlug, DPS and RED-DIFF implementations in (Shi et al., 2026). Table 2 summarizes the FFHQ and CT results.

For the noise schedule, we consider a design inspired by Theorem 2. Initially, we let $\rho_k = \sigma_k$, matching the noise level for the geometrical alignment. We then switch to a schedule where ρ_k satisfies the conditions of Theorem 1 for the tail of the iterations. We refer to Appendix J for specific hyperparameter settings.

For the FFHQ experiment, Table 2(a) presents the number of function evaluations (NFE), Peak-Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM) and Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al., 2018) for each inverse problem. Specifically, NFE measures the number of score evaluations. In some cases, score evaluations may not be the only computational overhead due to, e.g., inner solvers. Hence, we complement Table 2(a) with Figure 2 showcasing the computation time per image on the nonlinear phase retrieval problem. We observe that our approach can achieve comparable or better results with substantially fewer score evaluations compared to baseline methods. Furthermore, Figure 2 confirms that the closed-form measurement-space updates enable efficient reconstruction both in terms of speed and quality compared to baselines.

For the CT experiment, Table 2(b) reports NFE, PSNR and SSIM on the sparse-view transmission CT task. We observe that our method obtains the best reconstruction quality among the tested methods, achieving higher PSNR and substantially higher SSIM than RED-DIFF and DPS. Figures 3 and 4 show qualitative results on FFHQ and sparse CT reconstruction. Our approach is able to recover both a representative reconstruction of the original image and also retain finer details.

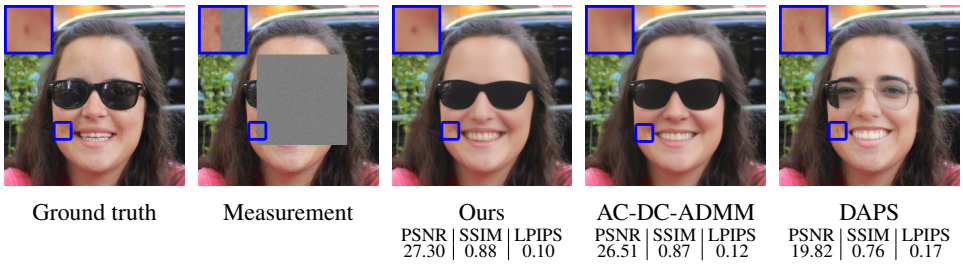


Figure 3: Inpainting with box constraint.

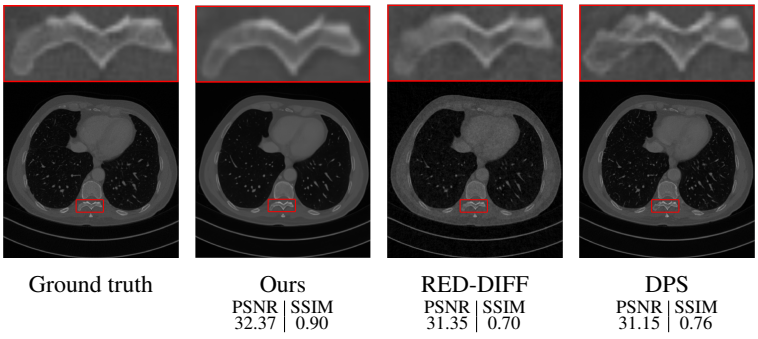


Figure 4: Sparse CT-reconstruction.

For additional simulation results, we refer to Appendix K.

6 Conclusions

We considered the connection between annealed MMSE denoising and MAP estimation in nonconvex and nonsmooth inverse problems. We proved that under suitable conditions, limits of the proposed algorithm are stationary points of the corresponding nonconvex MAP objective. We also showed that the geometry mismatch between optimization iterates and Gaussian perturbed data from the training data support can be quantified, revealing a tradeoff between noise level alignment and convergence. Empirically, the resulting method achieves comparable or better reconstruction quality across linear and nonlinear imaging tasks while using substantially fewer score evaluations for more difficult nonlinear tasks compared to several diffusion baselines.

AI USE STATEMENT

We have not used generative AI tools for generating synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose of refine hypotheses, design of provide feedback on research methodology or experiments, implement methods, assist with translation, clean and reformat dataset, support qualitative and thematic data analysis or interpreting results. Additionally, we used generative AI tools for editing the research paper to improve readability and for sourcing of some information. We have reviewed all AI-assisted work. The paper has been proofread by all authors and the information that has been sourced by AI has been critically evaluated. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Heinz H. Bauschke and Patrick L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, 2 edition, 2017.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013. doi: 10.1109/TPAMI.2013.50.

Dimitri P. Bertsekas and John N. Tsitsiklis. Gradient convergence in gradient methods with errors. *SIAM Journal on Optimization*, 10(3):627–642, 2000.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011. doi: 10.1561/2200000016.

Emmanuel J. Candès, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval from coded diffraction patterns. *Applied and Computational Harmonic Analysis*, 39(2):277–299, 2015. doi: 10.1016/j.acha.2014.09.004.

Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, 2011. doi: 10.1007/s10851-010-0251-1.

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.

Hyungjin Chung, Jeongsol Kim, Michael T. McCann, Marc L. Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *International Conference on Learning Representations*, 2023.

Damek Davis and Dmitriy Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019. doi: 10.1137/18M1178244.

Valentin De Bortoli. Convergence of denoising diffusion models under the manifold hypothesis. *Transactions on Machine Learning Research*, 2022. URL `https://openreview.net/forum?id=MhK5aXo3gB`.

Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011. doi: 10.1198/jasa.2011.tm11181.

Herbert Federer. Curvature measures. *Transactions of the American Mathematical Society*, 93(3):418–491, 1959.

Jeffrey A. Fessler. Statistical image reconstruction methods for transmission tomography. In Milan Sonka and J. Michael Fitzpatrick (eds.), *Handbook of Medical Imaging, Volume 2: Medical Image Processing and Analysis*, pp. 1–70. SPIE, 2000. doi: 10.1117/3.831079.ch1.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, pp. 2672–2680, 2014.

Alexandros Graikos, Esmeralda S Whitammer, Nebojsa Jojic, and Dimitris Samaras. Diffusion models as plug-and-play priors. In *Advances in Neural Information Processing Systems*, volume 35, pp. 14715–14728. Curran Associates, Inc., 2022. doi: 10.52202/068431-1070.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Tim Johnston, Iosif Lytras, Nikolaos Makras, and Sotirios Sabanis. The performance of the unadjusted langevin algorithm without smoothness assumptions. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=TTNeuyYdhg.

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 (CVPR)*, June 2019.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.

Housen Li, Johannes Schwab, Stephan Antholzer, and Markus Haltmeier. NETT: solving inverse problems with deep neural networks. *Inverse Problems*, 36(6):065005, 2020. doi: 10.1088/1361-6420/ab6d57.

Shudong Li, Xiao Jiang, Matthew Tivnan, Grace J. Gang, Yuan Shen, and J. Webster Stayman. Ct reconstruction using diffusion posterior sampling conditioned on a nonlinear measurement model. *Journal of Medical Imaging*, 11(4):043504, 2024. doi: 10.1117/1.JMI.11.4.043504.

Xiang Li, Soo Min Kwon, Shijun Liang, Ismail Alkhouri, Saiprasad Ravishankar, and Qing Qu. Dcdp: Decoupled data consistency via diffusion purification for inverse problems. In *2026 IEEE International Conference on Computational Photography (ICCP)*, pp. 1–15, 2026a. doi: 10.1109/ICCP69532.2026.11668875.

Xiang Li, Zebang Shen, Ya-Ping Hsieh, and Niao He. When scores learn geometry: Rate separations under the manifold hypothesis. In *International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=34V0IZytle.

Sebastian Lunz, Ozan Öktem, and Carola-Bibiane Schönlieb. Adversarial regularizers in inverse problems. In *Advances in Neural Information Processing Systems*, volume 31, pp. 8507–8516, 2018.

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.

Taylor R. Moen, Baiyu Chen, David R. Holmes III, Xinhui Duan, Zhicong Yu, Lifeng Yu, Shuai Leng, Joel G. Fletcher, and Cynthia H. McCollough. Low-dose ct image and projection dataset. *Medical Physics*, 48(2):902–911, 2021. doi: https://doi.org/10.1002/mp.14594.

Scott Pesme, Giacomo Meanti, Michael Arbel, and Julien Mairal. Map estimation with denoisers: Convergence rates and guarantees. *Advances in Neural Information Processing Systems*, 2025.

Marien Renaud, Jean Prost, Arthur Leclaire, and Nicolas Papadakis. Plug-and-play image restoration with stochastic deNOising regularization. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 42484–42520, 2024.

Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1530–1538. PMLR, 2015.

Herbert Robbins. An empirical Bayes approach to statistics. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1:157–163, 1956.

Herbert Robbins and David Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing Methods in Statistics*, pp. 233–257. Academic Press, 1971. doi: 10.1016/B978-0-12-604550-5.50015-8.

R. Tyrrell Rockafellar and Roger J.-B. Wets. *Variational Analysis*. Springer, Berlin, 1998. doi: 10.1007/978-3-642-02431-3.

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 *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *PMLR*, pp. 5546–5557, 2019.

Jiayang Shi, Daniël M. Pelt, and K. Joost Batenburg. DM4CT: Benchmarking diffusion models for computed tomography reconstruction. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=YE5scJekg5.

Rajesh Shrestha and Xiao Fu. Taming score-based denoisers in ADMM: A convergent plug-and-play framework. In *International Conference on Learning Representations*, 2026.

Zakhar Shumaylov, Jeremy Budd, Subhadip Mukherjee, and Carola-Bibiane Schönlieb. Weakly convex regularisers for inverse problems: Convergence of critical points and primal-dual optimisation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 45286–45314. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/shumaylov24a.html.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

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.

Jan Pawel Stanczuk, Georgios Batzolis, Teo Deveney, and Carola-Bibiane Schönlieb. Diffusion models encode the intrinsic dimension of data manifolds. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 46412–46440, 2024. URL https://proceedings.mlr.press/v235/stanczuk24a.html.

Tao Sun, Roberto Barrio, Lizhi Cheng, and Hao Jiang. Precompact convergence of the nonconvex primal–dual hybrid gradient algorithm. *Journal of Computational and Applied Mathematics*, 330: 15–27, 2018. doi: 10.1016/j.cam.2017.07.037.

Yodai Suzuki, Ryosuke Isono, and Shunsuke Ono. Convergent primal-dual plug-and-play image restoration: A general algorithm and applications. *IEEE Transactions on Computational Imaging*, 12:142–157, 2026. doi: 10.1109/TCI.2025.3644248.

W. van Aarle, W. J. Palenstijn, J. Cant, E. Janssens, F. Bleichrodt, A. Dabravolski, J. De Beenhouwer, K. J. Batenburg, and J. Sijbers. Fast and flexible x-ray tomography using the ASTRA toolbox. *Optics Express*, 24(22):25129–25147, 2016.

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 (GlobalSIP)*, pp. 945–948, 2013. doi: 10.1109/GlobalSIP.2013.6737048.

David Williams. *Probability with Martingales*. Cambridge Mathematical Textbooks. Cambridge University Press, 1991.

Bingliang Zhang, Wenda Chu, Julius Berner, Chenlin Meng, Anima Anandkumar, and Yang Song. Improving diffusion inverse problem solving with decoupled noise annealing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20895–20905, 2025.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

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*, pp. 1219–1229, 2023. doi: 10.1109/CVPRW59228.2023.00129.

A NOTATION

Table 3. Summary of notation conventions.

Symbol	Description
$\mathbb{R}^d$, $\mathbb{R}^m$	Euclidean spaces of dimension d and m.
I, I_d	Identity matrix. I_d is the $d \times d$ identity matrix.
$(\cdot)^\top$, $A^\top$	Transpose and adjoint for real linear operators.
$\langle u, v \rangle$	Euclidean inner product.
$\|u\|$	Euclidean vector norm.
$\|M\|$	Operator norm, i.e., spectral norm, of a matrix or linear operator M.
$\|z\|$	Absolute value or complex modulus
$uv^\top$	Outer product of the vectors u and v.
$h \circ A$	Function composition, $(h \circ A)(x) = h(Ax)$.
f^*	Fenchel conjugate of a function f.
$\partial f(x)$	Limiting subdifferential of f at x.
$\nabla f(x)$, $\nabla^2 f(x)$	Gradient and Hessian of f at x.
$\mathrm{prox}_{\alpha f}(u)$	Proximal operator of f with parameter α.
$\arg\min$, $\arg\max$	Set of minimizers or maximizers.
$\mathcal{N}(\mu, \Sigma)$	Gaussian distribution with mean μ and covariance Σ.
$\mathbb{E}[\cdot]$, $\mathbb{E}[\cdot \mid \cdot]$	Expectation and conditional expectation.
$D_{\mathrm{KL}}(P\|Q)$	Kullback–Leibler divergence from distribution P to distribution Q.
$*$	Convolution of densities or probability measures.
$\mathrm{supp}(p)$	Support of a density or probability measure.
$B(x, r)$	Euclidean ball of radius r centered at x.
σ_k, τ_k, ρ_k	Denoising level, primal stepsize, and injected-noise amplitude.
$O(a_k)$, $o(a_k)$	Big-O and little-o notation, bounded up to a constant multiple of a_k, and negligible relative to a_k, respectively.
$\liminf$, $\limsup$	Limit inferior and limit superior.
e^{-z}, $\exp(z)$	Componentwise exponential when z is vector-valued.

B PROOF OF PROPOSITION 1

Proof. Suppose that $x^k \to x^*$. Since $(w^k)_k$ is bounded, there exists a subsequence $(k_j)_j$ such that $w^{k_j} \to w^*$. The vanishing increments imply

$$x^{k_j+1} \to x^*, \qquad w^{k_j+1} \to w^*.$$

The dual update gives

$$z^{k_j+1} = Ax^{k_j} + \frac{w^{k_j} - w^{k_j+1}}{\gamma} \to Ax^*.$$

Define the proximal inputs

$$u_j := Ax^{k_j} + \frac{w^{k_j}}{\gamma} \to u^* := Ax^* + \frac{w^*}{\gamma},$$
$$v_j := x^{k_j} - \tau A^{\mathrm{T}} w^{k_j+1} \to v^* := x^* - \tau A^{\mathrm{T}} w^*.$$

By the minimizing property of the proximal updates, for every $z \in \mathrm{dom}\, h$ and $x \in \mathrm{dom}\, g$,

$$h(z^{k_j+1}) + \frac{\gamma}{2}\|z^{k_j+1} - u_j\|^2 \le h(z) + \frac{\gamma}{2}\|z - u_j\|^2,$$
$$g(x^{k_j+1}) + \frac{1}{2\tau}\|x^{k_j+1} - v_j\|^2 \le g(x) + \frac{1}{2\tau}\|x - v_j\|^2.$$

Taking lower limits and using lower semicontinuity yields

$$h(Ax^*) + \frac{\gamma}{2}\|Ax^* - u^*\|^2 \le h(z) + \frac{\gamma}{2}\|z - u^*\|^2,$$
$$g(x^*) + \frac{1}{2\tau}\|x^* - v^*\|^2 \le g(x) + \frac{1}{2\tau}\|x - v^*\|^2.$$

Properness ensures that the right-hand sides are finite for some z and x. Hence $h(Ax^*)$ and $g(x^*)$ are finite, and these inequalities show that

$$Ax^* \in \operatorname{prox}_{h/\gamma}(u^*), \qquad x^* \in \operatorname{prox}_{\tau g}(v^*).$$

The corresponding optimality conditions give

$$w^* \in \partial h(Ax^*), \qquad -A^{\mathrm{T}} w^* \in \partial g(x^*).$$

Consequently, by (13),

$$0 \in \partial g(x^*) + A^{\mathrm{T}} \partial h(Ax^*) \subseteq \partial(g + h \circ A)(x^*),$$

which proves the claim. □

C Proof of Proposition 2

Fix $v \in \mathbb{R}^d$ and let

$$U_{\sigma,v}(x) = g_0(x) + \frac{1}{2\sigma^2}\|x - v\|^2.$$

Since g_0 has an L_g-Lipschitz gradient, $U_{\sigma,v}$ is a_σ-strongly convex for

$$a_\sigma = \sigma^{-2} - L_g > 0.$$

Hence

$$x_\sigma(v) := \arg\min_x U_{\sigma,v}(x) = \operatorname{prox}_{\sigma^2 g_0}(v)$$

is uniquely defined and satisfies

$$x_\sigma(v) - v = -\sigma^2 \nabla g_0(x_\sigma(v)). \tag{18}$$

Let $X \sim p_0$, let $Z \sim \mathcal{N}(0, I_d)$ be independent, and condition on $X + \sigma Z = v$. By Bayes' rule, the conditional density is

$$\pi_{\sigma,v}(x) = \frac{1}{Z_{\sigma,v}} \exp\left(-g_0(x) - \frac{\|x - v\|^2}{2\sigma^2}\right) = \frac{1}{Z_{\sigma,v}} e^{-U_{\sigma,v}(x)}.$$

If $Y \sim \pi_{\sigma,v}$, Tweedie's identity gives

$$D_\sigma(v) = \mathbb{E}[Y]. \tag{19}$$

Strong convexity at $x_\sigma(v)$ gives

$$U_{\sigma,v}(x) \ge U_{\sigma,v}(x_\sigma(v)) + \frac{a_\sigma}{2}\|x - x_\sigma(v)\|^2.$$

Thus $\pi_{\sigma,v}$ is bounded by a Gaussian centered at $x_\sigma(v)$. In particular, the boundary terms in the calculations below vanish.

Since

$$\nabla \pi_{\sigma,v}(x) = -\nabla U_{\sigma,v}(x) \pi_{\sigma,v}(x),$$

integration by parts in each coordinate gives

$$\mathbb{E}[\nabla U_{\sigma,v}(Y)] = -\int_{\mathbb{R}^d} \nabla \pi_{\sigma,v}(x)\, dx = 0.$$

Similarly,

$$\begin{aligned}\mathbb{E}\left[\langle Y - x_\sigma(v), \nabla U_{\sigma,v}(Y)\rangle\right] &= -\sum_{i=1}^{d}\int_{\mathbb{R}^d}(x_i - x_{\sigma,i}(v))\partial_i \pi_{\sigma,v}(x)\,dx \\ &= \sum_{i=1}^{d}\int_{\mathbb{R}^d}\pi_{\sigma,v}(x)\,dx \\ &= d.\end{aligned}$$

On the other hand, strong monotonicity of $\nabla U_{\sigma,v}$ and $\nabla U_{\sigma,v}(x_\sigma(v)) = 0$ imply

$$a_\sigma\|x - x_\sigma(v)\|^2 \le \langle x - x_\sigma(v), \nabla U_{\sigma,v}(x)\rangle\,.$$

Taking expectations therefore yields

$$\mathbb{E}\|Y - x_\sigma(v)\|^2 \le \frac{d}{a_\sigma} = \frac{d\sigma^2}{1 - L_g\sigma^2}. \tag{20}$$

Expanding $\mathbb{E}[\nabla U_{\sigma,v}(Y)] = 0$ gives

$$\mathbb{E}[\nabla g_0(Y)] + \frac{\mathbb{E}[Y] - v}{\sigma^2} = 0.$$

Together with (19),

$$D_\sigma(v) - v = -\sigma^2\mathbb{E}[\nabla g_0(Y)].$$

Subtracting (18),

$$D_\sigma(v) - x_\sigma(v) = -\sigma^2\left(\mathbb{E}[\nabla g_0(Y)] - \nabla g_0(x_\sigma(v))\right).$$

Hence

$$\begin{aligned}\|D_\sigma(v) - x_\sigma(v)\| &\le L_g\sigma^2\mathbb{E}\|Y - x_\sigma(v)\| \\ &\le L_g\sigma^2\left(\mathbb{E}\|Y - x_\sigma(v)\|^2\right)^{1/2} \\ &\le \frac{L_g\sqrt{d}\,\sigma^3}{\sqrt{1 - L_g\sigma^2}}.\end{aligned}$$

Therefore, for every bounded $K \subset \mathbb{R}^d$,

$$\sup_{v\in K}\left\|D_\sigma(v) - \mathrm{prox}_{\sigma^2 g_0}(v)\right\| = O(\sigma^3).$$

For later use, the same argument gives a bound on the smoothed score. Since Tweedie's identity gives

$$\nabla g_\sigma(v) = \mathbb{E}[\nabla g_0(Y)],$$

we have

$$\begin{aligned}\|\nabla g_\sigma(v) - \nabla g_0(v)\| &\le L_g\mathbb{E}\|Y - v\| \\ &\le L_g\left(\mathbb{E}\|Y - x_\sigma(v)\|^2\right)^{1/2} + L_g\|x_\sigma(v) - v\|.\end{aligned}$$

The proximal optimality condition gives

$$\begin{aligned}\|x_\sigma(v) - v\| &= \sigma^2\|\nabla g_0(x_\sigma(v))\| \\ &\le \sigma^2\|\nabla g_0(v)\| + L_g\sigma^2\|x_\sigma(v) - v\|,\end{aligned}$$

and hence

$$\|x_\sigma(v) - v\| \le \frac{\sigma^2}{1 - L_g\sigma^2}\|\nabla g_0(v)\|.$$

Combining this with (20) gives

$$\|\nabla g_\sigma(v) - \nabla g_0(v)\| \le \frac{L_g\sqrt{d}\,\sigma}{\sqrt{1 - L_g\sigma^2}} + \frac{L_g\sigma^2}{1 - L_g\sigma^2}\|\nabla g_0(v)\|. \tag{21}$$

Thus, for every bounded $K \subset \mathbb{R}^d$,

$$\sup_{v\in K}\|\nabla g_\sigma(v) - \nabla g_0(v)\| = O(\sigma).$$

D PROOF OF THEOREM 1

The proof outline consists of first analyzing the z^{k+1} and w^{k+1} updates in Algorithm 1. We show that the residual $r_k = Ax^k - z^{k+1}$ vanishes when the primal increments tend to zero.

We then turn to the x^{k+1} update. Using the estimate from Proposition 2, we write this update as a perturbed descent step for the MAP objective. Since the MAP objective may be nonsmooth, we study the gradient of its Moreau envelope instead. The descent argument shows that this gradient converges to zero, and stationary points of the Moreau envelope coincide with stationary points of the MAP objective.

Throughout this appendix, set

$$a_k = Ax^k, \qquad r_k = a_k - z^{k+1}, \qquad \alpha = \gamma - \beta.$$

We begin with the z^{k+1} and w^{k+1} updates. The following lemma shows that boundedness of (Ax^k) is enough to bound both sequences.

Lemma D.1. *Suppose Assumption 2 holds, (Ax^k) is bounded, and $\gamma > 2\beta$. Then the z^{k+1}-update in Algorithm 1 is uniquely defined for every k, and the resulting sequences (w^k) and (z^{k+1}) satisfy*

$$\sup_k \|w^k\| < \infty, \qquad \sup_k \|z^{k+1}\| < \infty.$$

Proof. Set

$$f = h + \frac{\beta}{2}\|\cdot\|^2.$$

By Assumption 2, f is finite and convex. The measurement objective at iteration k can be written as

$$\begin{aligned} Q_k(z) &= h(z) + \langle w^k, a_k - z\rangle + \frac{\gamma}{2}\|a_k - z\|^2 \\ &= f(z) + \frac{\gamma-\beta}{2}\|z\|^2 - \langle \gamma a_k + w^k, z\rangle + \langle w^k, a_k\rangle + \frac{\gamma}{2}\|a_k\|^2. \end{aligned}$$

Since f has an affine lower bound and $\gamma > \beta$, Q_k is coercive and $(\gamma-\beta)$-strongly convex. Thus z^{k+1} is uniquely defined. Its optimality condition is

$$0 \in \partial h(z^{k+1}) + \gamma(z^{k+1} - a_k) - w^k.$$

Equivalently,

$$w^k + \gamma(a_k - z^{k+1}) \in \partial h(z^{k+1}).$$

The dual update

$$w^{k+1} = w^k + \gamma(a_k - z^{k+1})$$

therefore gives

$$w^{k+1} \in \partial h(z^{k+1}), \qquad z^{k+1} = a_k + \frac{w^k - w^{k+1}}{\gamma}. \tag{22}$$

Because f is finite and convex on $\mathbb{R}^m$, its subdifferential is nonempty and locally bounded. Hence, since

$$h(z) = f(z) - \frac{\beta}{2}\|z\|^2, \qquad \partial h(z) = \partial f(z) - \beta z,$$

the same properties hold for ∂h. See, e.g., (Rockafellar & Wets, 1998).

Let

$$R_a := \sup_k \|a_k\| < \infty$$

and choose

$$M_0 := \max\left\{1, \sup_{\substack{\|a\|\le R_a \\ s\in\partial h(a)}} \|s\|\right\} < \infty.$$

For each k, choose $s_k \in \partial h(a_k)$ with $\|s_k\| \le M_0$. Since

$$w^{k+1} + \beta z^{k+1} \in \partial f(z^{k+1}), \qquad s_k + \beta a_k \in \partial f(a_k),$$

monotonicity of ∂f gives

$$\begin{aligned} 0 &\le \left\langle w^{k+1} + \beta z^{k+1} - s_k - \beta a_k, z^{k+1} - a_k \right\rangle \\ &= -\langle w^{k+1} - s_k, r_k \rangle + \beta \|r_k\|^2. \end{aligned}$$

Hence

$$\langle w^{k+1}, r_k \rangle \le M_0 \|r_k\| + \beta \|r_k\|^2. \tag{23}$$

The dual update also gives

$$w^k = w^{k+1} - \gamma r_k.$$

Combining this identity with (23),

$$\begin{aligned} \|w^{k+1}\|^2 - \|w^k\|^2 &= 2\gamma \langle w^{k+1}, r_k \rangle - \gamma^2 \|r_k\|^2 \\ &\le 2\gamma M_0 \|r_k\| - \gamma(\gamma - 2\beta)\|r_k\|^2. \end{aligned} \tag{24}$$

Set

$$R := \frac{2M_0}{\gamma - 2\beta}.$$

If $\|r_k\| \ge R$, then (24) gives

$$\|w^{k+1}\| \le \|w^k\|.$$

If $\|r_k\| < R$, then

$$\|z^{k+1}\| = \|a_k - r_k\| \le R_a + R.$$

Local boundedness of ∂h, together with (22), therefore gives

$$\|w^{k+1}\| \le M_1, \qquad M_1 := \sup_{\substack{\|z\| \le R_a + R \\ s \in \partial h(z)}} \|s\| < \infty.$$

Thus every dual update either decreases the dual norm or places it below M_1. It follows by induction that

$$\|w^k\| \le W := \max\{\|w^0\|, M_1\} \qquad \text{for all } k.$$

Finally, (22) gives

$$\|z^{k+1}\| \le R_a + \frac{\|w^k\| + \|w^{k+1}\|}{\gamma} \le R_a + \frac{2W}{\gamma}.$$

□

We next consider how closely the z^{k+1}-update tracks Ax^k. Define

$$\ell_\gamma(a, w) = \min_z \left\{ h(z) + \langle w, a - z \rangle + \frac{\gamma}{2} \|a - z\|^2 \right\},$$

and

$$\Psi(a, w) = h(a) - \ell_\gamma(a, w), \qquad \Psi_k = \Psi(a_k, w^k).$$

Thus Ψ_k is the decrease in the z-subproblem obtained by moving from $z = a_k$ to its minimizer z^{k+1}. The next lemma relates this decrease to the residual $r_k = Ax^k - z^{k+1}$ and shows that, when the primal iterates move asymptotically little, both r_k and the increments $w^{k+1} - w^k$ vanish.

Lemma D.2. *Under the assumptions of Lemma D.1, ℓ_γ has a globally Lipschitz gradient. There is a constant $C > 0$ such that*

$$\frac{\alpha}{2} \|r_k\|^2 \le \Psi_k \le C \|r_k\|, \tag{25}$$

$$\Psi_{k+1} \le \Psi_k - \chi \|r_k\|^2 + C \|a_{k+1} - a_k\|, \qquad \chi = \frac{\gamma(\gamma - 2\beta)}{2(\gamma - \beta)} > 0. \tag{26}$$

If $a_{k+1} - a_k \to 0$, then

$$\Psi_k \to 0, \qquad r_k \to 0, \qquad w^{k+1} - w^k \to 0. \tag{27}$$

If h has a locally Lipschitz gradient, then also

$$\Psi_k \le C \|r_k\|^2. \tag{28}$$

Proof. Let $z(a, w)$ denote the unique minimizer in the definition of ℓ_γ. We first investigate its dependence on (a, w). Let

$$z = z(a, w), \qquad z' = z(a', w').$$

Their optimality conditions are

$$w + \gamma(a - z) \in \partial h(z), \qquad w' + \gamma(a' - z') \in \partial h(z').$$

Since $h + \frac{\beta}{2}\|\cdot\|^2$ is convex, monotonicity of its subdifferential gives

$$\alpha\|z' - z\|^2 \le \langle \gamma(a' - a) + (w' - w), z' - z\rangle,$$

where $\alpha = \gamma - \beta$. Hence

$$\|z' - z\| \le \frac{\gamma\|a' - a\| + \|w' - w\|}{\alpha}. \tag{29}$$

Since the minimizer is unique, Danskin's theorem applied to

$$\ell_\gamma(a, w) = \min_z \left\{ h(z) + \langle w, a - z\rangle + \frac{\gamma}{2}\|a - z\|^2 \right\}$$

gives

$$\nabla_a \ell_\gamma(a, w) = w + \gamma(a - z(a, w)), \qquad \nabla_w \ell_\gamma(a, w) = a - z(a, w).$$

Together with (29), these formulas show that $\nabla\ell_\gamma$ is globally Lipschitz. In particular, $\nabla_w \ell_\gamma(a, \cdot)$ is $1/\alpha$-Lipschitz.

We now turn to the gap. At iteration k, let

$$Q_k(z) = h(z) + \langle w^k, a_k - z\rangle + \frac{\gamma}{2}\|a_k - z\|^2.$$

Since $Q_k(a_k) = h(a_k)$ and $Q_k(z^{k+1}) = \ell_\gamma(a_k, w^k)$,

$$\Psi_k = Q_k(a_k) - Q_k(z^{k+1}).$$

The function Q_k is α-strongly convex and z^{k+1} is its minimizer. Therefore

$$\Psi_k \ge \frac{\alpha}{2}\|a_k - z^{k+1}\|^2 = \frac{\alpha}{2}\|r_k\|^2.$$

For the upper bound, choose $s_k \in \partial h(a_k)$. Lemma D.1 and boundedness of (a_k) give

$$\sup_k \|s_k\| < \infty, \qquad \sup_k \|w^k\| < \infty.$$

Weak convexity at a_k gives

$$h(z^{k+1}) \ge h(a_k) - \langle s_k, r_k\rangle - \frac{\beta}{2}\|r_k\|^2.$$

Using

$$\Psi_k = h(a_k) - h(z^{k+1}) - \langle w^k, r_k\rangle - \frac{\gamma}{2}\|r_k\|^2,$$

we obtain

$$\begin{aligned}\Psi_k &\le \langle s_k - w^k, r_k\rangle - \frac{\alpha}{2}\|r_k\|^2 \\ &\le C\|r_k\|.\end{aligned}$$

This proves (25).

We will also use the identity

$$B_k := \Psi_k - \frac{\alpha}{2}\|r_k\|^2 = h(a_k) - h(z^{k+1}) - \langle w^{k+1}, r_k\rangle + \frac{\beta}{2}\|r_k\|^2 \ge 0. \tag{30}$$

Indeed, the dual update gives $w^k = w^{k+1} - \gamma r_k$, and substitution in the expression for Ψ_k gives the equality.

Next, since

$$\nabla_w \ell_\gamma(a_k, w^k) = r_k, \qquad w^{k+1} - w^k = \gamma r_k,$$

the lower Taylor bound for the $1/\alpha$-smooth function $\ell_\gamma(a_k, \cdot)$ gives

$$\begin{aligned}\ell_\gamma(a_k, w^{k+1}) &\geq \ell_\gamma(a_k, w^k) + \gamma\|r_k\|^2 - \frac{\gamma^2}{2\alpha}\|r_k\|^2 \\ &= \ell_\gamma(a_k, w^k) + \chi\|r_k\|^2.\end{aligned}$$

Thus

$$\Psi(a_k, w^{k+1}) \leq \Psi_k - \chi\|r_k\|^2. \tag{31}$$

Since $h + \frac{\beta}{2}\|\cdot\|^2$ is finite and convex, h is locally Lipschitz. The function ℓ_γ is continuously differentiable. Lemma D.1 places (a_k, w^k, w^{k+1}) in a bounded set, so $\Psi = h - \ell_\gamma$ is Lipschitz on the corresponding bounded set. Hence

$$\Psi_{k+1} \leq \Psi(a_k, w^{k+1}) + C\|a_{k+1} - a_k\|.$$

Combining this with (31) gives (26).

Suppose now that $a_{k+1} - a_k \to 0$. From (25) and (26),

$$\Psi_{k+1} \leq \Psi_k - c\Psi_k^2 + \epsilon_k, \qquad c > 0, \qquad \epsilon_k \to 0.$$

Fix $\delta > 0$. For all sufficiently large k,

$$\epsilon_k \leq \min\left\{\frac{c\delta^2}{2}, \delta\right\}.$$

Whenever $\Psi_k \geq \delta$, the sequence decreases by at least $c\delta^2/2$, so it eventually enters $[0, \delta)$. Once there, the recursion keeps it below 2δ. Since δ is arbitrary, $\Psi_k \to 0$. The lower bound in (25) then gives $r_k \to 0$, and

$$w^{k+1} - w^k = \gamma r_k \to 0.$$

Finally, suppose that h has a locally Lipschitz gradient. Lemma D.1 places (a_k) and (z^{k+1}) in a closed ball. Let L_h be a Lipschitz constant for ∇h on this ball. Since

$$w^{k+1} = \nabla h(z^{k+1}),$$

the expression for Ψ_k can be written as

$$\Psi_k = h(a_k) - h(z^{k+1}) - \langle \nabla h(z^{k+1}), r_k\rangle + \frac{\gamma}{2}\|r_k\|^2.$$

The descent lemma gives

$$h(a_k) - h(z^{k+1}) - \langle \nabla h(z^{k+1}), r_k\rangle \leq \frac{L_h}{2}\|r_k\|^2.$$

Therefore

$$\Psi_k \leq \frac{L_h + \gamma}{2}\|r_k\|^2,$$

which proves (28). □

Moreau envelope of the MAP objective. Set

$$F_\lambda = h \circ A + \lambda g_0, \qquad \kappa = \beta\|A\|^2 + \lambda L_g.$$

The MAP objective F_λ may be nonsmooth, while its Moreau envelope is smooth under weak convexity. We therefore use the Moreau envelope as a smooth stationarity measure for F_λ. Its stationary points coincide with those of the original objective.

Since h is β-weakly convex and g_0 is L_g-smooth, F_λ is κ-weakly convex. Indeed,

$$\begin{aligned}F_\lambda(x) + \frac{\kappa}{2}\|x\|^2 = &\left(h(Ax) + \frac{\beta}{2}\|Ax\|^2\right) + \lambda\left(g_0(x) + \frac{L_g}{2}\|x\|^2\right) \\ &+ \frac{\beta}{2}\left(\|A\|^2\|x\|^2 - \|Ax\|^2\right),\end{aligned}$$

and each term on the right is convex.

Fix $\mu > 0$ with $\mu\kappa < 1$ and define

$$E_\mu(x) = \min_y \left\{ F_\lambda(y) + \frac{\|y - x\|^2}{2\mu} \right\}, \qquad y_k = \mathrm{prox}_{\mu F_\lambda}(x^k), \qquad m_k = \nabla E_\mu(x^k). \tag{32}$$

Standard properties of the Moreau envelope of a weakly convex function (Davis & Drusvyatskiy, 2019, Section 2.2) give

$$m_k = \frac{x^k - y_k}{\mu},$$

and ∇E_μ is Lipschitz. Moreover,

$$\nabla E_\mu(x) = 0 \quad \Longleftrightarrow \quad 0 \in \partial F_\lambda(x).$$

The next lemma relates the direction appearing in the x^{k+1}-update to the Moreau-envelope gradient. It shows that their alignment is positive up to the z^{k+1}-subproblem gap Ψ_k and the denoising error.

Lemma D.3. *Suppose the assumptions of Lemma D.2 hold and g_0 is globally L_g-smooth. Let $G_k = A^\top w^{k+1} + \lambda \nabla g_0(x^k)$ and $d_k = G_k + e_k$. For every sequence (e_k), there are constants $c_\mu, C > 0$, such that*

$$\langle d_k, m_k \rangle \geq c_\mu \|m_k\|^2 - C\Psi_k - C\|e_k\|^2. \tag{33}$$

Proof. The expansion of B_k in Lemma D.2 gives

$$B_k = h(a_k) - h(z^{k+1}) - \langle w^{k+1}, r_k \rangle + \frac{\beta}{2}\|r_k\|^2.$$

For $y \in \mathbb{R}^d$, write $t = y - x^k$. Then $Ay - z^{k+1} = At + r_k$. Weak convexity at z^{k+1}, using $w^{k+1} \in \partial h(z^{k+1})$, gives

$$\begin{aligned} h(Ay) &\geq h(z^{k+1}) + \langle w^{k+1}, At + r_k \rangle - \frac{\beta}{2}\|At + r_k\|^2 \\ &= h(a_k) + \langle A^\top w^{k+1}, t \rangle - \frac{\beta}{2}\|At\|^2 - \beta\langle r_k, At \rangle - B_k. \end{aligned}$$

The lower quadratic bound for g_0 is

$$\lambda g_0(y) \geq \lambda g_0(x^k) + \lambda \langle \nabla g_0(x^k), t \rangle - \frac{\lambda L_g}{2}\|t\|^2.$$

Adding the two inequalities and using $\|At\| \leq \|A\|\|t\|$ yields

$$\begin{aligned} F_\lambda(y) \geq F_\lambda(x^k) + \langle G_k, y - x^k \rangle - \frac{\kappa}{2}\|y - x^k\|^2 \\ - \beta\langle r_k, A(y - x^k) \rangle - B_k. \end{aligned} \tag{34}$$

Substitute $y = y_k = x^k - \mu m_k$. Proximal minimality gives

$$F_\lambda(y_k) + \frac{\mu}{2}\|m_k\|^2 \leq F_\lambda(x^k),$$

whereas (34) gives

$$F_\lambda(y_k) \geq F_\lambda(x^k) - \mu\langle G_k, m_k \rangle - \frac{\kappa\mu^2}{2}\|m_k\|^2 + \beta\mu\langle r_k, Am_k \rangle - B_k.$$

Combining these bounds and dividing by μ yields

$$\langle G_k, m_k \rangle \geq \frac{1 - \mu\kappa}{2}\|m_k\|^2 + \beta\langle r_k, Am_k \rangle - \frac{B_k}{\mu}.$$

Set $\theta = 1 - \mu\kappa > 0$. Young's inequality gives

$$\begin{aligned} \beta|\langle r_k, Am_k \rangle| &\leq \frac{\theta}{8}\|m_k\|^2 + \frac{2\beta^2\|A\|^2}{\theta}\|r_k\|^2, \\ |\langle e_k, m_k \rangle| &\leq \frac{\theta}{8}\|m_k\|^2 + \frac{2}{\theta}\|e_k\|^2. \end{aligned}$$

Using $B_k \le \Psi_k$ and $\|r_k\|^2 \le 2\Psi_k/\alpha$, we obtain

$$\langle d_k, m_k\rangle \ge \frac{\theta}{4}\|m_k\|^2 - \left(\frac{1}{\mu} + \frac{4\beta^2\|A\|^2}{\alpha\theta}\right)\Psi_k - \frac{2}{\theta}\|e_k\|^2.$$

Thus (33) holds with

$$c_\mu = \frac{\theta}{4}, \qquad C = \frac{1}{\mu} + \frac{4\beta^2\|A\|^2}{\alpha\theta} + \frac{2}{\theta}.$$

□

We next isolate a sequence argument for converting weighted square-summability of the Moreau-envelope gradient into pointwise convergence. The argument considers an interval over which $\|\nabla E_\mu(x^k)\|$ decreases from a fixed level ϵ to $\epsilon/2$. Weighted square-summability forces the accumulated stepsize over such an interval to vanish. Since the update directions are bounded and the perturbation series converges, the corresponding change in x^k also vanishes. Lipschitz continuity of ∇E_μ then rules out a fixed decrease of its norm over such intervals. This finite-interval argument follows (Bertsekas & Tsitsiklis, 2000).

Lemma D.4. *Let E_μ have a Lipschitz gradient. Suppose $x^{k+1} = x^k - \tau_k d_k + \nu_k$, where (d_k) is bounded, $\tau_k > 0$, $\sum_k \tau_k = \infty$, and $\sum_k \nu_k$ converges. If*

$$\sum_k \tau_k \|\nabla E_\mu(x^k)\|^2 < \infty,$$

then $\nabla E_\mu(x^k) \to 0$.

Proof. Write $m_k = \nabla E_\mu(x^k)$. The summability assumption first gives

$$\liminf_k \|m_k\| = 0.$$

Indeed, if $\|m_k\| \ge \epsilon > 0$ for every $k \ge K$, then

$$\sum_{k\ge K} \tau_k\|m_k\|^2 \ge \epsilon^2 \sum_{k\ge K}\tau_k = \infty,$$

a contradiction.

It remains to rule out excursions of $\|m_k\|$ away from zero. Suppose, toward a contradiction, that for some $\epsilon > 0$ there are arbitrarily large indices k_j with

$$\|m_{k_j}\| \ge \epsilon.$$

Let $\ell_j > k_j$ be the first subsequent index with $\|m_{\ell_j}\| \le \epsilon/2$. Such an index exists by the limit inferior just established. For $k_j \le i < \ell_j$, $\|m_i\| > \epsilon/2$, and therefore

$$\sum_{i=k_j}^{\ell_j-1}\tau_i \le \frac{4}{\epsilon^2}\sum_{i=k_j}^{\ell_j-1}\tau_i\|m_i\|^2 \le \frac{4}{\epsilon^2}\sum_{i=k_j}^{\infty}\tau_i\|m_i\|^2 \longrightarrow 0.$$

With $M = \sup_i \|d_i\|$, summing the recursion gives

$$\|x^{\ell_j} - x^{k_j}\| \le M\sum_{i=k_j}^{\ell_j-1}\tau_i + \left\|\sum_{i=k_j}^{\ell_j-1}\nu_i\right\| \longrightarrow 0.$$

The second term tends to zero by the Cauchy property of $\sum_i \nu_i$. Lipschitz continuity of ∇E_μ now gives

$$\frac{\epsilon}{2} \le \|m_{k_j}\| - \|m_{\ell_j}\| \le \|m_{k_j} - m_{\ell_j}\| \le L_\mu\|x^{k_j} - x^{\ell_j}\| \longrightarrow 0,$$

a contradiction. Hence $m_k \to 0$. □

We now combine these estimates along the iterates of Algorithm 1. The annealing schedule controls the error in the x^{k+1}-update, while Lemmas D.1 and D.2 control the z^{k+1} and w^{k+1} updates and their residual. Lemma D.3 then gives descent of the Moreau envelope up to summable errors. This yields weighted square-summability of its gradient, and Lemma D.4 turns this into $\nabla E_\mu(x^k) \to 0$.

Proof of Theorem 1. Since $\sum_k \sigma_k^3 < \infty$, we have $\sigma_k \to 0$ and $\tau_k = \sigma_k^2/\lambda \to 0$. Moreover,

$$\sum_k \tau_k = \frac{1}{\lambda}\sum_k \sigma_k^2 = \infty, \qquad \sum_k \tau_k^{3/2} = \lambda^{-3/2}\sum_k \sigma_k^3 < \infty.$$

For all sufficiently large k, $\sigma_k \le 1$, so

$$\sum_k \tau_k^2 = \lambda^{-2}\sum_k \sigma_k^4 < \infty.$$

The sequences (τ_k) and $(\sqrt{\tau_k})$ are positive, deterministic, and eventually nonincreasing. Set $\overline{\tau} = \sup_k \tau_k < \infty$.

Set $\nu_k = \rho_k n^k$ and let

$$\mathcal{F}_k = \sigma(x^0, w^0, n^0, \dots, n^{k-1})$$

denote the information available before n^k is sampled. The variables x^k, z^{k+1}, and w^{k+1} are $\mathcal{F}_k$-measurable, and n^k is independent of $\mathcal{F}_k$. The noise summability assumption implies, almost surely,

$$\sum_k \|\nu_k\|^2 < \infty, \qquad \sum_k \nu_k \text{ converges.} \tag{35}$$

Indeed, since the summands are nonnegative,

$$\mathbb{E}\left[\sum_k \|\nu_k\|^2\right] = \sum_k \mathbb{E}\|\nu_k\|^2 = d\sum_k \rho_k^2 < \infty.$$

Hence $\sum_k \|\nu_k\|^2 < \infty$ almost surely. For the second assertion, let

$$S_N = \sum_{k=0}^{N-1} \nu_k.$$

Since the ν_k are independent and centered, each coordinate of (S_N) is a martingale. Moreover,

$$\mathbb{E}\|S_N\|^2 = \sum_{k=0}^{N-1} \mathbb{E}\|\nu_k\|^2 = d\sum_{k=0}^{N-1} \rho_k^2 \le d\sum_k \rho_k^2 < \infty.$$

Thus (S_N) is L^2-bounded and converges almost surely, coordinatewise, by the L^2 martingale convergence theorem (Williams, 1991).

By Assumption 1 and Lemma D.1, the sequences (x^k), (w^k), and (z^{k+1}) are bounded almost surely. Equation (35) also gives $\sup_k \|\nu_k\| < \infty$. Hence the denoiser inputs

$$u^{k+1} = x^k - \tau_k A^\top w^{k+1} + \nu_k$$

satisfy

$$\sup_k \|u^{k+1}\| \le \sup_k \|x^k\| + \overline{\tau}\|A\| \sup_k \|w^{k+1}\| + \sup_k \|\nu_k\| < \infty.$$

Write $g_\sigma = -\log p_\sigma$. The estimate (21) in Appendix C gives, for every bounded set K,

$$\sup_{v \in K} \|\nabla g_\sigma(v) - \nabla g_0(v)\| \le C_K \sigma \quad \text{for sufficiently small } \sigma. \tag{36}$$

Along a bounded sample path, choose an integer-radius ball containing (x^k) and (u^{k+1}) and apply (36) on that ball.

Define

$$\begin{aligned} G_k &= A^\top w^{k+1} + \lambda \nabla g_0(x^k), \\ e_k &= \lambda\big(\nabla g_{\sigma_k}(u^{k+1}) - \nabla g_0(x^k)\big), \\ d_k &= G_k + e_k, \qquad x^{k+1} = x^k - \tau_k d_k + \nu_k. \end{aligned} \tag{37}$$

The last identity follows by expanding the Tweedie update:

$$\begin{aligned} x^{k+1} &= u^{k+1} - \sigma_k^2 \nabla g_{\sigma_k}(u^{k+1}) \\ &= x^k - \tau_k A^\top w^{k+1} + \nu_k - \lambda \tau_k \nabla g_{\sigma_k}(u^{k+1}) \\ &= x^k - \tau_k (G_k + e_k) + \nu_k. \end{aligned}$$

For all sufficiently large k, (36) and smoothness of g_0 give

$$\begin{aligned} \|e_k\| &\le \lambda \|\nabla g_{\sigma_k}(u^{k+1}) - \nabla g_0(u^{k+1})\| + \lambda \|\nabla g_0(u^{k+1}) - \nabla g_0(x^k)\| \\ &\le \lambda C_K \sigma_k + \lambda L_g \big(\tau_k \|A^\top w^{k+1}\| + \|\nu_k\|\big) \\ &\le C(\sigma_k + \tau_k + \|\nu_k\|). \end{aligned}$$

Moreover,

$$\|G_k\| \le \|A\| \sup_i \|w^{i+1}\| + \lambda \left(\|\nabla g_0(0)\| + L_g \sup_i \|x^i\| \right).$$

Since $\nu_k \to 0$, these estimates show that (G_k), (e_k), and (d_k) are bounded. Write $D_* = \sup_k \|d_k\| < \infty$. Using $\sigma_k^2 = \lambda \tau_k$ gives

$$\tau_k \|e_k\|^2 \le C\big(\lambda \tau_k^2 + \tau_k^3 + \tau_k \|\nu_k\|^2\big)$$

eventually, while the recursion gives

$$\|x^{k+1} - x^k\|^2 \le 2D_*^2 \tau_k^2 + 2\|\nu_k\|^2.$$

The mixed series are bounded by

$$\sum_k \tau_k^3 \le \overline{\tau} \sum_k \tau_k^2 < \infty, \qquad \sum_k \tau_k \|\nu_k\|^2 \le \overline{\tau} \sum_k \|\nu_k\|^2 < \infty.$$

After including the finite initial terms, we obtain

$$\begin{aligned} \sum_k \tau_k \|e_k\|^2 &\le C \sum_k \big(\lambda \tau_k^2 + \tau_k^3 + \tau_k \|\nu_k\|^2\big) < \infty, \\ \sum_k \|x^{k+1} - x^k\|^2 &\le 2D_*^2 \sum_k \tau_k^2 + 2 \sum_k \|\nu_k\|^2 < \infty. \end{aligned} \tag{38}$$

In particular, $\|a_{k+1} - a_k\| \le \|A\| \|x^{k+1} - x^k\| \to 0$. Lemma D.2 therefore gives $\Psi_k \to 0$, $r_k \to 0$, and $w^{k+1} - w^k \to 0$.

We next establish weighted summability of the dual gap. Let (b_k) be positive and deterministic, eventually nonincreasing, with $b_k \to 0$ and $\sum_k b_k \tau_k < \infty$. Set

$$L_k = \ell_\gamma(a_k, w^k), \qquad q_k = A^\top \nabla_a \ell_\gamma(a_k, w^{k+1}).$$

Continuity of ℓ_γ and its gradient on the bounded states implies $\sup_k |L_k| < \infty$ and $Q_* := \sup_k \|q_k\| < \infty$.

Let L_a be a global Lipschitz constant for $\nabla_a \ell_\gamma(\cdot, w)$, uniform in w. The bounds in Lemma D.2 permit $L_a = \gamma(1 + \gamma/\alpha)$. The lower Taylor bound gives

$$\begin{aligned} \ell_\gamma(a_{k+1}, w^{k+1}) \ge \ell_\gamma(a_k, w^{k+1}) \\ + \langle \nabla_a \ell_\gamma(a_k, w^{k+1}), A(x^{k+1} - x^k) \rangle - \frac{L_a \|A\|^2}{2} \|x^{k+1} - x^k\|^2. \end{aligned}$$

Combining this with (31) yields

$$L_{k+1} \ge L_k + \chi \|r_k\|^2 + \langle q_k, x^{k+1} - x^k \rangle - C\|x^{k+1} - x^k\|^2.$$

Substitute (37), rearrange, and multiply by b_k:

$$\begin{aligned} \chi b_k \|r_k\|^2 \le b_k (L_{k+1} - L_k) + b_k \tau_k \langle q_k, d_k \rangle - b_k \langle q_k, \nu_k \rangle \\ + C b_k \|x^{k+1} - x^k\|^2. \end{aligned} \tag{39}$$

To sum (39), we need convergence of the noise series $\sum_k b_k\langle q_k, \nu_k\rangle$. Since q_k is $\mathcal{F}_k$-measurable and n^k is independent of $\mathcal{F}_k$, the summands are conditionally Gaussian with mean zero. Their partial sums form a locally square-integrable martingale. Boundedness of $(b_k q_k)$ gives

$$\begin{aligned}\sum_k \mathbb{E}\big[b_k^2\langle q_k, \nu_k\rangle^2 \,\big|\, \mathcal{F}_k\big] &= \sum_k b_k^2\rho_k^2\|q_k\|^2 \\ &\le \left(\sup_k \|b_k q_k\|^2\right)\sum_k \rho_k^2 < \infty \quad \text{almost surely.}\end{aligned}$$

The conditional variances are summable, so the martingale convergence theorem gives

$$\sum_k b_k\langle q_k, \nu_k\rangle \quad \text{converges with probability one.} \tag{40}$$

The remaining series satisfy

$$\sum_k b_k\tau_k|\langle q_k, d_k\rangle| \le Q_* D_* \sum_k b_k\tau_k < \infty$$

and

$$\sum_k b_k\|x^{k+1} - x^k\|^2 \le \left(\sup_k b_k\right)\sum_k \|x^{k+1} - x^k\|^2 < \infty.$$

Choose K so that $(b_k)_{k\ge K}$ is nonincreasing. Summation by parts gives

$$\sum_{k=K}^{N-1} b_k(L_{k+1} - L_k) = b_{N-1}L_N - b_K L_K + \sum_{k=K+1}^{N-1}(b_{k-1} - b_k)L_k.$$

The first term tends to zero, and

$$\sum_{k=K+1}^{\infty}|(b_{k-1} - b_k)L_k| \le \sup_k |L_k| \sum_{k=K+1}^{\infty}(b_{k-1} - b_k) = b_K \sup_k |L_k| < \infty.$$

Thus the summed right-hand side of (39) converges. Its left-hand side has nonnegative terms, so

$$\sum_k b_k\|r_k\|^2 < \infty \quad \text{almost surely.} \tag{41}$$

Take $b_k = \sqrt{\tau_k}$. Then $\sum_k b_k\tau_k = \sum_k \tau_k^{3/2} < \infty$. Using (25), Cauchy-Schwarz, and (41), we obtain

$$\begin{aligned}\sum_k \tau_k\Psi_k &\le C\sum_k \tau_k\|r_k\| \\ &= C\sum_k \tau_k^{3/4}\big(\tau_k^{1/4}\|r_k\|\big) \\ &\le C\left(\sum_k \tau_k^{3/2}\right)^{1/2}\left(\sum_k \sqrt{\tau_k}\|r_k\|^2\right)^{1/2} < \infty.\end{aligned} \tag{42}$$

When h has a locally Lipschitz gradient, the choice $b_k = \tau_k$ also satisfies $\sum_k b_k\tau_k = \sum_k \tau_k^2 < \infty$. Equations (28) and (41) then give

$$\sum_k \tau_k\Psi_k \le C\sum_k \tau_k\|r_k\|^2 < \infty.$$

Fix a deterministic $\mu > 0$ with $\mu\kappa < 1$ and set

$$m_k = \nabla E_\mu(x^k).$$

Since (x^k) is bounded and ∇E_μ is Lipschitz,

$$\|m_k\| \le \|\nabla E_\mu(0)\| + L_\mu \sup_i \|x^i\| < \infty.$$

The descent lemma applied to (37) gives

$$\begin{aligned} E_\mu(x^{k+1}) &\le E_\mu(x^k) - \tau_k\langle m_k, d_k\rangle + \langle m_k, \nu_k\rangle + \frac{L_\mu}{2}\| - \tau_k d_k + \nu_k\|^2 \\ &\le E_\mu(x^k) - \tau_k\langle m_k, d_k\rangle + \langle m_k, \nu_k\rangle + L_\mu D_*^2\tau_k^2 + L_\mu\|\nu_k\|^2. \end{aligned}$$

Using Lemma D.3 yields

$$\begin{aligned} E_\mu(x^{k+1}) \le E_\mu(x^k) - c_\mu\tau_k\|m_k\|^2 + C\tau_k\Psi_k + \langle m_k, \nu_k\rangle \\ + C\big(\tau_k\|e_k\|^2 + \tau_k^2 + \|\nu_k\|^2\big). \end{aligned} \tag{43}$$

Set

$$\varepsilon_k = C\big(\tau_k\Psi_k + \tau_k\|e_k\|^2 + \tau_k^2 + \|\nu_k\|^2\big).$$

Equations (35), (38), and (42) show that $\sum_k \varepsilon_k < \infty$. Since $m_k = \nabla E_\mu(x^k)$ is $\mathcal{F}_k$-measurable and (m_k) is bounded almost surely, the argument for (40) applies with m_k in place of $b_k q_k$. Hence

$$\sum_k \langle m_k, \nu_k\rangle \quad \text{converges with probability one.}$$

Define

$$V_k = E_\mu(x^k) - \sum_{i=0}^{k-1}\langle m_i, \nu_i\rangle + \sum_{i=k}^{\infty}\varepsilon_i. \tag{44}$$

Then

$$V_{k+1} - V_k = E_\mu(x^{k+1}) - E_\mu(x^k) - \langle m_k, \nu_k\rangle - \varepsilon_k,$$

so (43) gives

$$V_{k+1} \le V_k - c_\mu\tau_k\|m_k\|^2. \tag{45}$$

Let $K_x = \overline{\{x^k : k \ge 0\}}$ and $M_k = \sum_{i=0}^{k-1}\langle m_i, \nu_i\rangle$. The set K_x is compact, E_μ is continuous, and (M_k) converges. Since $\varepsilon_i \ge 0$,

$$V_k \ge \min_{x\in K_x} E_\mu(x) - \sup_j |M_j| > -\infty.$$

Thus (V_k) is decreasing and bounded below. Summing (45) gives

$$c_\mu \sum_{k=0}^{N-1}\tau_k\|m_k\|^2 \le V_0 - V_N \le V_0 - \inf_j V_j < \infty.$$

Consequently,

$$V_k \text{ converges}, \qquad \sum_k \tau_k\|m_k\|^2 < \infty. \tag{46}$$

Equation (44) also gives

$$E_\mu(x^k) = V_k + M_k - \sum_{i=k}^{\infty}\varepsilon_i,$$

so $E_\mu(x^k)$ converges.

Lemma D.4, applied pathwise to (37), now gives $m_k \to 0$. Every cluster point $x^\star$ therefore satisfies $\nabla E_\mu(x^\star) = 0$, and hence $0 \in \partial F_\lambda(x^\star)$.

To identify this condition with the stationary set in the theorem, recall that $f = h + \frac{\beta}{2}\|\cdot\|^2$ is finite and convex, and therefore continuous everywhere. The convex linear-composition chain rule consequently gives

$$\partial(f \circ A)(x) = A^\top \partial f(Ax).$$

Applying the smooth sum rule to $h \circ A = f \circ A - \frac{\beta}{2}\|A\cdot\|^2$ yields

$$\begin{aligned} \partial(h \circ A)(x) &= A^\top \partial f(Ax) - \beta A^\top Ax \\ &= A^\top\big(\partial f(Ax) - \beta Ax\big) = A^\top \partial h(Ax). \end{aligned} \tag{47}$$

Adding the smooth prior gives

$$\partial F_\lambda(x) = A^\top \partial h(Ax) + \lambda \nabla g_0(x).$$

Thus every cluster point belongs to

$$\mathcal{C} = \{x : 0 \in A^\top \partial h(Ax) + \lambda \nabla g_0(x)\},$$

which is the first-order condition in Proposition 1 for the weighted prior λg_0. If a subsequence satisfied $\operatorname{dist}(x^{k_j}, \mathcal{C}) \geq \epsilon > 0$, boundedness would supply a further subsequence converging to a point $x^\star \in \mathcal{C}$. Along that subsequence,

$$\operatorname{dist}(x^{k_j}, \mathcal{C}) \leq \|x^{k_j} - x^\star\| \longrightarrow 0,$$

a contradiction. Hence $\operatorname{dist}(x^k, \mathcal{C}) \to 0$.

Finally, $y_k = x^k - \mu m_k$ satisfies

$$\|y_k\| \leq \sup_i \|x^i\| + \mu \sup_i \|m_i\| < \infty, \qquad \|y_k - x^k\| = \mu \|m_k\| \to 0.$$

The function $h = f - \frac{\beta}{2}\|\cdot\|^2$ is continuous, so $F_\lambda = h \circ A + \lambda g_0$ is continuous. It is therefore uniformly continuous on a compact ball containing (x^k) and (y_k), and

$$F_\lambda(x^k) - F_\lambda(y_k) \to 0.$$

By the envelope identity,

$$F_\lambda(x^k) - E_\mu(x^k) = F_\lambda(x^k) - F_\lambda(y_k) - \frac{\mu}{2}\|m_k\|^2 \longrightarrow 0.$$

Since $E_\mu(x^k)$ converges, $F_\lambda(x^k)$ converges as well. All conclusions hold with probability one. □

E Bounded Primal Iterates of Algorithm 1

The following result shows that the boundedness assumption in Assumption 1 for Theorem 1 holds in the applications that we consider in this work.

Proposition 3. *Suppose $p_0 > 0$ is a probability density on $\mathbb{R}^d$ and $g_0 = -\log p_0$ has an L_g-Lipschitz gradient. Let $h : \mathbb{R}^m \to \mathbb{R}$ be finite, bounded below, and β-weakly convex. Choose $\lambda > 0$, $\gamma > 2\beta$, and $a > 0$. Set*

$$\tau_k = \frac{a}{(k+1)\log(k+2)}, \qquad \sigma_k^2 = \lambda \tau_k, \qquad \sum_k \rho_k^2 < \infty, \tag{48}$$

where (ρ_k) is nonnegative and deterministic. Then iterates in Algorithm 1 satisfy, almost surely,

$$\sup_k \big(\|x^k\| + \|w^k\| + \|z^{k+1}\|\big) < \infty. \tag{49}$$

The conclusion holds from any finite deterministic initial state and after any finite initial stage of well-defined updates.

We first bound the growth of a primal-dual energy. This estimate makes the measurement errors summable and allows us to bound the Moreau envelope. Coercivity of the MAP objective then gives boundedness of the iterates.

Set $r_k = Ax^k - z^{k+1}$, $G_k = A^\top w^{k+1} + \lambda \nabla g_0(x^k)$, and $\nu_k = \rho_k n^k$. Let $\mathcal{F}_k = \sigma(x^0, w^0, n^0, \ldots, n^{k-1})$ denote the information available before n^k is sampled. Constants C are deterministic and may change between inequalities. Constants C_ω may depend on the trajectory. All small-stepsize estimates below hold from a fixed deterministic index onward.

The posterior estimate in Appendix C gives, for $\bar{x} = D_\sigma(u)$,

$$\bar{x} = u - \sigma^2 \nabla g_0(\bar{x}) + \zeta, \qquad \|\zeta\| \leq \frac{L_g \sqrt{d}\, \sigma^3}{\sqrt{1 - L_g \sigma^2}}.$$

Using $\sigma_k^2 = \lambda\tau_k$ and Lipschitz continuity of ∇g_0, we obtain

$$\begin{aligned} x^{k+1} &= x^k - \tau_k G_k + \nu_k + \eta_k, \\ \|\eta_k\| &\le C\big(\tau_k^2\|G_k\| + \tau_k\|\nu_k\| + \tau_k^{3/2}\big), \\ \|x^{k+1} - x^k\| &\le C\big(\tau_k\|G_k\| + \|\nu_k\| + \tau_k^{3/2}\big). \end{aligned} \tag{50}$$

Indeed, $\eta_k = -\lambda\tau_k(\nabla g_0(x^{k+1}) - \nabla g_0(x^k)) + \zeta_k$, so the term $\lambda L_g \tau_k \|x^{k+1} - x^k\|$ can be moved to the left. The same posterior estimate gives

$$\|D_{\sqrt{\lambda\tau}}(u)\|^2 \le (1 + C\tau)\|u\|^2 + C\tau \tag{51}$$

for sufficiently small τ, uniformly in u.

We begin with a growth estimate that does not require bounded iterates.

Lemma E.1. *Set $Y_k = 1 + \|x^k\|^2 + (\tau_k/\gamma)\|w^k\|^2$ and $W_k = \|w^{k+1}\|$. For some deterministic $p > 1$, let $\Pi_k = (\log(k+2))^p$. Then, almost surely,*

$$\sup_k \frac{Y_k}{\Pi_k} < \infty, \qquad \sum_k \frac{\tau_k W_k}{\Pi_k} < \infty. \tag{52}$$

Moreover, $\tau_k W_k^2 \le C Y_k$ for all sufficiently large k.

Proof. Choose $\beta < \bar\beta < \gamma/2$. Weak convexity tested at $s/\|s\|$, for $s \in \partial h(z) \setminus \{0\}$, gives

$$\langle s, z\rangle \ge \|s\| - C - \frac{\bar\beta}{2}\|z\|^2. \tag{53}$$

Here we used lower boundedness of h, its boundedness on the unit ball, and $\beta(1+\|z\|)^2 \le \bar\beta\|z\|^2 + C$. The case $s = 0$ also satisfies this bound.

Set $\widehat{x}^k = x^k - \tau_k A^\top w^{k+1}$. Expanding the squares and using $w^{k+1} - w^k = \gamma r_k$, $w^{k+1} \in \partial h(z^{k+1})$, and $\|z^{k+1}\|^2 \le 2\|Ax^k\|^2 + 2\|r_k\|^2$ gives

$$\begin{aligned} 1 + \|\widehat{x}^k\|^2 + \frac{\tau_k}{\gamma}W_k^2 \le Y_k + C\tau_k(1 + \|x^k\|^2) - 2\tau_k W_k \\ - (\gamma - 2\bar\beta)\tau_k\|r_k\|^2 + \tau_k^2\|A\|^2 W_k^2. \end{aligned}$$

Apply (51) at $\widehat{x}^k + \nu_k$. Since $\mathbb{E}[\nu_k \mid \mathcal{F}_k] = 0$ and $\mathbb{E}[\|\nu_k\|^2 \mid \mathcal{F}_k] = d\rho_k^2$, this yields

$$\mathbb{E}[Y_{k+1} \mid \mathcal{F}_k] \le (1 + C_0\tau_k)Y_k - 2\tau_k W_k + C_0\rho_k^2 \tag{54}$$

eventually. The remaining coefficient of W_k^2 is bounded above by $(1 + C\tau_k)\tau_k^2\|A\|^2 - (\tau_k - \tau_{k+1})/\gamma$, which becomes negative because $(\tau_k - \tau_{k+1})/\tau_k^2 \to \infty$. Also, (29) and the dual update give $W_k \le C(1 + \|x^k\| + \|w^k\|)$, hence $\tau_k W_k^2 \le C Y_k$.

Choose $p > \max\{1, aC_0\}$. Then $\Pi_{k+1}/\Pi_k \ge 1 + C_0\tau_k$ eventually, so

$$\mathbb{E}\left[\frac{Y_{k+1}}{\Pi_{k+1}}\,\middle|\,\mathcal{F}_k\right] \le \frac{Y_k}{\Pi_k} - \frac{2\tau_k W_k}{\Pi_{k+1}} + \frac{C_0\rho_k^2}{\Pi_{k+1}}.$$

The normalized energy and the subtracted term are nonnegative, and the last term is summable. The Robbins-Siegmund convergence theorem (Robbins & Siegmund, 1971) therefore gives convergence of Y_k/Π_k and finiteness of $\sum_k \tau_k W_k/\Pi_{k+1}$ almost surely. Since $\Pi_{k+1}/\Pi_k \to 1$, this proves (52). □

The next lemma bounds the two error sums needed in the envelope estimate.

Lemma E.2. *Almost surely,*

$$\sum_k \tau_k^2\|G_k\|^2 < \infty, \qquad \sum_k \tau_k Y_k\|r_k\| < \infty. \tag{55}$$

Proof. Set $b_k = \sqrt{\tau_k}/\Pi_k^2$. Lemma E.1 gives $Y_k \le C_\omega \Pi_k$ and $W_k \le C_\omega \sqrt{\Pi_k/\tau_k}$. In particular,

$$\tau_k \Pi_k W_k + b_k \Pi_k W_k \le C_\omega \big(\sqrt{\tau_k}\Pi_k^{3/2} + \Pi_k^{-1/2}\big)$$

is bounded. Multiplying by the summable sequence $\tau_k W_k/\Pi_k$ controls the W_k^2 terms below. The Y_k terms are controlled by $\sum_k (\tau_k^2 \Pi_k + \tau_k^{3/2}/\Pi_k) < \infty$. Thus

$$\sum_k (\tau_k^2 + b_k \tau_k)(Y_k + W_k^2) < \infty. \tag{56}$$

Since $\|G_k\|^2 \le C(Y_k + W_k^2)$, the first assertion follows.

Set $L_k = \ell_\gamma(Ax^k, w^k)$ and $q_k = A^\top \nabla_a \ell_\gamma(Ax^k, w^{k+1})$. The global Lipschitz estimate (29) and the formulas for $\nabla \ell_\gamma$ in Lemma D.2 give $\|q_k\|^2 \le C(Y_k + W_k^2)$. Combining the lower Taylor bound for $\ell_\gamma(\cdot, w^{k+1})$ with (31) and (50) gives

$$\begin{aligned} \chi b_k \|r_k\|^2 \le b_k(L_{k+1} - L_k) + C b_k \tau_k (Y_k + W_k^2) \\ + C b_k \|\nu_k\|^2 - b_k \langle q_k, \nu_k \rangle, \end{aligned} \tag{57}$$

where $\chi = \gamma(\gamma - 2\beta)/(2(\gamma - \beta)) > 0$. The nonnegative error terms are summable by (56) and (35). Moreover, q_k is $\mathcal{F}_k$-measurable and $\|b_k q_k\|^2 \le C_\omega/\Pi_k^3$, so

$$\sum_k \mathbb{E}\big[b_k^2 \langle q_k, \nu_k \rangle^2 \,\big|\, \mathcal{F}_k\big] = \sum_k \rho_k^2 \|b_k q_k\|^2 < \infty \quad \text{almost surely.}$$

The conditional-variance argument in (40) therefore gives convergence of $\sum_k b_k \langle q_k, \nu_k \rangle$.

Evaluating the minimum defining ℓ_γ at zero gives $L_k \le C_\omega \Pi_k/\sqrt{\tau_k}$. Summation by parts yields

$$\sum_{k=K}^{N-1} b_k (L_{k+1} - L_k) = b_{N-1} L_N - b_K L_K + \sum_{k=K+1}^{N-1} (b_{k-1} - b_k) L_k.$$

These partial sums are bounded above because $b_{N-1} L_N \le C_\omega/\Pi_N$ and $(b_{k-1} - b_k) L_k \le C_\omega/((k+1)\Pi_k)$, with $p > 1$. Summing (57) therefore gives $\sum_k b_k \|r_k\|^2 < \infty$. Finally,

$$\sum_k \tau_k Y_k \|r_k\| \le \left(\sum_k b_k \|r_k\|^2\right)^{1/2} \left(\sum_k \frac{\tau_k^2 Y_k^2}{b_k}\right)^{1/2} < \infty,$$

since $\tau_k^2 Y_k^2/b_k \le C_\omega \tau_k^{3/2} \Pi_k^4$ is summable. □

We now use these error bounds to control the Moreau envelope.

Proof of Proposition 3. We first show that $F_\lambda = h \circ A + \lambda g_0$ is coercive. Let $B = B(0,1)$. Averaging the smoothness bound $g_0(x+v) \le g_0(x) + \langle \nabla g_0(x), v \rangle + L_g \|v\|^2/2$ over B cancels the linear term, giving

$$\frac{1}{|B|} \int_B g_0(x+v)\, dv \le g_0(x) + \frac{L_g}{2}.$$

Jensen's inequality then gives

$$p_0(x) \le \frac{e^{L_g/2}}{|B|} \int_{B(x,1)} p_0(y)\, dy \longrightarrow 0 \qquad \text{as } \|x\| \to \infty.$$

The integral tends to zero because p_0 is integrable. Thus $g_0(x) \to \infty$, and lower boundedness of h gives

$$F_\lambda(x) \ge \inf h + \lambda g_0(x) \longrightarrow \infty.$$

Consequently, the continuous function F_λ attains a finite minimum F_*.

Choose $\mu > 0$ such that

$$\mu \left(\lambda L_g + \beta \|A\|^2 + \frac{4\beta^2 \|A\|^2}{\gamma - 2\beta}\right) \le \frac{1}{2},$$

and use $y_k = \mathrm{prox}_{\mu F_\lambda}(x^k)$ and $m_k = \nabla E_\mu(x^k) = (x^k - y_k)/\mu$ from (32). Proximal optimality and (47) provide $s_k \in \partial h(Ay_k)$ with $A^\top s_k = m_k - \lambda \nabla g_0(y_k)$. Since y_k and m_k have at most linear growth in x^k, (53) gives

$$\|s_k\| \le C + \langle A^\top s_k, y_k\rangle + C\|Ay_k\|^2 \le CY_k.$$

Weak convexity at z^{k+1} and Ay_k now yields

$$\begin{aligned}\mu\langle G_k, m_k\rangle \ge \mu(1-\mu\lambda L_g)\|m_k\|^2 + \langle w^{k+1} - s_k, r_k\rangle \\ - \beta\|\mu A m_k - r_k\|^2.\end{aligned}$$

Substitute $\langle w^{k+1}, r_k\rangle = (\|w^{k+1}\|^2 - \|w^k\|^2)/(2\gamma) + \gamma\|r_k\|^2/2$. Young's inequality and the choice of μ give

$$\langle G_k, m_k\rangle \ge \frac{1}{2}\|m_k\|^2 + \frac{\|w^{k+1}\|^2 - \|w^k\|^2}{2\gamma\mu} - CY_k\|r_k\|. \tag{58}$$

The dual difference is canceled by adding a dual term to the envelope. Define the nonnegative energy

$$U_k = E_\mu(x^k) - F_* + \frac{\tau_k}{2\gamma\mu}\|w^k\|^2.$$

Apply the envelope Taylor bound to (50), using (58) and

$$|\langle m_k, \eta_k\rangle| \le \frac{\tau_k}{4}\|m_k\|^2 + C\big(\tau_k^3\|G_k\|^2 + \tau_k\|\nu_k\|^2 + \tau_k^2\big).$$

Since $\tau_{k+1} \le \tau_k$ and $\mathbb{E}[\langle m_k, \nu_k\rangle \mid \mathcal{F}_k] = 0$, we obtain

$$\begin{aligned}\mathbb{E}[U_{k+1} \mid \mathcal{F}_k] \le U_k + \varepsilon_k, \\ \varepsilon_k = C\big(\tau_k Y_k\|r_k\| + \tau_k^2\|G_k\|^2 + \tau_k^2 + \rho_k^2\big).\end{aligned} \tag{59}$$

Here U_k and ε_k are nonnegative and $\mathcal{F}_k$-measurable. Lemma E.2 and (48) give $\sum_k \varepsilon_k < \infty$ almost surely. The Robbins-Siegmund convergence theorem again gives convergence of U_k to a finite limit.

It follows that $E_\mu(x^k)$ is bounded above. By the envelope identity,

$$F_\lambda(y_k) \le E_\mu(x^k), \qquad \|x^k - y_k\|^2 \le 2\mu\big(E_\mu(x^k) - F_*\big).$$

Coercivity of F_λ bounds y_k, and the second inequality then bounds x^k. Lemma D.1 bounds w^k and z^{k+1}, proving (49) with probability one.

A finite initial stage ends at a finite state. Conditioning on that state applies the preceding estimates from any sufficiently late deterministic index. The measurement map and the small-noise denoiser have linear growth, so the finite-time moments used above are finite for each fixed starting state. This proves the assertion concerning initial stages. Finally, (48) gives $\sum_k \tau_k = \infty$ and $\sum_k \tau_k^{3/2} < \infty$, so the schedule conditions of Theorem 1 also hold. □

Verification for the applications in this work. We verify that Proposition 3 applies to the quadratic, transmission-CT, and phase-retrieval applications.

For linear observations, $h(z) = \|z - y\|^2/(2s^2)$ is nonnegative and convex. For Poisson transmission CT,

$$h(z) = \sum_i (I_{0,i}e^{-z_i} + y_i z_i), \qquad I_{0,i} > 0, \quad y_i \ge 0,$$

we have $h_i''(z_i) > 0$ and $\inf h = \sum_{i:y_i>0} y_i(1 + \log(I_{0,i}/y_i)) > -\infty$. Zero-count terms have infimum zero. For Gaussian transmission CT,

$$h(z) = \frac{1}{2s^2}\sum_i (I_{0,i}e^{-z_i} - y_i)^2 \ge 0, \qquad h_i''(t) = \frac{2u^2 - y_i u}{s^2} \ge -\frac{(y_i)_+^2}{8s^2}, \quad u = I_{0,i}e^{-t}.$$

Thus $\beta = \|y_+\|_\infty^2/(8s^2)$ suffices. For phase retrieval, using real coordinates for each complex measurement,

$$h(z) = \frac{\eta}{2}\sum_i (|z_i|^2 - y_i)^2 \ge 0, \qquad \nabla^2 h_i(z_i) = 2\eta(|z_i|^2 - y_i)I + 4\eta z_i z_i^\top \succeq -2\eta(y_i)_+ I.$$

Hence $\beta = 2\eta\|y_+\|_\infty$ is valid. The nonsmooth intensity loss $h(z) = \eta\sum_i \big||z_i|^2 - y_i\big|$ is also nonnegative, and

$$h(z) + \eta\|z\|^2 = \eta\sum_i \max\{2|z_i|^2 - y_i,\ y_i\}$$

is convex. Here $\beta = 2\eta$ suffices for real or complex measurements.

F PRIORS ON AN AFFINE SUBSPACE

We extend Theorem 1 to priors supported on an affine subspace. Expressing the denoiser and measurement updates in orthonormal coordinates on the subspace allows us to extend the ideas of the Theorem 1 proof to the setting of our MAP objective restricted on the subspace.

Proposition 4. *Replace the prior in Assumption 1 by a positive density p_S on an affine subspace S of dimension $1 \le r \le d$, with respect to its induced volume measure. Suppose $g_S = -\log p_S$ has an L_g-Lipschitz intrinsic gradient. Define $F_{\lambda,S}(x) = h(Ax) + \lambda g_S(x)$ for $x \in S$ and $F_{\lambda,S}(x) = +\infty$ otherwise. Under the remaining assumptions and schedule conditions of Theorem 1, $x^k \in S$ for $k \ge 1$, and the convergence conclusions of that theorem hold for $F_{\lambda,S}$ with $\mathcal{C} = \operatorname{crit} F_{\lambda,S}$.*

Proof. Choose $b \in S$ and an orthonormal basis $U \in \mathbb{R}^{d\times r}$ for the direction space of S, so that

$$S = b + \operatorname{range} U, \qquad U^\top U = I_r.$$

The coordinate map $s \mapsto b + Us$ preserves distances and volume. Consequently,

$$q(s) = p_S(b + Us), \qquad g(s) = -\log q(s) = g_S(b + Us)$$

define a positive probability density on $\mathbb{R}^r$ and an energy with an L_g-Lipschitz gradient.

We first express the denoiser in these coordinates. For $t = U^\top(v - b)$, orthogonal decomposition gives

$$\|v - b - Us\|^2 = \|t - s\|^2 + \|(I - UU^\top)(v - b)\|^2.$$

The second term does not depend on s, so it cancels when the posterior density is normalized. The posterior in intrinsic coordinates is therefore proportional to $q(s)\exp(-\|t - s\|^2/(2\sigma^2))$. Writing D^q_σ for its mean gives

$$D_\sigma(v) = b + UD^q_\sigma(U^\top(v - b)). \tag{60}$$

In particular, every denoising update lies in S, so $x^k \in S$ for $k \ge 1$.

The same decomposition gives the ambient smoothed density

$$p_\sigma(v) = \frac{\exp\!\big(-\|(I - UU^\top)(v - b)\|^2/(2\sigma^2)\big)}{(2\pi\sigma^2)^{(d-r)/2}}\, q_\sigma(U^\top(v - b)), \tag{61}$$

where $q_\sigma = q * \mathcal{N}(0, \sigma^2 I_r)$. The normal Gaussian factor contributes $-(I - UU^\top)(v - b)/\sigma^2$ to the ambient score. Thus Tweedie's identity also gives (60).

We now express the iteration on S. For $k \ge 1$, write $x^k = b + Us^k$. Substitution of (60) into the primal update gives

$$s^{k+1} = D^q_{\sigma_k}\big(s^k - \tau_k(AU)^\top w^{k+1} + \rho_k\zeta^k\big), \qquad \zeta^k = U^\top n^k. \tag{62}$$

Since $U^\top U = I_r$, the projected noises are independent $\mathcal{N}(0, I_r)$ vectors, and each ζ^k is independent of the preceding iterates.

The offset Ab is absorbed into the fidelity by setting

$$\widehat{h}(z) = h(z + Ab), \qquad \widehat{z}^{k+1} = z^{k+1} - Ab.$$

Changing variables in the measurement proximal problem gives

$$\begin{aligned} \widehat{z}^{k+1} &= \operatorname{prox}_{\widehat{h}/\gamma}\left(AUs^k + \frac{w^k}{\gamma}\right), \\ w^{k+1} &= w^k + \gamma(AUs^k - \widehat{z}^{k+1}). \end{aligned} \tag{63}$$

Translation preserves finite-valuedness and β-weak convexity of h. Moreover, $\|s^k\| = \|x^k - b\|$, so the assumed boundedness of (x^k) gives boundedness of (s^k).

Equations (62) and (63) are Algorithm 1 in $\mathbb{R}^r$, with prior q, measurement operator AU, and fidelity $\widehat{h}$. Theorem 1 therefore applies to

$$f_\lambda(s) = \widehat{h}(AUs) + \lambda g(s) = h(Ab + AUs) + \lambda g(s).$$

We then obtain, almost surely,

$$\operatorname{dist}(s^k, \operatorname{crit} f_\lambda) \to 0, \qquad f_\lambda(s^k) \text{ converges.}$$

It remains to express these conclusions in the ambient space. For $x = b + Us$, the subdifferential is

$$\partial F_{\lambda,S}(x) = U\partial f_\lambda(s) + \operatorname{range}(U)^\perp.$$

The two terms lie in orthogonal subspaces, so

$$\begin{aligned} x \in \operatorname{crit} F_{\lambda,S} \quad &\iff \quad s \in \operatorname{crit} f_\lambda \\ &\iff \quad 0 \in (AU)^\top \partial h(Ab + AUs) + \lambda \nabla g(s). \end{aligned} \tag{64}$$

Thus $\mathcal{C} = b + U \operatorname{crit} f_\lambda$. Since the coordinate map preserves distances,

$$\operatorname{dist}(x^k, \mathcal{C}) = \operatorname{dist}(s^k, \operatorname{crit} f_\lambda), \qquad F_{\lambda,S}(x^k) = f_\lambda(s^k).$$

Both convergence conclusions follow. □

G Learned-Score Approximation

We now quantify how a bounded score error affects the MAP first-order condition. The proof retains this error in the Moreau-envelope estimate of Theorem 1 and shows that an accumulation point has a stationarity residual proportional to the score error.

Let s_θ be a fixed measurable score model and set $D_{\theta,\sigma}(u) = u + \sigma^2 s_\theta(u, \sigma)$. At the input $u^{k+1} = v^{k+1} + \rho_k n^k$, define

$$\delta_k = s_\theta(u^{k+1}, \sigma_k) - \nabla \log p_{\sigma_k}(u^{k+1}).$$

Proposition 5. *Assume the prior and fidelity regularity, parameters, and schedules of Theorem 1, and use D_{θ,σ_k} in Algorithm 1. Suppose the resulting primal iterates are almost surely bounded and $\sup_k \|\delta_k\| \le \varepsilon$ almost surely, where $\varepsilon \ge 0$. Then the sequence has, almost surely, an accumulation point $x^\star$ satisfying*

$$\operatorname{dist}\big(0, \partial F_\lambda(x^\star)\big) \le \lambda\varepsilon. \tag{65}$$

Proof. Set $\nu_k = \rho_k n^k$ and let $\mathcal{F}_k = \sigma(x^0, w^0, n^0, \dots, n^{k-1})$ denote the information available before n^k is sampled. With $g_\sigma = -\log p_\sigma$, define

$$G_k = A^\top w^{k+1} + \lambda \nabla g_0(x^k), \qquad e_k = \lambda\big(\nabla g_{\sigma_k}(u^{k+1}) - \nabla g_0(x^k)\big).$$

The identity $D_{\theta,\sigma_k}(u^{k+1}) = D_{\sigma_k}(u^{k+1}) + \sigma_k^2 \delta_k$ and $\sigma_k^2 = \lambda\tau_k$ give

$$x^{k+1} = x^k - \tau_k d_k + \nu_k, \qquad d_k = G_k + e_k - \lambda\delta_k. \tag{66}$$

Lemma D.1 bounds (w^k) and (z^{k+1}). By (35), (ν_k) is bounded and $\nu_k \to 0$, so the denoiser inputs are bounded as well. The score estimate (36) gives

$$\|e_k\| \le C(\sigma_k + \tau_k + \|\nu_k\|) \longrightarrow 0.$$

Thus (d_k) is bounded. The schedule conditions and (35) also give

$$\sum_k \|x^{k+1} - x^k\|^2 \le 2 \sup_k \|d_k\|^2 \sum_k \tau_k^2 + 2 \sum_k \|\nu_k\|^2 < \infty. \tag{67}$$

Since the primal increments vanish, Lemma D.2 gives

$$r_k = Ax^k - z^{k+1} \longrightarrow 0, \qquad \Psi_k = \Psi(Ax^k, w^k) \longrightarrow 0.$$

We next compare the update direction with the envelope gradient. Fix $\mu > 0$ with $\mu\kappa < 1$, where $\kappa = \beta\|A\|^2 + \lambda L_g$, and use the Moreau envelope E_μ from (32). Set

$$y_k = \operatorname{prox}_{\mu F_\lambda}(x^k), \qquad m_k = \nabla E_\mu(x^k) = \frac{x^k - y_k}{\mu}.$$

Proximal optimality gives $m_k \in \partial F_\lambda(y_k)$. Weak convexity at y_k then yields

$$F_\lambda(x^k) - F_\lambda(y_k) \geq \mu\left(1 - \frac{\mu\kappa}{2}\right)\|m_k\|^2. \tag{68}$$

Combining this with the lower model (34) at y_k and $0 \leq B_k \leq \Psi_k$ from (30) gives

$$\langle G_k, m_k\rangle \geq (1-\mu\kappa)\|m_k\|^2 + \beta\langle r_k, Am_k\rangle - \frac{\Psi_k}{\mu}. \tag{69}$$

The score-error bound gives $\lambda\langle\delta_k, m_k\rangle \leq \lambda\varepsilon\|m_k\|$. Since (x^k) is bounded and ∇E_μ is Lipschitz, (m_k) is bounded. The remaining errors satisfy

$$\eta_k = \big(\|e_k\| + \beta\|A\|\|r_k\|\big)\|m_k\| + \frac{\Psi_k}{\mu} \longrightarrow 0.$$

The descent lemma applied to (66) therefore gives

$$\begin{aligned} E_\mu(x^{k+1}) \leq E_\mu(x^k) &- \tau_k\big((1-\mu\kappa)\|m_k\|^2 - \lambda\varepsilon\|m_k\| - \eta_k\big) \\ &+ \langle m_k, \nu_k\rangle + \frac{L_\mu}{2}\|x^{k+1} - x^k\|^2, \end{aligned} \tag{70}$$

where L_μ is a Lipschitz constant for ∇E_μ. The quadratic remainder is summable by (67). Moreover, m_k is $\mathcal{F}_k$-measurable and bounded, so the martingale argument for (40) gives convergence of $\sum_k \langle m_k, \nu_k\rangle$ with probability one.

We claim that

$$\liminf_{k\to\infty} \|\nabla E_\mu(x^k)\| \leq \frac{\lambda\varepsilon}{1-\mu\kappa}. \tag{71}$$

Otherwise, the expression in parentheses in (70) would eventually be bounded below by a positive constant, since $\eta_k \to 0$. Summing that inequality and using $\sum_k \tau_k = \infty$ would force $E_\mu(x^k) \to -\infty$. This contradicts its lower bound on the bounded trajectory.

To obtain a subgradient of F_λ itself, choose a deterministic sequence $\mu_j \downarrow 0$ with $\mu_j\kappa < 1$. The bound (71) holds simultaneously for this countable sequence with probability one. Choose increasing indices k_j such that

$$\|\nabla E_{\mu_j}(x^{k_j})\| \leq \frac{\lambda\varepsilon}{1-\mu_j\kappa} + \frac{1}{j},$$

and set

$$y_j = \operatorname{prox}_{\mu_j F_\lambda}(x^{k_j}), \qquad \xi_j = \nabla E_{\mu_j}(x^{k_j}).$$

Then

$$\xi_j \in \partial F_\lambda(y_j), \qquad \|x^{k_j} - y_j\| = \mu_j\|\xi_j\| \longrightarrow 0, \qquad \limsup_j \|\xi_j\| \leq \lambda\varepsilon. \tag{72}$$

By boundedness, pass to a subsequence with $x^{k_j} \to x^\star$ and $\xi_j \to \xi^\star$. Then $y_j \to x^\star$ as well. For every $z \in \mathbb{R}^d$, weak convexity gives

$$F_\lambda(z) \geq F_\lambda(y_j) + \langle \xi_j, z - y_j\rangle - \frac{\kappa}{2}\|z - y_j\|^2.$$

Passing to the limit and using continuity of F_λ gives $\xi^\star \in \partial F_\lambda(x^\star)$. Together with $\|\xi^\star\| \leq \lambda\varepsilon$, this proves (65) with probability one. □

H Proof of Theorem 2

Case (ii) requires a bound on the distance of the denoised iterates from S. We first establish this bound and then use the Gaussian KL formula to prove the theorem. Throughout, let $\mathcal{F}_k = \sigma(x^0, w^0, n^0, \ldots, n^{k-1})$ denote the information available before n^k is sampled.

Lemma H.1. *Under case (ii) of Theorem 2, suppose $\rho_k = \sigma_k$. Then*

$$\operatorname{dist}(x^{k+1}, S) = O\big(\sigma_k^2 \log(e/\sigma_k)\big) \qquad \textit{with probability one.}$$

Proof. We first show that the iterates come arbitrarily close to S. We then show that denoising keeps them close at the stated rate.

Let Y have the Gaussian posterior at input u and level σ, with normalizing constant

$$Z_{\sigma,u} = \int_S e^{-\|y-u\|^2/(2\sigma^2)} p_0(dy).$$

Since $\mathbb{E}e^{\|Y-u\|^2/(2\sigma^2)} = 1/Z_{\sigma,u}$, Jensen's inequality gives

$$\mathbb{E}\|Y-u\|^2 \le -2\sigma^2 \log Z_{\sigma,u}. \tag{73}$$

Choose R with $m = p_0(B(0,R)) > 0$. The bound $Z_{\sigma,u} \ge m e^{-(\|u\|+R)^2/(2\sigma^2)}$ implies

$$\mathbb{E}\|Y-u\|^2 \le (\|u\|+R)^2 + 2\sigma^2 \log(1/m).$$

Consequently, $\|D_\sigma(u)\| \le 2\|u\| + C$ for $0 < \sigma \le 1$.

The posterior variance controls the distance of its mean from S. Differentiating the posterior mean gives

$$\sigma^2 \nabla D_\sigma(u) = \mathrm{Cov}(Y), \qquad \mathrm{dist}(D_\sigma(u), S)^2 \le \mathbb{E}\|Y - \mathbb{E}Y\|^2 = \sigma^2 \,\mathrm{div}\, D_\sigma(u).$$

The moment bound also gives polynomial growth of ∇D_σ for each fixed σ. Gaussian integration by parts therefore yields, for $n \sim \mathcal{N}(0, I_d)$ and $\|v\| \le B$,

$$\mathbb{E}\,\mathrm{dist}(D_\sigma(v+\sigma n), S)^2 \le \sigma \mathbb{E}\langle n, D_\sigma(v+\sigma n)\rangle \le C_B \sigma. \tag{74}$$

Lemma D.1 bounds (w^k), so (v^{k+1}) is bounded with probability one. For large k, applying (74) conditionally on $\mathcal{F}_k$ gives

$$\mathbb{E}\big[\mathbf{1}_{\{\|v^{k+1}\| \le B\}}\, \mathrm{dist}(x^{k+1}, S)^2\big] \le C_B \sigma_k \longrightarrow 0.$$

By Fatou's lemma, we have

$$\begin{aligned} 0 &\le \mathbb{E}\left[\liminf_{k\to\infty} \mathbf{1}_{\{\|v^{k+1}\| \le B\}}\, \mathrm{dist}(x^{k+1}, S)^2\right] \\ &\le \liminf_{k\to\infty} \mathbb{E}\left[\mathbf{1}_{\{\|v^{k+1}\| \le B\}}\, \mathrm{dist}(x^{k+1}, S)^2\right] \\ &\le \liminf_{k\to\infty} C_B \sigma_k = 0 \end{aligned}$$

Boundedness of (v^{k+1}) therefore gives

$$\liminf_k \mathrm{dist}(x^k, S) = 0 \qquad \text{with probability one.} \tag{75}$$

We next establish two geometric estimates near S. The growth bound above places the relevant nearest points in a compact part of S. On this part, the exterior-ball property of a C^2 embedded manifold gives a uniform radius $R_0 > 0$ (Federer, 1959). Specifically, if q is a nearest point to $u \ne q$, the open ball of radius R_0 centered at $q + R_0(u-q)/\|u-q\|$ does not intersect S. Thus, for every $y \in S$,

$$R_0^2 \le \left\| y - q - R_0 \frac{u-q}{\|u-q\|} \right\|^2 = \|y-q\|^2 + R_0^2 - \frac{2R_0\langle u-q, y-q\rangle}{\|u-q\|}.$$

Rearranging gives the first estimate below, with $K = 1/(2R_0)$. For $u = q$, it holds trivially.

For the second estimate, small coordinate neighborhoods have volume bounded below by a constant times t^r, including at boundary points. The positive continuous density is bounded below on the compact region. Consequently, for uniform constants $K, c, t_0 > 0$,

$$\begin{aligned} \langle u-q, y-q\rangle &\le K\|u-q\|\,\|y-q\|^2 \qquad (y \in S), \\ p_0(B(q,t)) &\ge c t^r \qquad (0 < t < t_0), \end{aligned} \tag{76}$$

where $r = \dim S$.

Choose $\epsilon > 0$ with $2K\epsilon \le 1/2$. For a bounded input u with $\operatorname{dist}(u, S) < \epsilon$, set $L(y) = \|y - u\|^2 - \|q - u\|^2$. Then

$$L(y) \ge \tfrac{1}{2}\|y - q\|^2, \qquad L(y) \le \sigma^4 + 2\epsilon\sigma^2 \quad \text{when } \|y - q\| \le \sigma^2.$$

For small σ, factoring out the likelihood at q gives

$$\begin{aligned}\widetilde{Z}_{\sigma,u} := e^{\|q-u\|^2/(2\sigma^2)} Z_{\sigma,u} &= \int_S e^{-L(y)/(2\sigma^2)} p_0(dy) \\ &\ge e^{-\epsilon-\sigma^2/2} p_0(B(q,\sigma^2)) \\ &\ge c\, e^{-\epsilon-\sigma^2/2}\sigma^{2r} \ge c'\sigma^{2r}.\end{aligned}$$

Subtracting $\|q - u\|^2$ in (73) gives

$$\mathbb{E}\|Y - q\|^2 \le 2\mathbb{E}L(Y) \le -4\sigma^2 \log \widetilde{Z}_{\sigma,u} \le C\sigma^2 \log(e/\sigma). \tag{77}$$

Write $b = \mathbb{E}Y$ and $\delta = \operatorname{dist}(b, S)$. Equation (77) first gives $\delta \le \|b - q\| \to 0$ uniformly for these inputs. Applying (76) at a nearest point s to b, averaging over Y, and dividing by $\delta > 0$ gives

$$\delta \le K\mathbb{E}\|Y - s\|^2 = K\mathbb{E}\|Y - b\|^2 + K\delta^2 \le C\sigma^2 \log(e/\sigma) + K\delta^2.$$

For small enough σ, $K\delta \le 1/2$. Moving the last term to the left gives

$$\operatorname{dist}(D_\sigma(u), S) \le C\sigma^2 \log(e/\sigma). \tag{78}$$

The case $\delta = 0$ satisfies the same bound.

It remains to show that the iterates eventually stay in this neighborhood. The assumption $\sum_k \sigma_k^3 < \infty$ gives $\sum_k \mathbb{E}\|\sigma_k n^k\|^3 < \infty$, hence $\sigma_k n^k \to 0$ with probability one. For the input $u^{k+1} = v^{k+1} + \sigma_k n^k$, we therefore have

$$u^{k+1} - x^k = -\frac{\sigma_k^2}{\lambda} A^\top w^{k+1} + \sigma_k n^k \longrightarrow 0.$$

The inputs are bounded, so (78) holds uniformly near S in a region containing them. Eventually, both $\|u^{k+1} - x^k\|$ and $C\sigma_k^2 \log(e/\sigma_k)$ are less than $\epsilon/3$. Equation (75) supplies a later iterate within $\epsilon/3$ of S. Its input is within $2\epsilon/3$, and (78) places the next iterate within $\epsilon/3$. Induction proves the claimed bound with probability one. □

Proof of Theorem 2. Lemma D.1 gives

$$\|v^{k+1} - x^k\| = \frac{\sigma_k^2}{\lambda}\|A^\top w^{k+1}\| = O(\sigma_k^2) \qquad \text{with probability one.} \tag{79}$$

Choose $s^k \in S$ with

$$\|s^k - x^k\| < \operatorname{dist}(x^k, S) + \sigma_k^2.$$

Taking the first qualifying element of a fixed countable dense subset of S makes this choice $\mathcal{F}_k$-measurable.

Conditionally on $\mathcal{F}_k$, the denoiser input has distribution $\mathcal{N}(v^{k+1}, \rho_k^2 I_d)$. For $\rho_k > 0$, the Gaussian KL formula gives

$$J_k = \frac{\|s^k - v^{k+1}\|^2}{2\rho_k^2} + \frac{d}{2}\left(\frac{\sigma_k^2}{\rho_k^2} - 1 - \log\frac{\sigma_k^2}{\rho_k^2}\right). \tag{80}$$

Suppose first that $\rho_k = \sigma_k$. In case (i), $\operatorname{dist}(x^k, S) = 0$ eventually. In case (ii), Lemma H.1 and the bound on consecutive noise ratios give

$$\operatorname{dist}(x^k, S) \le C\sigma_{k-1}^2 \log(e/\sigma_{k-1}) \le C'\sigma_k^2 \log(e/\sigma_k) = o(\sigma_k).$$

Together with (79), either case gives

$$\frac{\|s^k - v^{k+1}\|}{\sigma_k} \le \frac{\operatorname{dist}(x^k, S)}{\sigma_k} + \sigma_k + \frac{\sigma_k}{\lambda}\|A^\top w^{k+1}\| \longrightarrow 0.$$

The covariance term in (80) vanishes, so $J_k \to 0$ with probability one.

Conversely, if $\rho_k = 0$, the second Gaussian is a point mass, whereas the first has positive variance. Hence $J_k = +\infty$. Thus $J_k \to 0$ would require $\rho_k > 0$ eventually and the covariance term in (80) to tend to zero. The function $t - 1 - \log t$ has its unique zero at 1 and diverges as $t \to 0$ or $t \to \infty$, so this would imply $\sigma_k^2/\rho_k^2 \to 1$. In particular, $\rho_k^2 \geq \sigma_k^2/2$ eventually. This contradicts the schedule conditions

$$\sum_k \sigma_k^2 = \infty, \qquad \sum_k \rho_k^2 < \infty$$

of Theorem 1. Consequently, $J_k \not\to 0$ under those conditions, regardless of the reference points in S. □

I Connection to Subgradient Langevin Dynamics

We retain the prior and fidelity regularity of Theorem 1, with $\lambda > 0$, $\gamma > 2\beta$. Let $\rho_k = \sigma_k$, $\tau_k = \sigma_k^2/\lambda$, and $\sigma_k \to 0$. The comparison below identifies local drift and noise coefficients as the measurement residual $r_k = Ax^k - z^{k+1}$ vanishes.

Write $u^{k+1} = v^{k+1} + \sigma_k n^k$ and $g_\sigma = -\log p_\sigma$. Tweedie's identity gives the recursion from (37),

$$x^{k+1} = x^k - \tau_k(G_k + e_k) + \sqrt{\lambda\tau_k}\, n^k, \tag{81}$$

where

$$G_k = A^\top w^{k+1} + \lambda \nabla g_0(x^k), \qquad e_k = \lambda\big(\nabla g_{\sigma_k}(u^{k+1}) - \nabla g_0(x^k)\big).$$

On bounded sets of current primal-dual states, $u^{k+1} - x^k = -\tau_k A^\top w^{k+1} + \sigma_k n^k \to 0$ in probability. The uniform score estimate in the proof of Proposition 2, together with Lipschitz continuity of ∇g_0, therefore gives $e_k \to 0$ in probability.

The measurement optimality condition (22) gives $w^{k+1} \in \partial h(z^{k+1})$. Consider any convergent sequence of primal-dual states $(x^{k_j}, w^{k_j+1}) \to (\bar{x}, \bar{w})$ along which $r_{k_j} \to 0$. Then $z^{k_j+1} \to A\bar{x}$. Since h is finite and weakly convex, its subdifferential has a closed graph, so $\bar{w} \in \partial h(A\bar{x})$. The chain rule (47) now yields

$$G_{k_j} \longrightarrow \bar{\xi} := A^\top \bar{w} + \lambda \nabla g_0(\bar{x}) \in \partial F_\lambda(\bar{x}).$$

Thus vanishing measurement residuals identify the limiting drift with subgradients of the original MAP objective, without differentiability of h.

Subgradient Langevin dynamics at temperature T take the form (Johnston et al., 2025)

$$dX_t = -\xi_t\, dt + \sqrt{2T}\, dB_t, \qquad \xi_t \in \partial F_\lambda(X_t), \tag{82}$$

where B_t is standard Brownian motion. Over a time interval of length τ_k, the Brownian increment has covariance $\tau_k I_d$. Comparing (81) with (82) gives

$$T_k = \frac{\rho_k^2}{2\tau_k} = \frac{\sigma_k^2}{2\tau_k} = \frac{\lambda}{2}.$$

Consequently, when the splitting and denoising errors vanish locally, matched noise gives the drift and noise scaling of subgradient Langevin dynamics at fixed temperature $\lambda/2$.

J Simulation Settings and Hyperparameters

J.1 Data and acquisition settings

All experiments were run on Google Colab using an NVIDIA A100 GPU with 40GB of memory, a 12-vCPU Intel Xeon CPU @ 2.20GHz, and 83 GiB of system memory. For FFHQ tasks we consider six inverse problems with the following configurations:

- **Super-resolution.** The low-resolution measurements are generated by $4\times$ downsampling using cubic interpolation.

- **Gaussian deblurring.** The degraded observation is obtained by applying a Gaussian blur with a kernel size of 61 and std 3.
- **Motion deblurring.** The measurement is formed by blurring the image with a motion-blur kernel of size 61, using blur std 0.5.
- **Box inpainting.** A square missing region of size 128×128 is placed near the center of the image, while preserving a spatial margin of 32 pixels from the image boundaries.
- **Random inpainting.** The observation is generated by removing 70% of the pixels uniformly at random.
- **Phase retrieval.** The measurements are generated from phaseless observations using an oversampling factor of 2.

The FFHQ measurements have Gaussian noise with variance $\sigma_n^2 = 0.05$. For the baselines used in FFHQ, we use the default hyperparameter settings for each inverse problem as stated in the papers.

For CT, we consider the following sparse-view reconstruction setting:

- **Sparse-view transmission CT.** We use 100 axial CT slices. Measurements are simulated with the ASTRA (van Aarle et al., 2016) parallel-beam CT operator using 80 projection angles and 512 detector bins. We use an intensity $I_0 = 5000$ with an average transmittance of 0.5 for each slice.

J.2 Schedule and parameters

We use a two-part annealing schedule. For the first $K - 20$ iterations, the denoising level decreases linearly from $\sigma_{\max} = 10$ to σ_0,

$$\sigma_k = 10 - \frac{k}{K-20}(10 - \sigma_0), \qquad 0 \le k < K - 20.$$

During these iterations, we set

$$\rho_k = \sigma_k, \qquad \tau_k = \frac{\sigma_0^2}{\lambda}.$$

For the final 20 iterations, we switch to the schedule used in Theorem 1. Indexing these iterations by $j = 0, \dots, 19$, we use

$$\sigma_j = \frac{\sigma_0}{\sqrt{(j+1)\log(j+2)}}, \qquad \tau_j = \frac{\sigma_j^2}{\lambda}, \qquad \rho_j = \frac{\sigma_0}{j+1}.$$

Hyperparameter tuning was performed on 100 validation images from each dataset. The resulting values of σ_0, λ, γ, and K used for each inverse problem are reported in Table 4.

K Additional Experiments

K.1 Ablation Study

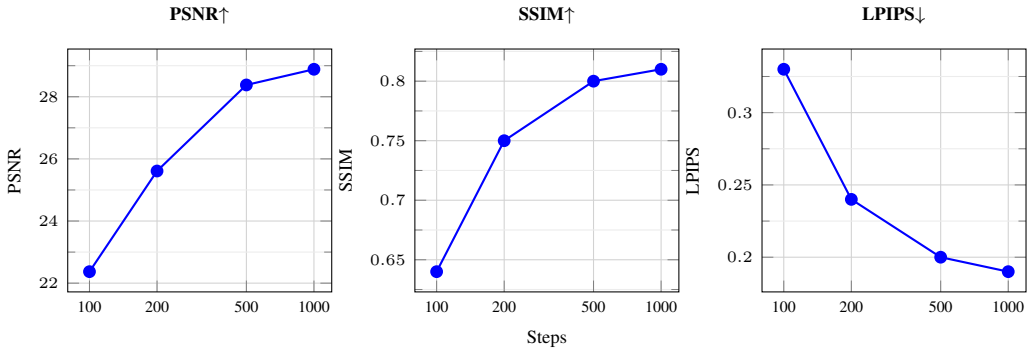


Figure 5: Effect of the number of iterations on reconstruction metrics for phase retrieval.

Table 4: Parameters used in the experiments. The denoising level is decreased linearly from 10 to σ_0, followed by 20 iterations using the Theorem 1 schedule.

Task	σ_0	λ	γ	K
Box inpainting	0.05	2.5	500	200
Random inpainting	0.05	2.5	500	100
Gaussian deblurring	0.10	1	500	200
Super-resolution $4\times$	0.01	0.01	500	200
Phase retrieval	0.001	10^{-4}	1100	500
Motion deblurring	0.02	0.16	500	200
Sparse-view transmission CT	0.10	1	500	200

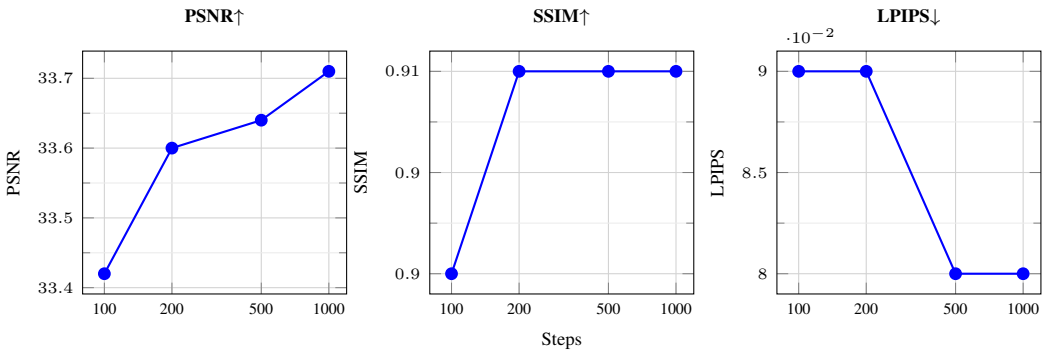


Figure 6: Effect of the number of iterations on reconstruction metrics for random inpainting.

The ablations examine sensitivity to the number of iterations and the dual parameter γ. Since each primal-dual iteration uses one score-model evaluation, varying the number of iterations K directly varies the number of function evaluations.

Figure 5 reports the effect of K on the nonlinear phase-retrieval task. We evaluate $K \in \{100, 200, 500, 1000\}$ and report PSNR, SSIM, and LPIPS for the final reconstruction. The reconstruction quality improves consistently as the number of steps increases. PSNR and SSIM increase, while LPIPS decreases. The largest gains occur when increasing from 100 to 500 steps, whereas the improvement from 500 to 1000 steps is more moderate. This indicates that phase retrieval benefits from a larger iteration budget, but that the performance begins to saturate after several hundred iterations.

Figure 6 shows the same step-count ablation for random inpainting, which is a linear inverse problem. In this case the metrics are much less sensitive to the number of steps. Increasing K improves PSNR and LPIPS slightly, while SSIM quickly saturates. This suggests that, for easier linear degradation models such as random inpainting, our approach obtains high-quality reconstructions with a small number of score evaluations.

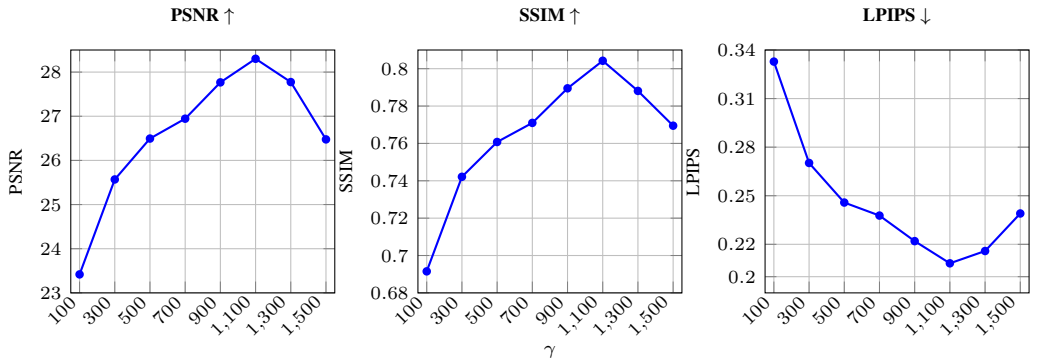


Figure 7: Reconstruction metrics for various dual parameters γ on 100 FFHQ images with phase retrieval.

We also examine the influence of the PDHG dual parameter γ on the phase-retrieval problem. Figure 7 reports a sweep over several values of γ. The best value of γ is obtained in an intermediate range, as performance seems to deteriorate when γ is too small or too high.

K.2 Linear Experiment

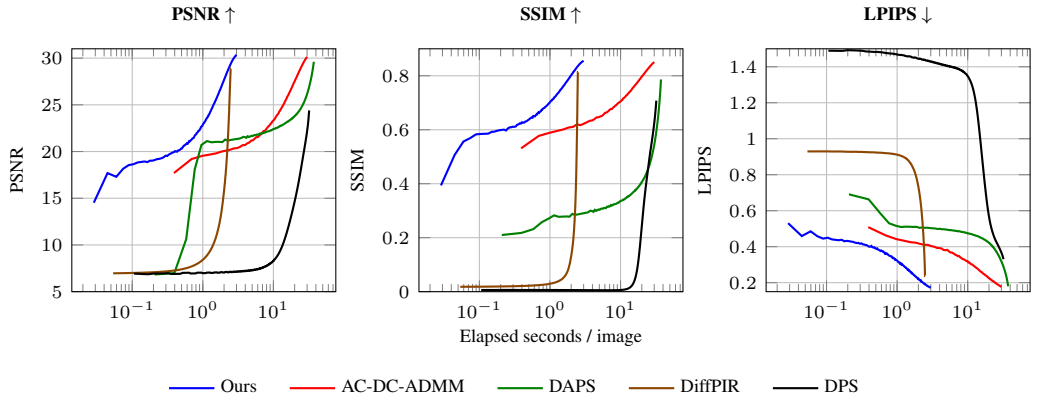


Figure 8: Reconstruction metrics over elapsed time for 100 images with Gaussian blur.

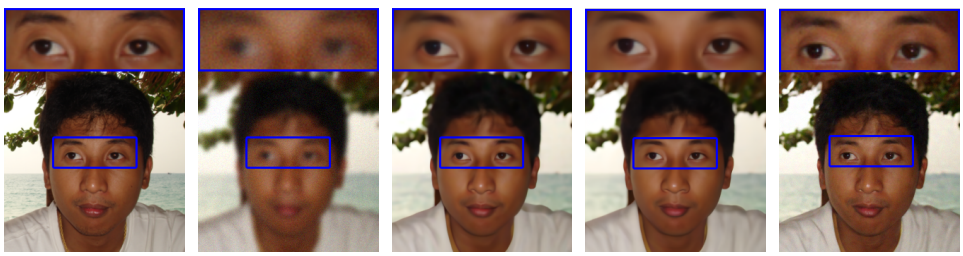


Figure 9: Gaussian blur.

We further report a runtime comparison on the Gaussian deblurring task, which is a linear inverse problem. Figure 8 plots PSNR, SSIM, and LPIPS as functions of elapsed time per image for the primal-dual implementation and the diffusion-based baselines. Our primal-dual approach reaches high PSNR and SSIM values within a shorter time window than the competing methods, even in the linear case. The LPIPS curve shows a similar trend in decrease.

Figure 9 gives a qualitative comparison with the diffusion baselines.

K.3 Reconstruction Statistics

We report histograms for the spread of PSNR, SSIM and LPIPS in various inverse problems. Since the spreads for the linear problems are similar, we use Gaussian blur as a representative. We also consider phase retrieval and CT transmission.

For Gaussian blur, the histograms show consistent performance across the 100 test images. PSNR is centered around 30 dB, SSIM is concentrated around 0.85, and LPIPS is mostly around 0.17. The spread is moderate, indicating that the method behaves similarly across most samples.

For phase retrieval, the histograms show much larger variation. Many samples have PSNR above 30 dB, SSIM close to 0.9, and low LPIPS, but there is a clear group of reconstructions with low PSNR, low SSIM, and high LPIPS. This is likely due to the highly ambiguous nature of phase retrieval.

For CT, the PSNR and SSIM histograms are narrow. PSNR is centered around 33.5 dB, and SSIM is centered around 0.915. This indicates stable reconstruction quality across the 100 test images.

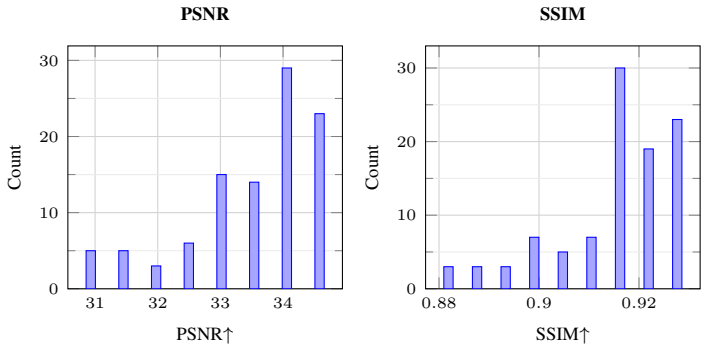


Figure 10: Histogram of reconstruction metrics for sparse-view transmission CT.

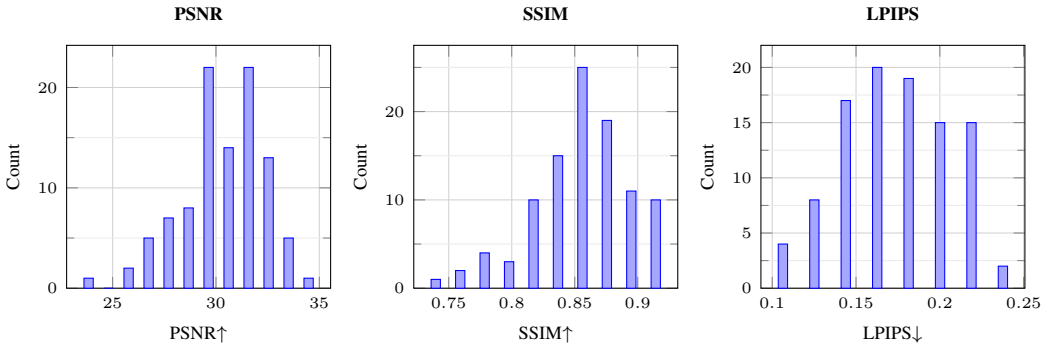


Figure 11: Histogram of reconstruction metrics for Gaussian blur.

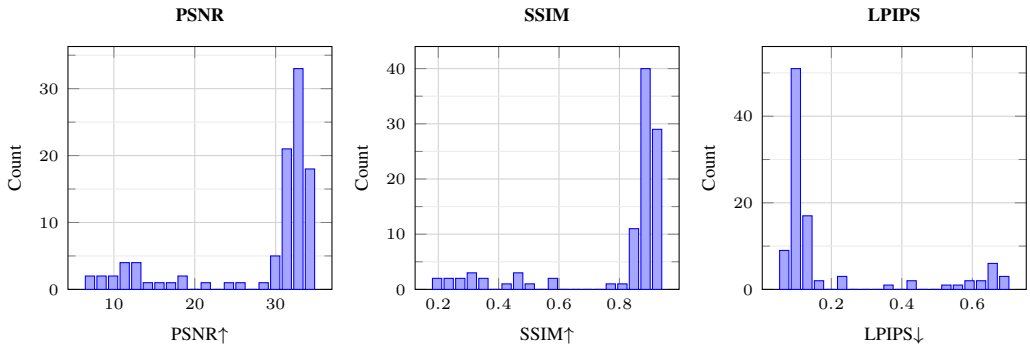


Figure 12: Histogram of reconstruction metrics for phase retrieval.