SIGNAL DENOISING IN NON-GAUSSIAN NOISE VIA DIFFUSION REGULARIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

When the signal prior is known but the measurement-noise distribution is unknown and non-Gaussian, a key challenge is how to incorporate data-driven noise information into a principled denoising objective while explicitly preserving the signal prior. We develop a noise-domain variational framework for this setting. Specifically, a noise-domain optimization problem is rigorously derived from the finite-dispersion variational formulation of regularization by denoising diffusion process (RED-diff), with the learned noise distribution treated as the RED-diff prior and the signal structure retained explicitly. The resulting objective admits a maximum a posteriori (MAP) interpretation under a normalized effective noise likelihood. This framework is instantiated as Regularization by Denoising with an Atomic-Norm Prior (RED–Atomic) for gridless line spectral estimation, where the atomic norm explicitly encodes spectral sparsity. The corresponding optimization problem is solved by primal–dual splitting with a Toeplitz semidefinite representation. Numerical experiments demonstrate competitive or improved performance across a range of non-Gaussian noise models.

1 INTRODUCTION

In many scientific and engineering sensing problems, the quantities of interest are governed by well-specified physical or parametric structure. Representative examples include pulse-Doppler radar (Heckel et al., 2016); spectral estimation (Tang et al., 2013; Chen & Chi, 2013); neural or heat source localization (Michel et al., 2004; Li et al., 2014). In such problems, properties such as sparsity, low rank, nonnegativity, and Toeplitz/Hankel structure constitute prior knowledge about the signal and can therefore be incorporated explicitly through regularization or feasibility constraints (Chandrasekaran et al., 2012). More importantly, the objective in such sensing tasks is often to select a particular structured explanation among candidate solutions consistent with the observations, such as the sparsest representation. The resulting estimates may further serve as direct inputs to downstream tasks such as detection, localization, and tracking. Therefore, these inference tasks are naturally formulated as structured optimization problems.

Despite the availability of explicit structural knowledge, practical sensing systems may still be affected by correlated interference, impulsive disturbances, outliers, heterogeneous backgrounds, or multimodal noise mechanisms, for which the commonly assumed i.i.d. Gaussian model can be substantially misspecified (Wen et al., 2017; Carrillo et al., 2010; Zoubir et al., 2012). Classical model-based estimators can effectively incorporate known signal structure through regularizers or constraints, but their data-fidelity terms are typically built on prespecified and tractable parametric noise models, which leads to substantial measurement-model mismatch. This raises a natural question: *Can a noise likelihood be constructed from representative noise samples while explicitly retaining the known signal structure?*

Diffusion models provide a flexible mechanism for learning complex probability distributions through time-dependent score functions and generating samples through reverse diffusion dynamics (Ho et al., 2020; Song et al., 2020b; Kingma et al., 2021; Dhariwal & Nichol, 2021; Song et al., 2020a). In inverse problems, this capability has predominantly been used to represent the unknown signal distribution and to approximately sample from the posterior under a prescribed measurement model (Chung et al., 2022; Song et al., 2023; Daras et al., 2024; Kawar et al., 2022; Song et al., 2023; Chung et al., 2023). For the structured sensing problems considered here, however, the mod-

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.

eling situation is largely reversed. This motivates using diffusion models to learn the measurement noise distribution rather than the signal prior. Recent work has begun to explore this direction. A joint conditional approximate posterior sampling method is proposed to efficiently remove structured noise by learning separate score models for the signal and noise distributions using diffusion models (Stevens et al., 2025). Legin et al. (2023) propose Score-based Likelihood Characterization (SLIC), which learns the score of a non-Gaussian noise distribution from noise realizations to perform Langevin sampling from the unbiased likelihood. These studies demonstrate the utility of data-driven noise modeling for inverse inference, but an explicit deterministic optimization problem incorporating the learned measurement-noise information remains lacking.

RED-diff (Mardani et al., 2024) provides a perspective toward this connection. It derives a variational formulation of diffusion-based posterior sampling and yields a multiscale score-matching regularizer. However, its stochastic optimization procedure is designed for posterior sampling. The zero-dispersion limit that yields a deterministic estimator, together with its statistical interpretation, has not been formally characterized. We develop this connection in the setting where the signal prior is explicitly specified while the measurement-noise distribution is learned from data. Specifically, we rigorously derive a noise-domain denoising objective from the RED-diff variational framework, retaining the explicit signal prior while using the learned noise model as the data-fidelity term. The resulting deterministic optimization problem admits a MAP interpretation under a normalized effective noise likelihood.

For a concrete instantiation, we consider the denoising problem in gridless line spectral estimation, where the spectral sparsity of the signal is explicitly encoded through the atomic norm. The corresponding noise-domain optimization problem is called Regularization by Denoising with an Atomic-Norm Prior (RED–Atomic). Under Gaussian noise, the proposed formulation reduces to atomic norm denoising (AND) (Bhaskar et al., 2013). We reformulate it through the Toeplitz semidefinite representation of the atomic norm. The corresponding optimization problem is solved using a primal–dual splitting algorithm.

Contributions. Our contributions are twofold. First, we rigorously derive the zero-dispersion limiting noise-domain objective from the finite-dispersion RED-diff formulation and establish its MAP interpretation under a normalized effective noise likelihood, with further characterization under Gaussian noise and diffusion-time weighting. Second, we apply the proposed framework to the continuous line spectral denoising problem, leading to RED–Atomic, and derive a single-loop primal–dual implementation with a Toeplitz semidefinite representation.

2 Related Work

Conceptually closest to our work are SLIC (Legin et al., 2023) and the structured-noise removal framework of Stevens et al. (2025), both of which learn measurement-noise statistics with diffusion models. SLIC estimates the score of a non-Gaussian noise distribution from noise realizations and performs Langevin sampling from the resulting likelihood, whereas Stevens et al. (2025) jointly use separate signal and noise scores for conditional posterior sampling, whereas Stevens et al. (2025) jointly use separate signal and noise scores for conditional posterior sampling. These works demonstrate that diffusion models can capture complex measurement-noise distributions from data and leverage the learned noise statistics for denoising. Their inference procedures, however, remain sampling based. In contrast, we use the learned noise information to define an explicit effective likelihood and combine the resulting data fidelity with a prescribed structural signal prior in a deterministic optimization problem.

Our theoretical development is most directly connected to RED-diff (Mardani et al., 2024), which formulates diffusion-based posterior inference through a finite-dispersion variational distribution and represents the diffusion prior through a multiscale score-matching regularizer. Posterior sampling is then realized by stochastic optimization over the variational parameters. Our work is based on this variational perspective but uses a diffusion model to represent the measurement-noise distribution rather than the signal prior. More importantly, we characterize the zero-dispersion limit of the finite-dispersion formulation, show that the resulting objective corresponds to a proper effective noise likelihood, and use it as the data-fidelity term of a structured estimator.

Our formulation is structurally related to classical model-based signal estimation, where known structure is encoded explicitly through regularizers or constraints. Examples include model-based compressed sensing (Baraniuk et al., 2010; Donoho, 2006) and gridless spectral estimation with atomic norms (Tang et al., 2013; Bhaskar et al., 2013). Robust variants accommodate non-Gaussian measurements by replacing quadratic fidelity with prescribed losses or statistical models, including ℓ_1-type and Huber penalties (Carrillo et al., 2010; Wen et al., 2017; Huber, 1992). These approaches specify the form of the measurement fidelity analytically. Our framework instead retains the explicit structural regularizer while estimating the fidelity from noise-only data through a diffusion model.

3 PRELIMINARIES

In this section, we first introduce diffusion models and score estimation in Section 3.1, and then review the basic RED-diff framework in Section 3.2.

3.1 DIFFUSION MODELS AND SCORE ESTIMATION

Diffusion models are widely used generative models allowing sampling from data distributions by progressively denoising samples drawn from a simple reference distribution, typically Gaussian (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2020b;a). Let $z_0 \sim p_0$ denote a data sample. Under the variance-preserving (VP) diffusion, the forward process is a prescribed noising process which can be described by a stochastic differential equation (SDE) (Song et al., 2020b)

$$dz_t = -\frac{1}{2}\beta(t) z_t \, dt + \sqrt{\beta(t)}\, dw_t, \qquad t \in [0, T], \tag{1}$$

where $\beta(t) > 0$ is the noise schedule and w_t is a standard Wiener process. Its transition kernel has the closed-form representation $z_t = \alpha_t z_0 + \sigma_t \epsilon$, $\epsilon \sim \mathcal{N}(0, I)$, where $\alpha_t = \exp\Big(-\frac{1}{2}\int_0^t \beta(s)\, ds\Big)$, $\sigma_t^2 = 1 - \alpha_t^2$. The signal-to-noise ratio (SNR) at diffusion time t is defined as $\mathrm{SNR}_t = \alpha_t^2/\sigma_t^2$. Generation is performed by reversing this process. The corresponding reverse-time SDE is

$$dz_t = \left[-\frac{1}{2}\beta(t) z_t - \beta(t)\nabla_{z_t} \log p_t(z_t)\right] dt + \sqrt{\beta(t)}\, d\bar{w}_t, \tag{2}$$

integrated from T to 0. Hence, the reverse dynamics are determined by the time-dependent score $s^\star(z_t, t) = \nabla_{z_t} \log p_t(z_t)$, which can be estimated through denoising score matching. The relationship between the score function and the diffusion noise at diffusion time t can be expressed as

$$-\sigma_t \nabla_{z_t} \log p_t(z_t) = \mathbb{E}[\epsilon \mid z_t, t] = \arg\min_a \mathbb{E}\left[\|\epsilon - a\|_2^2 \mid z_t, t\right]. \tag{3}$$

A time-conditioned denoiser $\epsilon_\theta(z_t, t)$ is therefore trained using the noise-prediction objective

$$\min_\theta \; \mathbb{E}_{z_0, t, \epsilon}\left[w_{\mathrm{tr}}(t)\, \|\epsilon - \epsilon_\theta(\alpha_t z_0 + \sigma_t \epsilon, t)\|_2^2\right], \tag{4}$$

and provides the score estimate $s_\theta(z_t, t) = -\epsilon_\theta(z_t, t)/\sigma_t \approx \nabla_{z_t} \log p_t(z_t)$.

3.2 RED-DIFF VARIATIONAL REGULARIZATION

RED-diff (Mardani et al., 2024) provides a variational framework of diffusion-based sampling for inverse problems. Consider the observation model

$$y = f(x_0) + v, \qquad v \sim \mathcal{N}(0, \sigma_v^2 I), \tag{5}$$

where f is a known forward operator and a pretrained diffusion model represents the prior distribution $p(x_0)$. RED-diff approximates the posterior $p(x_0 \mid y)$ with a Gaussian distribution $q_\delta^\mu(x_0 \mid y) = \mathcal{N}\left(\mu, \delta^2 I\right)$ and minimizes the Kullback–Leibler (KL) divergence

$$\min_{\mu, \delta} \; D_{\mathrm{KL}}\left(q_\delta^\mu(x_0 \mid y) \,\|\, p(x_0 \mid y)\right). \tag{6}$$

Using Bayes' rule, the objective can be separated as

$$D_{\mathrm{KL}}\left(q_\delta^\mu(x_0 \mid y) \,\|\, p(x_0 \mid y)\right) = -\mathbb{E}_{q_\delta^\mu}\left[\log p(y \mid x_0)\right] + D_{\mathrm{KL}}\left(q_\delta^\mu(x_0 \mid y) \,\|\, p(x_0)\right) + C_y, \tag{7}$$

where C_y does not depend on the variational parameters. The first term enforces consistency with the measurement model, whereas the second term favors solutions that are consistent with the learned diffusion prior. By formally setting the zero-dispersion specialization $\delta = 0$, the resulting optimization problem with respect to the variational mean μ takes the generic form

$$\min_{\mu} \ \|y - f(\mu)\|_2^2 + \mathbb{E}_{t,\epsilon}\left[\lambda_t \left\|\epsilon_\theta\left(\alpha_t \mu + \sigma_t \epsilon, t\right) - \epsilon\right\|_2^2\right] \tag{8}$$

where λ_t is a time-dependent weight. Thus, diffusion information enters the optimization through a multi-time score-matching based regularizer. RED-diff interprets posterior sampling as stochastic optimization over the variational mean μ. This variational formulation avoids direct approximation of the conditional likelihood $p(y \mid x_t)$, which is generally intractable. The authors of RED-diff further propose an inverse-SNR-related weighting strategy to balance contributions from different diffusion levels.

4 Noise-Domain Variational Framework

Consider the following measurement model

$$y = x + n, \tag{9}$$

where $x \in \mathbb{R}^d$ represents the signal and n denotes measurement noise. Assume that the signal is governed by a known physical or parametric model that induces an explicit structural prior $p_X(x) = Z_X^{-1} \exp\{-\mathcal{R}_X(x)\}$, whereas the noise distribution p_N is unknown and is learned from representative noise samples using a diffusion model. The aim is to remove the measurement noise while preserving the known signal structure through a deterministic optimization objective. We first revisit RED-diff through a generalized multiscale KL formulation, which further clarifies how diffusion-time weighting determines the corresponding optimization objective and provides a principled criterion for selecting the weights. Then we reformulate the denoising problem in the noise domain.

4.1 MultiScale KL Regularization and Time Weighting

We introduce a multiscale KL formulation to characterize the role of diffusion-time weighting. Let q_0 and p_0 be two distributions propagated by the same VP diffusion process, and denote their marginals at diffusion time t by q_t and p_t, respectively. Define $D_t = D_{\mathrm{KL}}(q_t \| p_t)$, $I_t = \mathbb{E}_{q_t}\left[\|\nabla \log q_t - \nabla \log p_t\|_2^2\right]$. The relationship between D_t and I_t can be given as $\frac{d}{dt} D_t = -\frac{\beta(t)}{2} I_t$ under VP diffusion (Mardani et al., 2024; Song et al., 2021). Let $\pi(t)$ be a probability density over the diffusion time interval $[0, T]$ with $\pi(t) \geq 0$, $\int_0^T \pi(t)\, dt = 1$. Let $\omega(t) = \int_0^t \pi(r)\, dr$ be its cumulative weight. The corresponding multiscale KL divergence is defined as

$$\mathcal{D}_\pi(q, p) := \int_0^T D_{\mathrm{KL}}(q_t \| p_t)\, \pi(t)\, dt. \tag{10}$$

By setting $q_T = p_T$, Eq. 10 can also be equivalently represented in terms of I_t,

$$\mathcal{D}_\pi(q, p) = \int_0^T \frac{\beta(t)}{2} \omega(t) I_t\, dt. \tag{11}$$

Hence, a time-weighted score-matching regularizer with $\widetilde{\omega}(t) = \beta(t)\omega(t)/2$ admits an equivalent multiscale KL interpretation. This representation is consistent with the generalized time weighting used in RED-diff (Mardani et al., 2024). Although such reweighting generally changes the original MAP target under the physical noise likelihood, it offers a useful property for robust estimation.

Diffusion-time weighting. For the weighting analysis, we fix $\delta > 0$. Define $D_{t,\delta}(\mu) = D_{\mathrm{KL}}(q_{t,\delta}^{\mu} \| p_t)$, $D_{\pi,\delta}(\mu) = \int_0^T \pi(t) D_{t,\delta}(\mu)\, dt$, $\mathcal{E}_{\epsilon,\delta}(t;\mu) = \mathbb{E}_{z\sim q_{t,\delta}^{\mu}}\left[\|\epsilon_\theta(z,t) - \epsilon_t^\star(z)\|_2^2\right]$, where $q_{t,\delta}^{\mu} = \mathcal{N}\left(\alpha_t \mu,\, (\alpha_t^2 \delta^2 + \sigma_t^2) I\right)$, $\epsilon_\theta(z,t)$ is a pre-trained denoiser. The learned multiscale gradient satisfies

$$\left\|\widehat{\nabla D}_{\pi,\delta}(\mu) - \nabla D_{\pi,\delta}(\mu)\right\|_2^2 \leq C_{\pi,\delta}(\mu) := \int_0^T \pi(t) C_\delta(t;\mu)\, dt, \tag{12}$$

where $C_\delta(t;\mu) = \mathrm{SNR}_t \mathcal{E}_{\epsilon,\delta}$. Since $\mathrm{SNR}_t \to \infty$ as $t \to 0$, Eq. 12 may become unbounded unless the score approximation error decreases sufficiently fast. Thus, the original KL scale $D_{0,\delta}(\mu)$ is statistically most faithful but may be highly sensitive to learned-score errors.

Equation 12 provides a direct robustness interpretation of the inverse-SNR-related weighting used in RED-diff (Mardani et al., 2024). In particular, consider the following normalized probability density $\pi_{\mathrm{IS}}(t) = \mathrm{SNR}_t^{-1} / \int_0^T \mathrm{SNR}_s^{-1}\, ds$. Substituting $\pi_{\mathrm{IS}}(t)$ into Eq. 12 yields $C_{\pi_{\mathrm{IS}},\delta}(\mu) = \int_0^T \mathcal{E}_{\epsilon,\delta}(t;\mu)\, dt / \int_0^T \mathrm{SNR}_s^{-1}\, ds$. Thus, inverse-SNR weighting suppresses the potentially unstable high-SNR regime without requiring explicit knowledge of the population score error. We use this weighting as the default RED-diff choice in the subsequent formulation and experiments.

4.2 Noise-Domain Variational Formulation

We now reformulate the denoising problem in the noise domain. Since $n = y - x$, the posterior can be equivalently written as

$$p(n \mid y) \propto p_N(n)\, p_X(y - n), \tag{13}$$

so that the unknown noise distribution p_N plays the role of the prior in RED-diff, while the known signal structure remains explicitly encoded by $p_X(y - n)$. We retain the notation in Section 3.2, where μ now denotes the variational mean of the noise variable n and δ is the variational dispersion.

Following RED-diff, we approximate Eq. 13 by $q^{\mu}_{0,\delta}$, whose diffused marginal is $q^{\mu}_{t,\delta} = \mathcal{N}\left(\alpha_t \mu,\, v^2_{t,\delta} I\right)$, $v^2_{t,\delta} = \alpha_t^2 \delta^2 + \sigma_t^2$. Therefore, for each fixed $\delta > 0$, the resulting optimization problem for μ under the RED-diff framework is

$$\min_{\mu}\ \mathbb{E}_\eta \left[\mathcal{R}_X(y - \mu - \delta\eta)\right] + \int_0^T D_{\mathrm{KL}}\left(q^{\mu}_{t,\delta} \,\|\, p_{N,t}\right) \pi(t)\, dt, \tag{14}$$

where $\eta \sim \mathcal{N}(0, I)$.

Define

$$U_{t,\delta}(\mu) = -\mathbb{E}_\epsilon \log\left[\alpha_t^d p_{N,t}\left(\alpha_t \mu + \sqrt{\alpha_t^2 \delta^2 + \sigma_t^2}\,\epsilon\right)\right],\ \ U_t(\mu) = -\mathbb{E}_\epsilon \log\left[\alpha_t^d p_{N,t}\left(\alpha_t \mu + \sigma_t \epsilon\right)\right], \tag{15}$$

and $U_{\pi,\delta}(\mu) = \int_0^T U_{t,\delta}(\mu)\pi(t)\, dt$, $U_\pi(\mu) = \int_0^T U_t(\mu)\pi(t)\, dt$. The following theorems characterize the zero-dispersion limit of Eq. 14 and prove that the resulting limiting optimization problem admits a MAP interpretation.

Theorem 1. *Let $x = y - \mu$. Suppose that $\mathcal{R}_X$ is continuous with at most polynomial growth and that, for each fixed μ, the family $\{U_{t,\delta}(\mu)\}_{0<\delta\le\delta_0}$ is dominated by a π-integrable function of t for some $\delta_0 > 0$. Then, as $\delta \to 0$, after removing terms independent of μ, the objective in Eq. 14 converges pointwise to*

$$U_\pi(y - x) + \mathcal{R}_X(x). \tag{16}$$

Theorem 2. *Under the assumption of Theorem 1, $p^{\mathrm{eff}}_\pi(n) = Z_\pi^{-1} e^{-U_\pi(n)}$ defines a proper probability density with $Z_\pi = \int_{\mathbb{R}^d} e^{-U_\pi(n)}\, dn$. The corresponding optimization problem*

$$\widehat{x}_\pi \in \arg\min_x \left\{U_\pi(y - x) + \mathcal{R}_X(x)\right\}, \tag{17}$$

is a MAP estimator under the effective noise likelihood $p^{\mathrm{eff}}_\pi(y - x)$ and the original signal prior $p_X(x)$.

Corollary 1. *Consider the general additive-noise inverse problem $y = \mathcal{A}(x) + n$, where $\mathcal{A}$ is a known forward operator. The MAP estimator under p^{eff}_π and the original signal prior p_X is given by*

$$\widehat{x}_\pi \in \arg\min_x \left\{U_\pi(y - \mathcal{A}(x)) + R_X(x)\right\}. \tag{18}$$

In particular, the denoising model considered in Theorem 1 is recovered by setting $\mathcal{A} = I$.

Proof. Since the measurement noise is additive, the effective likelihood satisfies $p^{\mathrm{eff}}_\pi(y \mid x) = p^{\mathrm{eff}}_\pi(y - \mathcal{A}(x))$. By Bayes' rule, the posterior is proportional to $p^{\mathrm{eff}}_\pi(y - \mathcal{A}(x)) p_X(x)$. Taking the negative logarithm and discarding terms independent of x yields the stated MAP objective. □

Theorem 3. *Suppose that $p_N = \mathcal{N}(0, \Sigma)$, $\Sigma \succ 0$ and define $r_t = \frac{\sigma_t^2}{\alpha_t^2}$. Then*

$$U_\pi(n) = \frac{1}{2} n^\mathsf{T} W_\pi n + c_\pi, \qquad W_\pi = \int_0^T \pi(t) \left(\Sigma + r_t I\right)^{-1} dt, \tag{19}$$

where c_π is independent of n. Hence $p_\pi^{\text{eff}} = \mathcal{N}\left(0,\, W_\pi^{-1}\right)$. Moreover, $1 \le \operatorname{cond}(W_\pi) \le \operatorname{cond}(\Sigma)$.

Theorem 3 shows that positive diffusion scales can improve the conditioning of the Gaussian data-fidelity term. In particular, if $\Sigma = \sigma_N^2 I$, then W_π is a scalar multiple of the identity. As the weighting concentrates at the zero-diffusion limit, $W_\pi \to \Sigma^{-1}$ and the physical Gaussian likelihood is recovered.

The proofs of Theorems 1–3 are given in Appendix A.1.

5 RED–Atomic for Gridless Line Spectral Estimation

In this section, we apply the proposed framework to gridless line spectral denoising task, where the signal is spectrally sparse with frequencies lying in a continuous domain. The resulting method is termed Regularization by Denoising with an Atomic-Norm Prior (RED–Atomic). Although the theoretical development is presented for real-valued variables, the complex-valued RED–Atomic formulation follows directly by identifying $\mathbb{C}^M$ with $\mathbb{R}^{2M}$.

5.1 Problem Formulation

A spectrally sparse signal with K components can be written as (Candès & Fernandez-Granda, 2014; Tang et al., 2013)

$$x = \sum_{k=1}^{K} c_k a(f_k), \qquad a(f) = \left[1, e^{\mathrm{i}2\pi f}, \ldots, e^{\mathrm{i}2\pi(M-1)f}\right]^\mathsf{T}, \tag{20}$$

where $f_k \in [0, 1)$ are continuous frequencies and $c_k \in \mathbb{C}$ are complex amplitudes. Neither the frequencies nor the number of components is assumed known in advance. Define the atomic set $\mathcal{A} = \left\{e^{\mathrm{i}\phi} a(f) : f \in [0,1),\ \phi \in [0, 2\pi)\right\}$ and the corresponding atomic norm (Tang et al., 2013; Yang & Xie, 2015; 2016)

$$\|x\|_\mathcal{A} := \inf \left\{ \sum_j d_j : x = \sum_j d_j b_j,\ d_j \ge 0,\ b_j \in \mathcal{A} \right\}. \tag{21}$$

The atomic norm provides a convex surrogate for spectral sparsity over the continuous frequency domain. Setting $p_X(x) = Z_X^{-1} \exp\{-\lambda_A \|x\|_\mathcal{A}\}$ gives RED–Atomic:

$$\widehat{x} \in \arg\min_{x \in \mathbb{C}^M} \left\{ U_\pi(y - x) + \lambda_A \|x\|_\mathcal{A} \right\}. \tag{22}$$

For Gaussian noise, Eq. 22 is equivalent to conventional atomic norm denoising (Bhaskar et al., 2013) up to a rescaling of the regularization parameter.

The atomic norm Eq. 21 admits the Toeplitz semidefinite representation (Candès & Fernandez-Granda, 2014; Tang et al., 2013)

$$\|x\|_\mathcal{A} = \min_{u,\rho} \frac{u_0 + \rho}{2} \quad \text{s.t.} \quad \mathcal{Z}(x, u, \rho) := \begin{bmatrix} \mathcal{T}(u) & x \\ x^\mathsf{H} & \rho \end{bmatrix} \succeq 0, \tag{23}$$

where $\mathcal{T}(u)$ is the Hermitian Toeplitz matrix, $u = [u_0, u_1, \ldots, u_{M-1}]^\mathsf{T}$. Therefore, Eq. 22 is equivalent to

$$\begin{aligned} \min_{x,u,\rho} \quad & U_\pi(y - x) + \frac{\lambda_A}{2}(u_0 + \rho) \\ \text{s.t.} \quad & \mathcal{Z}(x, u, \rho) \succeq 0. \end{aligned} \tag{24}$$

The Toeplitz representation isolates the continuous spectral structure in a single positive-semidefinite constraint, while the learned noise model enters only through the smooth gradient of U_π.

Algorithm 1 RED–Atomic via primal–dual splitting

Require: Observation y, denoiser ϵ_θ, $\{t_j, w_j\}_{j=1}^J$, $\{\epsilon_b\}_{b=1}^B$, λ_A, step sizes η, ζ
1: Initialize $z^0 = (x^0, u^0, \rho^0)$ and Y^0
2: **for** $k = 0, 1, \ldots$ **do**
3: $n^k \leftarrow y - x^k$
4: Evaluate $\widehat{g}_{\pi,B}(n^k)$ using 30
5: $\widehat{\nabla f}(z^k) \leftarrow (-\widehat{g}_{\pi,B}(n^k), 0, 0)$
6: Update z^{k+1} using 28
7: Update Y^{k+1} using 29
8: **if** stationarity and feasibility tolerances are satisfied **then**
9: **break**
10: **end if**
11: **end for**
12: Recover $\widehat{x}$ and extract continuous frequencies from the estimated spectral structure

5.2 Primal–Dual Splitting Algorithm

Primal–dual splitting methods solve constrained optimization problems by introducing a dual variable and recasting the problem as a saddle-point problem (Condat, 2013; Vũ, 2013). In particular, a problem of the form

$$\min_z \; f(z) + \psi(z) + \chi(\mathcal{K}z) \tag{25}$$

can be equivalently written as

$$\min_z \max_Y \; f(z) + \psi(z) + \langle \mathcal{K}z, Y\rangle - \chi^*(Y), \tag{26}$$

where χ^* is the Fenchel conjugate of χ and Y is the corresponding dual variable. We apply this framework to Eq. 24. Let $z = (x, u, \rho)$, $\mathcal{K}z = \mathcal{Z}(x, u, \rho)$ and define $f(z) = U_\pi(y - x)$, $\psi(z) = \frac{\lambda_A}{2}(u_0 + \rho)$, $\chi = \iota_{\mathbb{H}_+^{M+1}}$, where $\iota_{\mathbb{H}_+^{M+1}}$ is the indicator function of the Hermitian positive-semidefinite cone. Since $\chi^* = \iota_{\mathbb{H}_-^{M+1}}$, the dual variable is constrained to the negative-semidefinite cone.

Let $\mathcal{K}^*$ denote the adjoint of $\mathcal{K}$. Using the learned diffusion gradient,

$$\widehat{\nabla f}(z) = \left(-\widehat{\nabla U_\pi}(y - x),\, 0,\, 0\right), \tag{27}$$

the primal–dual iterations are

$$z^{k+1} = \mathrm{prox}_{\eta\psi}\left(z^k - \eta\left[\widehat{\nabla f}(z^k) + \mathcal{K}^* Y^k\right]\right), \tag{28}$$

$$Y^{k+1} = \Pi_{\mathbb{H}_-^{M+1}}\left[Y^k + \zeta\,\mathcal{K}\left(2z^{k+1} - z^k\right)\right], \tag{29}$$

where $\eta, \zeta > 0$ are the primal and dual step sizes, respectively. Each iteration requires one diffusion-gradient evaluation and one semidefinite cone projection. The algorithm workflow is shown in Appendix B.4.

For numerical implementation, $\nabla U_\pi(n)$ is approximated by discretizing the diffusion-time expectation over $\{t_j, w_j\}_{j=1}^J$ and using B Gaussian samples at each level

$$\widehat{\nabla U_\pi}(n) = \sum_{j=1}^{J} w_j \frac{\alpha_{t_j}}{B\sigma_{t_j}} \sum_{b=1}^{B}\left[\epsilon_\theta\left(\alpha_{t_j} n + \sigma_{t_j}\epsilon_b, t_j\right) - \epsilon_b\right], \qquad \epsilon_b \sim \mathcal{C}N(0, I). \tag{30}$$

6 Experiments

6.1 Different noise distributions

Setup. We evaluate gridless line spectral recovery from $y = x + \sigma z$ with $M = 32$ samples and $K = 3$ spectral components. Frequencies are sampled uniformly on $[0, 1)$ with circular separation at least $4/M$; complex amplitudes have magnitudes in $[0.8, 1.2]$ and uniformly random phases.

Each clean signal is normalized to $\|x\|_2^2/M = 1$. We use 50 paired Monte Carlo (MC) observations per condition and report frequency root-mean-square error (RMSE). For the VP diffusion, we parameterize the numerical smoothing level by $\tau(t) = \sigma_t/\alpha_t$.

Noise models. The correlated-Gaussian benchmark uses $z \sim \mathcal{CN}(0, R_{50}(0.23))$, where $R_\kappa(f) = M[I+(\kappa-1)v(f)v(f)^H]/(M+\kappa-1)$. The non-Gaussian colored benchmark is a moment-matched covariance-switching mixture

$$z \sim \tfrac{1}{2}\mathcal{CN}(0, C_1) + \tfrac{1}{2}\mathcal{CN}(0, C_2), \qquad C_{1,2} = I \pm 0.9(P_1 - P_2), \tag{31}$$

where P_1 and P_2 are fixed orthogonal rank-10 random projectors. Hence $(C_1 + C_2)/2 = I$: the two mixture components are strongly colored, but their unconditional covariance is exactly isotropic. The sparse impulsive benchmark uses $z_i = (g_i + 10 b_i h_i)/\sqrt{6}$, with $g_i, h_i \sim \mathcal{CN}(0, 1)$ and $b_i \sim \text{Bernoulli}(0.05)$. We additionally evaluate five other noise distributions, with the experimental settings and results reported in Appendices B.1 and B.3.

Baselines. We compare AND (Bhaskar et al., 2013), *covariance-based AND* (Cov-AND), *Huber-loss-based AND* (Huber-AND), *normalizing-flow-based AND* (NF-AND), and RED-Atomic. Huber-AND combines the Huber loss (Huber, 1992) with the same atomic penalty, while NF-AND uses a Real NVP likelihood (Dinh et al., 2016). Cov-AND uses a covariance-weighted quadratic fidelity. All methods use the same continuous Toeplitz–PSD atomic representation.

For the color-switching model, RED is trained on 200k noise-only samples with 10k validation samples. Its score network combines a global covariance skip, 64 learned quadratic features, and a 256–256 MLP; the learned features do not use mixture labels or the true projectors. RED uses eight geometric smoothing scales in $[0.03, 0.6]$, weights proportional to τ^2, four antithetic perturbations. NF-AND uses a separately trained four-layer Real NVP. The Bernoulli–Gaussian RED model uses 100k noise-only samples. The atomic penalty coefficient is $0.6\sigma\sqrt{M \log M}$ and the Huber threshold multiplier is 1.5.

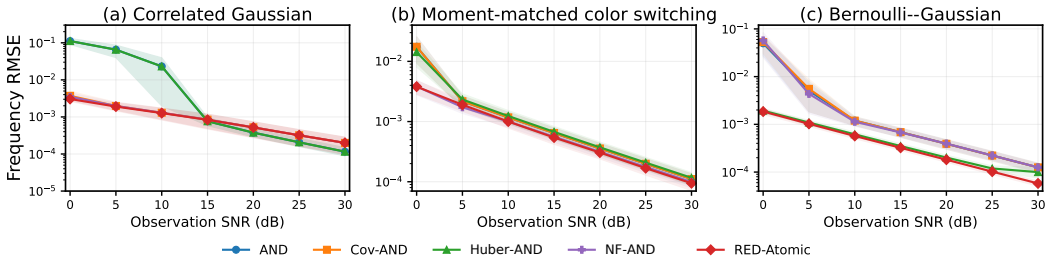


Figure 1: Frequency RMSE under three noise distributions. All panels span 0–30 dB and include AND, Cov-AND, Huber-AND, NF-AND, and RED-Atomic. Panel (b) uses the moment-matched rank-10 covariance-switching mixture in Eq. 31; its 5/10-dB 50-trial values are reported separately in Table 1.

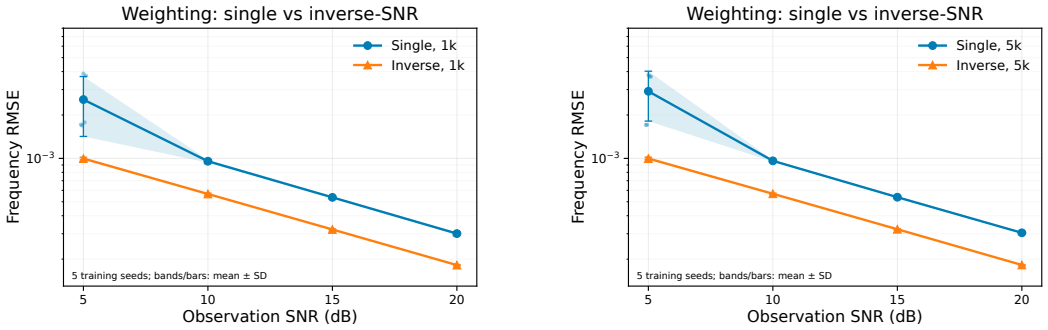
Results. Under correlated Gaussian noise, Cov-AND, NF-AND, and RED-Atomic nearly coincide: at 5 dB their RMSEs are 1.978, 1.975, and 1.903×10^{-3}. This is the matched-Gaussian regime, where second-order information already captures the noise law.

The color-switching mixture isolates information beyond the unconditional covariance. Cov-AND reduces to the same quadratic model as AND and cannot identify the active covariance state because $\text{Cov}(z) = I$. Across the full SNR sweep in Figure 1(b), RED-Atomic and NF-AND remain below the quadratic and Huber baselines. In the 50-trial 5/10-dB benchmark, RED-Atomic improves over Cov-AND by 19.8% and 18.0%, respectively, while NF-AND attains nearly the same accuracy. This agreement is expected when the explicit likelihood model is expressive enough to capture the smooth mixture density.

The Bernoulli–Gaussian experiment produces a different ordering. RED-Atomic obtains $1.025/0.571 \times 10^{-3}$ at 5/10 dB, compared with $4.355/1.134 \times 10^{-3}$ for NF-AND and $1.090/0.618 \times 10^{-3}$ for Huber-AND. The comparison highlights a practical distinction between the two learned

Table 1: Frequency RMSE ($\times 10^{-3}$) at 5/10 dB, 50 paired trials per setting.

Method	Correlated Gaussian 5 dB	Correlated Gaussian 10 dB	Color switch 5 dB	Color switch 10 dB	Bernoulli–Gaussian 5 dB	Bernoulli–Gaussian 10 dB
AND	65.343	23.130	2.334	1.287	5.045	1.197
Cov-AND	1.978	1.298	2.334	1.287	5.541	1.197
Huber-AND	65.206	23.080	2.331	1.288	1.090	0.618
NF-AND	1.975	1.297	**1.856**	**1.051**	4.355	1.134
RED-Atomic	**1.903**	**1.270**	1.872	1.056	**1.025**	**0.571**



(a) $N_{\text{train}} = 1000$ (b) $N_{\text{train}} = 5000$

Figure 2: Single-scale versus inverse-SNR weighting under Bernoulli–Gaussian noise. The multi-scale objective uses eight geometrically spaced smoothing levels over $\tau \in [0.01, 0.6]$. Panels (a) and (b) correspond to $N_{\text{train}} = 1000$ and $N_{\text{train}} = 5000$, respectively. Curves report the mean over five independently trained score models, with error bars indicating ± 1 standard deviation across training seeds.

models: NF-AND must learn an explicit normalized density $p_\theta(z)$, whereas RED-Atomic directly learns the smoothed score $\nabla_z \log p_\tau(z)$ required by the reconstruction. This difference is most visible for sparse impulsive noise, where accurately modeling rare tail events is more demanding than learning the local score field after smoothing.

6.2 Time Weighting

Inverse-SNR weighting. We compare the single-scale objective at $t = 0.002$ ($\tau(t) \approx 0.0155$) with inverse-SNR multiscale weighting over $\tau \in [0.01, 0.6]$ under the same Bernoulli–Gaussian noise model above. Figure 2(a) shows the frequency RMSE results at $N_{\text{train}} = 1000$, which are averaged over five independently trained score models and 50 test observations per model. It can be seen from Figure 2(a) that inverse-SNR weighting reduces the frequency RMSE from 2.559 to 0.997×10^{-3} at 5 dB, from 0.956 to 0.566×10^{-3} at 10 dB, from 0.536 to 0.321×10^{-3} at 15 dB, and from 0.301 to 0.182×10^{-3} at 20 dB. The same trend holds at $N_{\text{train}} = 5000$. Thus, aggregating across scales is consistently more effective than relying on a single near-zero scale.

7 Conclusion

We proposed a noise-domain variational framework based on RED-diff for signal denoising problem in which the signal prior is explicitly known but the measurement-noise distribution is unknown. Starting from the finite-dispersion RED-diff formulation, we rigorously derived the corresponding zero-dispersion limiting objective. The resulting minimizer admits a MAP interpretation under a normalized effective noise likelihood. We applied the framework to gridless line spectral denoising problem, leading to RED–Atomic, which combines a learned noise model with an explicit sparsity prior. We developed a primal–dual splitting algorithm based on the semidefinite representation of the atomic norm for solving it. Numerical experiments under different noise distributions demonstrate the effectiveness of the proposed method.

AI USE STATEMENT

ChatGPT (GPT 5.6) were used to search for and identify relevant papers, polish the language of the manuscript, check the manuscript for errors, assist in checking mathematical derivations, assist in designing the noise distributions and experimental settings, and help implement the experimental code. The research questions, methodological choices, theoretical claims, experimental settings, and final manuscript were determined, reviewed, and validated by the authors. All AI-assisted derivations, code, and experimental results were independently verified, and the authors take full responsibility for the correctness and final content of the paper.

REFERENCES

Richard G Baraniuk, Volkan Cevher, Marco F Duarte, and Chinmay Hegde. Model-based compressive sensing. *IEEE Transactions on information theory*, 56(4):1982–2001, 2010.

Badri Narayan Bhaskar, Gongguo Tang, and Benjamin Recht. Atomic norm denoising with applications to line spectral estimation. *IEEE Transactions on Signal Processing*, 61(23):5987–5999, 2013.

Emmanuel J Candès and Carlos Fernandez-Granda. Towards a mathematical theory of super-resolution. *Communications on pure and applied Mathematics*, 67(6):906–956, 2014.

Rafael E Carrillo, Kenneth E Barner, and Tuncer C Aysal. Robust sampling and reconstruction methods for sparse signals in the presence of impulsive noise. *IEEE Journal of Selected Topics in Signal Processing*, 4(2):392–408, 2010.

Venkat Chandrasekaran, Benjamin Recht, Pablo A Parrilo, and Alan S Willsky. The convex geometry of linear inverse problems. *Foundations of Computational mathematics*, 12(6):805–849, 2012.

Yuxin Chen and Yuejie Chi. Spectral compressed sensing via structured matrix completion. In *International conference on machine learning*, pp. 414–422. PMLR, 2013.

Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687*, 2022.

Hyungjin Chung, Jeongsol Kim, and Jong Chul Ye. Direct diffusion bridge using data consistency for inverse problems. *Advances in Neural Information Processing Systems*, 36:7158–7169, 2023.

Laurent Condat. A primal–dual splitting method for convex optimization involving lipschitzian, proximable and linear composite terms. *Journal of optimization theory and applications*, 158(2): 460–479, 2013.

Giannis Daras, Hyungjin Chung, Chieh-Hsin Lai, Yuki Mitsufuji, Jong Chul Ye, Peyman Milanfar, Alexandros G Dimakis, and Mauricio Delbracio. A survey on diffusion models for inverse problems. *arXiv preprint arXiv:2410.00083*, 2024.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

David L Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006.

Reinhard Heckel, Veniamin I Morgenshtern, and Mahdi Soltanolkotabi. Super-resolution radar. *Information and Inference: A Journal of the IMA*, 5(1):22–75, 2016.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution*, pp. 492–518. Springer, 1992.

Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. *Advances in neural information processing systems*, 35:23593–23606, 2022.

Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *Advances in neural information processing systems*, 34:21696–21707, 2021.

Ronan Legin, Alexandre Adam, Yashar Hezaveh, and Laurence Perreault-Levasseur. Beyond gaussian noise: A generalized approach to likelihood analysis with non-gaussian noise. *The Astrophysical Journal Letters*, 949(2):L41, 2023.

Yingying Li, Stanley Osher, and Richard Tsai. Heat source identification based on constrained minimization. *Inverse Problems and Imaging*, 8(1):199–221, 2014.

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*, volume 2024, pp. 28027–28053, 2024.

Christoph M Michel, Micah M Murray, Göran Lantz, Sara Gonzalez, Laurent Spinelli, and Rolando Grave De Peralta. Eeg source imaging. *Clinical neurophysiology*, 115(10):2195–2222, 2004.

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

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. pmlr, 2015.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020a.

Jiaming Song, Arash Vahdat, Morteza Mardani, and Jan Kautz. Pseudoinverse-guided diffusion models for inverse problems. In *International conference on learning representations*, 2023.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020b.

Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. *Advances in neural information processing systems*, 34:1415–1428, 2021.

Tristan Stevens, Hans Van Gorp, Faik C Meral, Junseob Shin, Jason Yu, Jean-luc Robert, and Ruud Van Sloun. Removing structured noise using diffusion models. *Transactions on Machine Learning Research*, 2025.

Gongguo Tang, Badri Narayan Bhaskar, Parikshit Shah, and Benjamin Recht. Compressed sensing off the grid. *IEEE transactions on information theory*, 59(11):7465–7490, 2013.

Bng Công Vũ. A splitting algorithm for dual monotone inclusions involving cocoercive operators. *Advances in Computational Mathematics*, 38(3):667–681, 2013.

Fei Wen, Ling Pei, Yuan Yang, Wenxian Yu, and Peilin Liu. Efficient and robust recovery of sparse signal and image using generalized nonconvex regularization. *IEEE Transactions on Computational Imaging*, 3(4):566–579, 2017.

Zai Yang and Lihua Xie. On gridless sparse methods for line spectral estimation from complete and incomplete data. *IEEE Transactions on Signal Processing*, 63(12):3139–3153, 2015.

Zai Yang and Lihua Xie. Exact joint sparse frequency recovery via optimization methods. *IEEE Transactions on Signal Processing*, 64(19):5145–5157, 2016.

Abdelhak M Zoubir, Visa Koivunen, Yacine Chakhchoukh, and Michael Muma. Robust estimation in signal processing: A tutorial-style treatment of fundamental concepts. *IEEE Signal Processing Magazine*, 29(4):61–80, 2012.

A PROOFS

A.1 PROOF OF THEOREM 1

We first establish the zero-dispersion limit of 14. For each fixed $t > 0$, $q^{\mu}_{t,\delta} = \mathcal{N}(\alpha_t \mu, v^2_{t,\delta} I)$ with $v^2_{t,\delta} = \alpha_t^2 \delta^2 + \sigma_t^2$. By the definition of KL divergence,

$$D_{\mathrm{KL}}\left(q^{\mu}_{t,\delta} \,\|\, p_{N,t}\right) = \mathbb{E}_{q^{\mu}_{t,\delta}}\left[\log q^{\mu}_{t,\delta}(z)\right] - \mathbb{E}_{q^{\mu}_{t,\delta}}\left[\log p_{N,t}(z)\right]. \tag{A.1}$$

Since $q^{\mu}_{t,\delta}$ is Gaussian,

$$\mathbb{E}_{q^{\mu}_{t,\delta}}\left[\log q^{\mu}_{t,\delta}(z)\right] = -\frac{d}{2}\log(2\pi e\, v^2_{t,\delta}).$$

Using $z = \alpha_t \mu + \sqrt{\alpha_t^2 \delta^2 + \sigma_t^2}\,\epsilon$, $\epsilon \sim \mathcal{N}(0, I)$, and the definition of $U_{t,\delta}$, we therefore obtain the exact decomposition

$$D_{\mathrm{KL}}\left(q^{\mu}_{t,\delta} \,\|\, p_{N,t}\right) = U_{t,\delta}(\mu) + C_{t,\delta}, \qquad C_{t,\delta} = -\frac{d}{2}\log\left[2\pi e\left(\delta^2 + \frac{\sigma_t^2}{\alpha_t^2}\right)\right]. \tag{A.2}$$

The term $C_{t,\delta}$ is independent of μ. Hence, for every fixed $\delta > 0$, the multiscale KL term differs from $U_{\pi,\delta}(\mu)$ only by a term independent of the optimization variable.

We next show that $U_{t,\delta}(\mu) \to U_t(\mu)$ for every fixed $t > 0$. Since $\lim_{\delta\to 0}\sqrt{\alpha_t^2\delta^2 + \sigma_t^2} = \sigma_t$, it remains only to justify passage of the limit through the Gaussian expectation. For every fixed $t > 0$, the diffused noise density is

$$p_{N,t}(z) = \int_{\mathbb{R}^d} \frac{1}{(2\pi\sigma_t^2)^{d/2}} \exp\left(-\frac{\|z - \alpha_t u\|_2^2}{2\sigma_t^2}\right) dP_N(u). \tag{A.3}$$

Thus $p_{N,t}(z) \le (2\pi\sigma_t^2)^{-d/2}$. Moreover, since P_N is a probability distribution, there exists $R < \infty$ such that $m_R := P_N(\|u\|_2 \le R) > 0$. Restricting Eq. A.3 to this ball gives

$$\begin{aligned} p_{N,t}(z) &\ge m_R (2\pi\sigma_t^2)^{-d/2} \exp\left(-\frac{(\|z\|_2 + \alpha_t R)^2}{2\sigma_t^2}\right) \\ &\ge c_t \exp(-b_t \|z\|_2^2) \end{aligned} \tag{A.4}$$

for some $c_t, b_t > 0$. Therefore, for finite constants a_t, b_t,

$$|\log p_{N,t}(z)| \le a_t + b_t \|z\|_2^2. \tag{A.5}$$

For sufficiently small δ, $z_\delta = \alpha_t \mu + \sqrt{\alpha_t^2\delta^2 + \sigma_t^2}\epsilon$ satisfies $\|z_\delta\|_2^2 \le C_{t,\mu}(1 + \|\epsilon\|_2^2)$ for some $C_{t,\mu} < \infty$ independent of δ. Hence Eq. A.5 provides an integrable dominating function under $\epsilon \sim \mathcal{N}(0, I)$. The dominated convergence theorem yields

$$\lim_{\delta\to 0} U_{t,\delta}(\mu) \longrightarrow U_t(\mu), \tag{A.6}$$

for every fixed $t > 0$.

By the domination condition in Theorem 1, for each fixed μ there exists $G_\mu(t)$ such that $|U_{t,\delta}(\mu)| \le G_\mu(t)$ for sufficiently small δ and $\int_0^T G_\mu(t)\pi(t)\,dt < \infty$. Combining this with Eq. A.6, dominated convergence theorem gives

$$\lim_{\delta\to 0} U_{\pi,\delta}(\mu) = U_\pi(\mu). \tag{A.7}$$

For the signal-prior term, the continuity of $\mathcal{R}_X$ gives $\mathcal{R}_X(y - \mu - \delta\eta) \to \mathcal{R}_X(y - \mu)$ pointwise. Since $\mathcal{R}_X$ has at most polynomial growth, $|\mathcal{R}_X(z)| \le a + b\|z\|_2^m$ for some finite a, b, m. For sufficiently small δ, $|\mathcal{R}_X(y - \mu - \delta\eta)|$ is therefore bounded by $C_\mu(1 + \|\eta\|_2^m)$. Hence,

$$\lim_{\delta\to 0} \mathbb{E}_\eta\left[\mathcal{R}_X(y - \mu - \delta\eta)\right] = \mathcal{R}_X(y - \mu). \tag{A.8}$$

Combining Eq. A.2, A.7, A.8 and discarding terms independent of μ, the μ dependent part of $\mathcal{J}_\delta(\mu)$ converges pointwise to

$$\mathcal{J}_0(\mu) := \mathcal{R}_X(y - \mu) + U_\pi(\mu). \tag{A.9}$$

A.2 Proof of Theorem 2

We prove the effective-likelihood statement. For fixed (t, ϵ), let $f_{t,\epsilon}(n) = \alpha_t^d p_{N,t}(\alpha_t n + \sigma_t \epsilon)$. The change of variables $z = \alpha_t n + \sigma_t \epsilon$ gives $\int_{\mathbb{R}^d} f_{t,\epsilon}(n)\, dn = 1$. By the definition of U_π and Jensen's inequality,

$$e^{-U_\pi(n)} = \exp(\mathbb{E}_{t,\epsilon} \log f_{t,\epsilon}(n)) \le \mathbb{E}_{t,\epsilon} f_{t,\epsilon}(n).$$

Tonelli's theorem therefore gives

$$Z_\pi = \int_{\mathbb{R}^d} e^{-U_\pi(n)}\, dn \le \mathbb{E}_{t,\epsilon} \int_{\mathbb{R}^d} f_{t,\epsilon}(n)\, dn = 1. \tag{A.10}$$

Moreover, taking $\delta \to 0$ in the domination bound implies $|U_t(n)| \le G_n(t)$ and hence $|U_\pi(n)| < \infty$ for every fixed n. Thus $e^{-U_\pi(n)} > 0$ and $Z_\pi > 0$. Consequently, $0 < Z_\pi \le 1$ and $p_\pi^{\text{eff}}(n) = Z_\pi^{-1} e^{-U_\pi(n)}$ is a proper probability density.

Finally, setting $x = y - \mu$ in A.9 gives

$$\widehat{x}_\pi \in \arg\min_x \left\{ U_\pi(y - x) + \mathcal{R}_X(x) \right\}.$$

Since $p_X(x) \propto e^{-\mathcal{R}_X(x)}$ and $p_\pi^{\text{eff}}(y - x) \propto e^{-U_\pi(y-x)}$, this is exactly the MAP estimator under the effective noise likelihood and the original signal prior.

A.3 Proof of Theorem 3

Proof. For Gaussian noise, we have

$$p_{N,t} = \mathcal{N}(0, \Sigma_t), \qquad \Sigma_t = \alpha_t^2 \Sigma + \sigma_t^2 I = \alpha_t^2 (\Sigma + r_t I). \tag{A.11}$$

Therefore,

$$-\log \left[\alpha_t^d p_{N,t}(\alpha_t n + \sigma_t \epsilon) \right] = \frac{1}{2} (n + \sqrt{r_t}\epsilon)^\mathsf{T} (\Sigma + r_t I)^{-1} (n + \sqrt{r_t}\epsilon) + c_t, \tag{A.12}$$

where c_t is independent of n. Taking expectation over $\epsilon \sim \mathcal{N}(0, I)$ yields

$$\mathbb{E}_\epsilon[\cdot] = \frac{1}{2} n^\mathsf{T} (\Sigma + r_t I)^{-1} n + \widetilde{c}_t. \tag{A.13}$$

Averaging over $t \sim \pi$ gives

$$U_\pi(n) = \frac{1}{2} n^\mathsf{T} W_\pi n + c_\pi, \qquad W_\pi = \int_0^T \pi(t) (\Sigma + r_t I)^{-1} dt, \tag{A.14}$$

which proves 19. It follows directly that

$$p_\pi^{\text{eff}} = \mathcal{N}(0, W_\pi^{-1}). \tag{A.15}$$

Let

$$\Sigma = Q\ \text{diag}(\lambda_1, \ldots, \lambda_d) Q^\mathsf{T}, \tag{A.16}$$

with $0 < \lambda_{\min} \le \lambda_i \le \lambda_{\max}$. Then

$$W_\pi = Q\ \text{diag}(h_\pi(\lambda_1), \ldots, h_\pi(\lambda_d))\, Q^\mathsf{T}, \tag{A.17}$$

where

$$h_\pi(\lambda) := \int_0^T \frac{\pi(t)}{\lambda + r_t}\, dt. \tag{A.18}$$

Since h_π is decreasing,

$$\text{cond}(W_\pi) = \frac{h_\pi(\lambda_{\min})}{h_\pi(\lambda_{\max})} \ge 1. \tag{A.19}$$

For every $r_t \ge 0$,

$$\frac{1}{\lambda_{\min} + r_t} \le \frac{\lambda_{\max}}{\lambda_{\min}} \frac{1}{\lambda_{\max} + r_t}. \tag{A.20}$$

Averaging over π yields

$$\text{cond}(W_\pi) \le \frac{\lambda_{\max}}{\lambda_{\min}} = \text{cond}(\Sigma). \tag{A.21}$$

If $\Sigma = \sigma_N^2 I$, then $W_\pi = cI$ for some $c > 0$, so the resulting objective differs from the conventional white-Gaussian objective only by a scalar rescaling of the regularization parameter. Finally, if the weighting concentrates at $r_t = 0$, then $W_\pi \to \Sigma^{-1}$. □

B SUPPLEMENTARY EXPERIMENTS AND IMPLEMENTATION DETAILS

B.1 COMMON-INITIALIZATION BENCHMARK AND NOISE MODELS

We extend the main three-noise evaluation to eight distributions at observation SNRs of 5 and 10 dB. All methods use the same signal realizations, the same paired test-noise realizations within each condition, the same continuous atomic representation, and the same structural initialization.

Signal model, data splits, and evaluation. We retain $M = 32$ and $K = 3$. Frequencies are sampled uniformly on $[0, 1)$ subject to circular separation at least $4/M$; amplitude magnitudes are uniform on $[0.8, 1.2]$ and phases are uniform on $[0, 2\pi)$. Each clean signal is normalized to $\|x\|_2^2/M = 1$. For every noise family we generate 50 independent signal–noise pairs and reuse the standardized noise realizations at the two observation SNRs. Noise normalization is population based. Each distribution supplies 200,000 noise-only training vectors and 10,000 independent validation vectors for the learned methods.

All reconstruction methods use the same rank-K Hankel–ESPRIT frequency readout after signal estimation, followed by minimum-cost circular matching. The reconstruction objective itself does not use the true frequencies or K; only the evaluation readout uses $K = 3$. With $d_{\mathrm{circ}}(a, b) = \min(|a - b|, 1 - |a - b|)$, we report

$$\mathrm{RMSE}_f = \left[\frac{1}{50K} \sum_{r=1}^{50} \min_{\pi \in S_K} \sum_{k=1}^{K} d_{\mathrm{circ}}(\widehat{f}_{r,\pi(k)}, f_{r,k})^2 \right]^{1/2}. \tag{B.1}$$

All test outputs enter this statistic. We additionally report trial-level median and P90 errors. Pointwise confidence intervals in the released CSV files use 5,000 trial-bootstrap resamples, paired comparisons resample identical trial indices for the two methods being compared.

Eight noise distributions. We present the selected noise distributions in Table 2. $g, h \sim \mathcal{CN}(0, I_M)$ and are independent with each other. For the correlated Gaussian and elliptical Student-t models, one fixed random vector $v \sim \mathcal{CN}(0, I_M)$ is generated independently of all signal realizations, normalized to $\|v\|_2 = 1$, and used to define

$$C = \frac{M}{M + 49} \left(I + 49 v v^H \right), \qquad \mathrm{tr}(C)/M = 1, \qquad \mathrm{cond}(C) = 50. \tag{B.2}$$

Thus the strong covariance direction is not itself a line-spectrum atom. The projectors P_1, P_2 have mutually orthogonal rank-10 ranges generated by one fixed complex QR factorization. A separate fixed random unit vector u defines the directional-burst geometry. None of these directions depends on the test signals.

Table 2: Normalized noise laws used in the supplementary benchmark.

Noise	Distribution / generation rule
White Gaussian	$\xi = g$.
Correlated Gaussian	$\xi = C^{1/2} g$, with C in Eq. B.2.
Covariance switching	$\xi \sim \frac{1}{2}\mathcal{CN}(0, C_1) + \frac{1}{2}\mathcal{CN}(0, C_2)$, $C_{1,2} = I \pm 0.9(P_1 - P_2)$; $(C_1 + C_2)/2 = I$.
Bernoulli–Gaussian	$\xi_i = (g_i + 10 b_i h_i)/\sqrt{6}$, $b_i \sim \mathrm{Bernoulli}(0.05)$.
Laplace	$\Re\xi_i, \Im\xi_i$ independently follow $\mathrm{Laplace}(0, 1/2)$.
Independent Student-t	$\xi_i = (a_i + i b_i)/\sqrt{6}$, with independent $a_i, b_i \sim t_3$.
Elliptical Student-t	$\xi = \sqrt{3/q}\, C^{1/2} g$, $q \sim \chi^2_5$ shared across the complete vector; $\mathrm{Cov}(\xi) = C$.
Directional burst	$\xi = (g + ba u)/\sqrt{7.4}$, $b \sim \mathrm{Bernoulli}(0.1)$ and $a \sim \mathcal{CN}(0, 64M)$.

B.2 ESTIMATORS, TRAINING, AND HYPERPARAMETER SELECTION

Five estimators with the same atomic penalty. Atomic norm denoising (AND) (Bhaskar et al., 2013), covariance-based AND (Cov-AND), Huber-loss-based AND (Huber-AND), normalizing-

flow-based AND (NF-AND), and RED–Atomic all solve

$$\min_x F_m(y-x) + \lambda_A \|x\|_{\mathcal{A}}, \qquad \lambda_A = 0.6s\sqrt{M \log M}. \tag{B.3}$$

Writing $r = y - x$, AND uses $F_{\mathrm{AND}}(r) = \frac{1}{2}\|r\|_2^2$, Cov-AND uses $F_{\mathrm{Cov}}(r) = \frac{1}{2} r^H \widehat{C}^{-1} r$, and Huber-AND uses $F_{\mathrm{H}}(r) = \sum_i \rho_\delta(|r_i|)$, where $\rho_\delta(a) = a^2/2$ for $a \le \delta$ and $\delta a - \delta^2/2$ otherwise (Huber, 1992). With $C_{\mathrm{emp}} = N_{\mathrm{train}}^{-1} \sum_i \xi_i \xi_i^H$, the covariance estimate is

$$\widehat{C} = 0.998 C_{\mathrm{emp}} + 0.002 \ \mathrm{tr}(C_{\mathrm{emp}}) I/M. \tag{B.4}$$

NF-AND uses

$$F_{\mathrm{NF}}(r) = -\frac{s^2}{2} \log p_\theta(r/s), \tag{B.5}$$

with a four-block Real NVP (Dinh et al., 2016). The factors of $1/2$ in the Gaussian and flow fidelities are chosen so that the unit-covariance complex Gaussian gradient is r.

Diffusion coordinates and multiscale discretization. We use the variance-preserving convention (Song et al., 2020b) with $T = 1$, $\beta(t) = 0.1 + 19.9t$, $\alpha_t = \exp(-0.05t - 4.975t^2)$, and $\sigma_t^2 = 1 - \alpha_t^2$. The numerical additive-smoothing coordinate is

$$\tau(t) = \frac{\sigma_t}{\alpha_t} = \sqrt{\exp(0.1t + 9.95t^2) - 1}. \tag{B.6}$$

Both score training and inference use the common support $\tau \in [0.01, 0.6]$. The final RED field uses eight geometrically spaced nodes

$$\tau_j = \mathrm{geomspace}(0.01, 0.6, 8)_j, \qquad w_j = \frac{\tau_j^2}{\sum_\ell \tau_\ell^2}, \tag{B.7}$$

where $w_j \propto \tau_j^2$ corresponds to inverse internal diffusion SNR on this grid and is unrelated to observation SNR. If $\epsilon_\theta(z, \tau)$ predicts the injected $\mathcal{CN}(0, I)$ perturbation of $\xi + \tau\epsilon$, the implemented residual field is

$$\widehat{g}_{\mathrm{RED}}(r) = s \sum_{j=1}^{8} \frac{w_j}{4\tau_j} \sum_{b=1}^{4} \left[\epsilon_\theta(r/s + \tau_j \epsilon_b, \tau_j) - \epsilon_b\right]. \tag{B.8}$$

Two independent Gaussian perturbations and their negatives form the four samples at each scale, so the explicit $-\epsilon_b$ terms cancel exactly. The perturbations are fixed during optimization and paired across methods and SNRs where applicable.

Noise-model training. For independent-coordinate noise (white Gaussian, Bernoulli–Gaussian, Laplace, and independent Student-t), RED-Atomic uses a shared seven-feature coordinate network with two 64-unit hidden layers and a two-dimensional output. Training uses denoising score matching for 3,000 Adam steps, 4,096 coordinate pairs per batch, learning rate 10^{-3}, and gradient clipping at norm 5. For dependent noise, RED-Atomic uses an empirical-covariance linear skip, 64 trainable quadratic features, and a 256–256 vector multilayer perceptron (MLP). This model is trained for 6,000 Adam steps with 512 vectors per batch and learning rate 8×10^{-4}. No mixture-component labels, burst indicators, true projectors, or true covariance eigenvectors are provided to the score networks.

NF-AND uses empirical-covariance whitening and four affine coupling blocks with 64–64 tanh subnetworks. It is trained for 3,000 Adam steps with batch size 512, learning rate 5×10^{-4}, and gradient clipping at norm 5. Checkpoints, including the initial linear/Gaussian model, are evaluated every 500 steps. RED checkpoints are selected by independent validation denoising loss; NF checkpoints are selected by independent validation negative log-likelihood.

Shared initialization and hyperparameter rules. For each observation, 250 AND iterations with coefficient $0.3s\sqrt{M \log M}$ generate one lifted primal state; every final method receives a copy of this primal state and its dual variable is reset to zero. No method receives mode-path continuation in this supplementary benchmark. The final coefficient 0.6 in Eq. B.3 is inherited from the main benchmark and is fixed across methods, noise families, and SNRs. Its $s\sqrt{M \log M}$ scaling follows

the Gaussian atomic-noise calibration of AND (Bhaskar et al., 2013). The coefficient 0.6 is therefore a common experimental constant rather than a method-specific test-set choice. The Huber threshold is

$$\delta = 1.5s \ \mathrm{median}(|\xi_{\mathrm{train}}|)/\sqrt{\log 2}. \tag{B.9}$$

Covariance estimation, robust-scale estimation, and neural checkpoint selection use only noise-only training or validation data.

B.3 Recovery results and numerical diagnostics

Table 3: Frequency RMSE ($\times 10^{-3}$) over 50 paired observations per noise model and SNR. All methods start from the same AND initialization.

Noise	dB	AND	Cov-AND	Huber-AND	NF-AND	RED–Atomic
White Gaussian	5	2.135	2.134	**2.130**	2.134	2.138
	10	1.188	1.187	**1.186**	1.187	1.196
Correlated Gaussian	5	2.468	1.515	2.520	1.515	**1.465**
	10	1.325	0.829	1.362	0.829	**0.807**
Covariance switching	5	2.156	2.155	2.161	1.782	**1.565**
	10	1.160	1.159	1.161	0.995	**0.873**
Bernoulli–Gaussian	5	5.957	6.516	1.097	11.568	**1.013**
	10	1.494	1.494	0.623	1.476	**0.532**
Laplace	5	2.194	2.194	2.082	2.194	**2.022**
	10	1.215	1.215	1.153	1.215	**1.123**
Independent Student-t_3	5	2.635	2.627	1.675	2.388	**1.651**
	10	1.157	1.153	0.923	1.234	**0.909**
Elliptical Student-t_5	5	**3.463**	8.330	5.631	6.192	7.731
	10	1.287	0.846	1.257	0.975	**0.838**
Directional burst	5	37.554	0.839	40.401	0.833	**0.755**
	10	43.049	0.466	44.337	0.463	**0.423**

Table 4: Trial-level median/P90 frequency RMSE ($\times 10^{-3}$). Each entry is median/P90 over the same 50 observations used in Table 3.

Noise	dB	AND	Cov-AND	Huber-AND	NF-AND	RED–Atomic
White Gaussian	5	1.850/3.113	1.856/3.110	1.848/3.110	1.856/3.110	1.838/3.087
	10	1.033/1.720	1.040/1.717	1.029/1.718	1.040/1.717	1.072/1.718
Correlated Gaussian	5	1.988/3.075	1.434/2.006	2.035/3.039	1.434/2.006	1.333/1.999
	10	1.087/1.692	0.795/1.086	1.085/1.668	0.795/1.086	0.748/1.105
Covariance switching	5	2.037/2.867	2.032/2.869	2.017/2.861	1.639/2.574	1.447/2.215
	10	1.067/1.589	1.073/1.602	1.053/1.588	0.931/1.467	0.809/1.248
Bernoulli–Gaussian	5	1.588/4.883	1.609/4.888	0.867/1.617	1.606/4.360	0.771/1.261
	10	0.868/2.370	0.886/2.370	0.495/0.943	0.803/2.100	0.427/0.685
Laplace	5	1.973/3.129	1.959/3.138	1.780/2.996	1.959/3.138	1.864/2.690
	10	1.094/1.690	1.085/1.689	1.015/1.592	1.085/1.689	1.065/1.515
Independent Student-t_3	5	1.666/3.012	1.649/3.033	1.430/2.199	1.930/3.681	1.380/2.489
	10	0.914/1.698	0.906/1.669	0.787/1.268	1.066/1.860	0.776/1.373
Elliptical Student-t_5	5	1.523/3.440	1.019/2.891	1.542/3.308	0.994/2.867	0.944/2.764
	10	0.906/1.822	0.565/1.451	0.891/1.800	0.578/1.372	0.548/1.438
Directional burst	5	0.691/1.658	0.732/1.199	0.699/1.618	0.720/1.177	0.663/1.049
	10	0.387/0.931	0.405/0.675	0.395/0.832	0.396/0.668	0.377/0.596

Table 5: Analytic calibration for the two rerun dependent-noise models (frequency RMSE, $\times 10^{-3}$). "Exact NLL" uses the physical likelihood; "Exact smoothed" replaces the learned RED field by the exact diffusion-smoothed score while keeping the same multiscale weights and solver.

Noise	dB	Exact NLL	Exact smoothed	Learned RED
Correlated Gaussian	5	1.515	1.464	1.465
	10	0.829	0.807	0.807
Elliptical Student-t_5	5	3.515	7.913	7.731
	10	1.138	0.851	0.838

Table 6: Terminal numerical diagnostics pooled over 800 outputs per method. Stop is the percentage satisfying the iterate-change and feasibility criteria.

Method	Stop (%)	Median r_s	Median r_c	Max. PSD violation
AND	99.8	1.9×10^{-6}	4.1×10^{-9}	6.4×10^{-7}
Cov-AND	99.8	2.5×10^{-6}	8.0×10^{-9}	8.0×10^{-7}
Huber-AND	99.4	5.8×10^{-6}	1.2×10^{-8}	1.4×10^{-6}
NF-AND	99.6	1.6×10^{-6}	3.2×10^{-9}	1.9×10^{-5}
RED–Atomic	100.0	3.0×10^{-5}	2.0×10^{-8}	2.4×10^{-6}

Gaussian benchmarks. White-Gaussian results provide a matched-model sanity check: all five methods are nearly indistinguishable at both SNRs. Under correlated Gaussian noise, exploiting anisotropy is important. At 5 dB, AND has RMSE 2.468×10^{-3}, whereas Cov-AND and RED–Atomic obtain 1.515 and 1.465×10^{-3}; at 10 dB the corresponding values are 1.325, 0.829, and 0.807×10^{-3}. The same ordering is visible in the median/P90 summaries. Table 5 confirms that the learned RED field closely tracks the analytic Gaussian targets: its RMSE agrees with the exact smoothed score to three displayed decimals at both SNRs and remains close to the exact Gaussian likelihood. This reproduces the expected behavior that covariance-aware and score-based fidelities exploit the same second-order geometry in the Gaussian case.

Non-Gaussian and structured noise. The covariance-switching mixture separates distributional information from unconditional second-order statistics. Since $(C_1 + C_2)/2 = I$, Cov-AND reduces to the same unconditional quadratic model as AND, while NF-AND and RED–Atomic can respond to component-dependent structure. RED–Atomic reaches $1.565/0.873 \times 10^{-3}$ at 5/10 dB, compared with $2.155/1.159 \times 10^{-3}$ for Cov-AND. Under Bernoulli–Gaussian contamination, Huber-AND provides a strong robust baseline, while RED–Atomic achieves $1.013/0.532 \times 10^{-3}$ versus $1.097/0.623 \times 10^{-3}$ for Huber-AND. Laplace and independent Student-t_3 show the same broad pattern: robust or learned nonlinear fidelities improve on the quadratic baselines, with RED–Atomic providing the lowest displayed pooled RMSE. Directional-burst noise is dominated by joint covariance structure; Cov-AND and NF-AND are already strong, and RED–Atomic further reduces the RMSE to $0.755/0.423 \times 10^{-3}$.

Elliptical Student-t and tail-sensitive metrics. The elliptical Student-t_5 model differs from the independent Student-t_3 case because one random radial scale multiplies the entire vector. At 5 dB, the trial-level median/P90 values favor RED–Atomic ($0.944/2.764 \times 10^{-3}$), followed closely by NF-AND and Cov-AND, but a few high-energy radial draws dominate the squared aggregation used by pooled RMSE. Consequently, the pooled 5-dB RMSE is 7.731×10^{-3} for RED–Atomic and 6.192×10^{-3} for NF-AND, while the simpler AND value is 3.463×10^{-3}. At 10 dB the tail effect is much weaker and RED–Atomic attains the lowest pooled RMSE, 0.838×10^{-3}. Reporting median and P90 together with pooled RMSE therefore distinguishes typical localization accuracy from a small number of heavy-tail failures.

Analytic likelihood and exact-score calibration. Table 5 separates score-learning error from the effect of the multiscale objective itself. For correlated Gaussian noise, the learned RED result is essentially identical to the exact smoothed score and closely matches the exact likelihood, showing that the dependent-noise score model captures the Gaussian geometry accurately. For elliptical

Student-t_5, the exact smoothed score gives $7.913/0.851\times10^{-3}$ at 5/10 dB, while learned RED gives $7.731/0.838\times10^{-3}$. Their close agreement indicates that the remaining low-SNR tail failures are primarily associated with the chosen multiscale effective likelihood and the heavy-tailed realizations, rather than with insufficient score-network training. The exact physical Student-t likelihood behaves differently ($3.515/1.138\times10^{-3}$), illustrating that diffusion smoothing changes the effective fidelity as characterized by the main-text theory.

Numerical behavior. Table 6 summarizes terminal stationarity, complementarity, and PSD-feasibility diagnostics for the implemented solver. More than 99% of runs satisfy the iterate-change and feasibility stopping criteria for every method, and the reported residuals remain small relative to the optimization scale. These numerical summaries are reported separately from the recovery metrics because a well-resolved optimization equilibrium and an accurate spectral estimate are distinct properties.

B.4 Primal–dual splitting: convex guarantees and learned-fidelity use

Implemented splitting. The lifted matrix state is $Z=\left[\begin{smallmatrix} T & x \\ x^H & \rho \end{smallmatrix}\right]$, where T is Hermitian Toeplitz. Let $\mathcal{S}$ denote this linear structural subspace and define $D_\lambda=\mathrm{diag}(\lambda_A/(2M),\dots,\lambda_A/(2M),\lambda_A/2)$. If $g(r)$ denotes the residual gradient/field, then $G(Z)$ has signal blocks $-g(r)/2$ and $-g(r)^H/2$ and zero diagonal blocks. The matrix implementation of the primal–dual update is

$$\begin{aligned} Z^{k+1} &= \Pi_{\mathcal{S}}\left[Z^k - P\{D_\lambda + G(Z^k) + Y^k\}\right], \\ Y^{k+1} &= \Pi_{\mathbb{H}_-}\left[Y^k + \zeta(2Z^{k+1} - Z^k)\right]. \end{aligned} \tag{B.10}$$

Here $\zeta=0.6$. The diagonal preconditioner uses $0.97/\zeta$ on structural entries and $0.97/(\zeta+L/4)$ on the two signal blocks. We set $L=1$ for AND and Huber-AND, $L=\|\widehat{C}^{-1}\|_2$ for Cov-AND, and $L=100$ for RED. NF starts from $L=25$ and doubles it for an observation until the local quadratic upper-bound test is satisfied. Each final solve has a 2,000-iteration cap. Every 100 iterations the implementation checks relative primal change, relative dual change, and relative PSD violation against 3×10^{-5}.

Convex case. Condat–Vũ splitting was developed for convex composite optimization (Condat, 2013; Vũ, 2013). For a problem

$$\min_z f(z)+\psi(z)+\chi(\mathcal{K}z), \tag{B.11}$$

assume that ψ and χ are proper closed convex functions, f is convex with an L_f-Lipschitz gradient, the proximal maps are evaluated exactly, and the primal–dual solution set is nonempty. Under the standard step-size/metric conditions, the primal–dual iterates converge to a primal–dual solution. For the unrelaxed scalar-step form, a sufficient strict condition is

$$\eta^{-1}-\zeta\|\mathcal{K}\|^2 > L_f/2. \tag{B.12}$$

The diagonal implementation uses the corresponding positive-definite metric condition. AND, Cov-AND, and Huber-AND retain this convex structure regardless of the actual test-noise distribution. For exact Gaussian noise, the population RED potential is also quadratic and convex, consistent with the Gaussian result in the main text; a learned score model, however, need not be exactly convex merely because its training data are Gaussian.

Learned nonconvex fidelities. For NF-AND and general learned RED–Atomic, the smooth fidelity may be nonconvex, so the classical convex theorem does not provide a global convergence result for the implemented iteration. The same splitting is used as a structured fixed-point method. A useful conditional characterization follows directly from the update map: if the iterates converge, the limiting step metric is positive definite, and the fixed learned residual field is continuous, then the limiting pair $(\bar{Z},\bar{Y})$ satisfies

$$\Pi_{\mathcal{S}}(D_\lambda+\widehat{G}(\bar{Z})+\bar{Y})=0,\quad \bar{Z}\succeq 0,\quad \bar{Y}\preceq 0,\quad \mathrm{Re}\langle\bar{Y},\bar{Z}\rangle_F=0. \tag{B.13}$$

If $\widehat{G}$ is the gradient of a differentiable learned fidelity, Eq. B.13 gives the corresponding first-order KKT conditions. For a generic learned score field that is not exactly conservative, it instead characterizes an equilibrium between the learned vector field and the cone constraint (Reehorst & Schniter, 2018).

Reported numerical residuals. In addition to the iterate-change and PSD-feasibility tests, we record the normalized stationarity residual

$$r_s = \frac{\|\Pi_{\mathcal{S}}(D_\lambda + \widehat{G} + Y)\|_F}{\max\{1, \|D_\lambda\|_F + \|\widehat{G}\|_F + \|Y\|_F\}}, \tag{B.14}$$

and the normalized cone-complementarity residual

$$r_c = \frac{|\operatorname{Re}\langle Y, Z\rangle_F|}{\max\{1, \|Y\|_F \|Z\|_F\}}. \tag{B.15}$$

Table 6 summarizes these diagnostics across all 4,000 reconstruction outputs. They complement the recovery metrics by separately tracking the equilibrium and cone conditions of the implemented solver.