Preconditioned Implicit Midpoint Langevin Sampling for Non-Smooth Bayesian Imaging

Sichen Wang[1] **Zhipeng Lu**[1,2]
[1]Department of Engineering, Shenzhen MSU-BIT University
[2]Guangdong Laboratory of Machine Perception and Intelligent Computing
Shenzhen, China
Correspondence: zhipeng.lu@smbu.edu.cn

Abstract

We introduce P-IMLA, a preconditioned implicit-midpoint Langevin sampler for high-dimensional log-concave targets with non-smooth regularization. The obstruction is that the large steps enabled by matrix preconditioning are precisely where preconditioned Euler–Maruyama inflates stationary variance. Implicit midpoint removes this surplus via a Cayley-transform identity, making P-IMLA Gaussian-exact at any step size. Beyond Gaussians, whitening transfers the sharp IMLA contraction rate to κ_{eff}, and backward-error analysis gives drift-only rather than Euler-type diffusion bias under explicit smoothness assumptions. An M-norm Moreau envelope yields a proximal non-smooth variant with a three-term error budget. Controlled Gaussian and TV-regularized imaging experiments confirm the predicted separation: the unpreconditioned midpoint scheme retains long transients as κ_f grows, whereas P-IMLA remains κ_{eff}-governed and near variance-exact.

1 Introduction

Large-scale Bayesian inverse problems—image deblurring, computed tomography, magnetic resonance imaging [Stuart, 2010]—give rise to posterior distributions $\pi(x) \propto \exp(-f(x) - g(x))$ in which f is a smooth data-fit term and g is a non-smooth regulariser such as total variation or wavelet-ℓ_1 [Pereyra, 2016, Durmus et al., 2018, Habring et al., 2024]. At image resolutions of 256×256 and above, the Hessian of f routinely exhibits condition numbers κ_f on the order of 10^5–10^7, inflating the mixing time of the unadjusted Langevin algorithm [ULA; Roberts and Tweedie, 1996] for smooth strongly log-concave targets to $O(\kappa_f/\varepsilon^2)$ iterations [Dalalyan, 2017, Durmus and Moulines, 2019, Cheng and Bartlett, 2018, Vempala and Wibisono, 2019, Chewi, 2026]—in practice, computationally out of reach.

On a smooth potential U, ULA takes $X_{n+1} = X_n - \delta\nabla U(X_n) + \sqrt{2\delta}\,\xi_n$, with $\xi_n \sim \mathcal{N}(0, I)$. Preconditioning is the classical remedy. Replacing the Langevin drift by $-M^{-1}\nabla f$ with $M \approx \nabla^2 f$ equalises curvatures across coordinates and, in principle, lowers the effective condition number $\kappa_{\text{eff}} = L_M/m_M$ toward $O(1)$, where m_M, L_M bound the eigenvalues of $M^{-1/2}\nabla^2 f M^{-1/2}$, echoing long-established accelerations in iterative linear solvers and optimisation [Girolami and Calderhead, 2011, Titsias, 2023, Hird and Livingstone, 2025]. Yet the preconditioned unadjusted Langevin algorithm (P-ULA) faces an obstruction: at large dimensionless steps $\delta\ell$, its explicit-Euler discretisation incurs variance inflation $1/(1-\delta\ell/2)$ on a whitened mode of precision ℓ, offsetting the intended conditioning gain.

An algebraic resolution. On a scalar Gaussian target with precision ℓ, every θ-scheme reduces the Langevin diffusion to an AR(1) recursion $z_{n+1} = r\,z_n + \sigma\,\xi_n$ with stationary variance $\sigma^2/(1-r^2)$. Matching this to the target variance $v = \ell^{-1}$ is equivalent to the *variance-budget identity* $r^2 + \sigma^2/v =$

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

	No preconditioning ($M = I$)	**Preconditioned** ($M \approx \nabla^2 f$)
Explicit Euler	slow-mixing ULA: VR $= 0.24$	unstable P-ULA: VR $\to \infty$
Implicit midpoint	long transient IMLA: VR $= 48.6$	**rapid relaxation** **P-IMLA: VR** $= 1.0246 \pm 0.0001$

Figure 1: **The 2×2 landscape at a fixed simulation budget.** Gaussian target with mean μ, precision H and $\kappa_f = 4 \times 10^7$ (App. C.2). VR denotes the time-averaged normalised mean-square displacement from the target mean, $d^{-1}\overline{(X-\mu)^\top H(X-\mu)}$, including the cold-start transient. It is not a stationary variance estimate. Both midpoint schemes are Gaussian-exact with exact solves; 98.7% of IMLA's displayed 48.6 is transient mean error (App. D.5). Preconditioning makes that transient short at the reported budget.

1. Direct computation, with $a := \delta\ell/2$, gives two qualitatively different outcomes:

$$\text{explicit Euler:} \quad r^2 + \sigma^2/v \;=\; (1-2a)^2 + 4a \;=\; 1 + 4a^2, \tag{1}$$

$$\text{implicit midpoint } (\theta = \tfrac{1}{2}): \quad r^2 + \sigma^2/v \;=\; \frac{(1-a)^2 + 4a}{(1+a)^2} \;\equiv\; 1. \tag{2}$$

The $4a^2$ surplus in (1) is precisely what preconditioning's large steps amplify; its exact cancellation in (2)—unique to $\theta = 1/2$ among θ-schemes, traceable to the Cayley transform $r(a) = (1-a)/(1+a)$—is the algebraic cancellation that this paper builds on, making implicit-midpoint integration a natural ingredient for realising large-step preconditioning. Figure 1 maps the resulting 2×2 landscape.

Contributions. We introduce the *Preconditioned Implicit Midpoint Langevin Algorithm* (P-IMLA) and its non-smooth variant P-IMLA$_\lambda$ based on an M-norm Moreau envelope; to our knowledge this is the first matrix-preconditioned implicit-midpoint proximal Langevin sampler. Our contributions are:

1. **Algorithm (§3).** An inner solver computes $\mathrm{prox}^M_{(\delta/2)U_\lambda}$ for $U_\lambda = f + g^{\lambda,M}$ at $O(K_{\mathrm{P}} K_{\mathrm{in}} d \log d)$ per outer iteration on convolutional models with TV priors, matching MYULA's per-step cost; an engineering variant JP-IMLA (App. B) gives $\sim 60\times$ per-iteration speedup.
2. **Gaussian exactness and contraction (Theorems 2–3).** On Gaussian targets, P-IMLA's stationary distribution equals π exactly at every step size—via the identity in (2). On general m-strongly log-concave targets, whitening transfers the reflected-proximal bound of Klatzer et al. [2024] to an M-norm Wasserstein contraction rate $\rho^\star = (\sqrt{\kappa_{\mathrm{eff}}} - 1)/(\sqrt{\kappa_{\mathrm{eff}}} + 1)$, with quadratic targets attaining the bound.
3. **Sharp drift-only bias (Theorem 5, primary novelty).** A backward-error analysis under the regularity and Poisson assumptions of Theorem 5 gives the asymptotic bias formula

$$\pi_\delta(\varphi) - \pi(\varphi) \;=\; \tfrac{\delta}{4}\,\mathrm{Cov}_\pi\big(\,\mathrm{tr}(M^{-1}\nabla^2 f),\, \varphi\big) + o(\delta), \tag{3}$$

   whose bias operator is first-order (drift-only), in structural contrast to P-ULA's second-order drift-*and*-diffusion bias. The drift-only form forces the bias to vanish on Gaussians (where $\mathrm{tr}(M^{-1}\nabla^2 f)$ is constant) while retaining an $O(\delta)$ stationary weak bias in general. Here "first-order" describes the differential operator, not an improvement in the order in δ.
4. **End-to-end error budget (Theorem 6).** A three-term W_2 decomposition—contraction + discretisation + Moreau smoothing—with an explicit $C^{1,1}$ discretisation remainder and coupled (δ, λ) dependence; see Rem. 7 for the fixed-λ (P1) vs full-ε (P2) regimes. We further verify (App. C.9) the smoothing term's $\sqrt{\lambda}$ exponent is non-tight on polyhedral g (convex functions with polyhedral epigraphs, including ℓ_1 penalties [Bauschke and Combettes, 2011]; empirical slopes 0.97 asymmetric / 1.95 symmetric, both strictly above the worst-case exponent 0.5).
5. **Imaging-scale experiments at fixed iteration budgets (§5).** On 256×256 TV deconvolution at a $100\times$ iteration-budget ratio, P-IMLA$_\lambda$ reaches cold-start per-pixel $\hat{R} < 1.05$ at iteration 10 (contraction timescale $N \approx 22$, not a theoretical $\hat{R}$ threshold at $\kappa^\lambda_{\mathrm{eff}} = 201$) and $\hat{R} \le 1.001$ at $N = 2{,}000$ with SSE $= 855.22 \pm 0.66$, while SP-MYULA reaches near-unity $\hat{R}$ at $\sim 4\times$ larger

SSE; a MALA comparison (App. C.12) gives covariance-trace ratios 7.68 vs 0.95, complemented by overdispersed-start diagnostics and target-exact small-image Gibbs checks (App. D), and MYULA / SK-ROCK fail to mix. The $3.5\times$–$4.0\times$ advantage is stable across geometric, Shepp–Logan, and cameraman phantoms.

Related work. Prior work splits into two regimes. The *preconditioned-Langevin* thread [Girolami and Calderhead, 2011, Titsias, 2023, Hird and Livingstone, 2025] reshapes sampling geometry, while explicit Euler retains a stability and bias trade-off. SK-ROCK [Pereyra et al., 2020] stabilises the explicit step via Chebyshev stages, and Corbineau et al. [2019] combines preconditioning with non-smooth imaging priors. The *implicit-integrator* thread [Hodgkinson et al., 2021, Klatzer et al., 2024] studies implicit Langevin discretisations, alongside the geometric-integrator-for-sampling tradition of Bou-Rabee and Sanz-Serna [2018]; the semi-implicit pCNL [Cotter et al., 2013] treats a Gaussian-prior covariance implicitly with Metropolis correction. Underdamped Langevin [Cheng et al., 2018, Leimkuhler et al., 2024], HMC/NUTS [Neal, 2011, Hoffman and Gelman, 2014], and Wasserstein-gradient-flow samplers [Wibisono, 2018] provide complementary acceleration mechanisms outside our controlled integrator comparison. IMLA [Klatzer et al., 2024] ($M = I$, $\tilde{O}(\sqrt{\kappa_f})$) is the closest predecessor. At large κ_f, its low-curvature modes retain long initial transients; the VR $= 48.6$ in Figure 1 is a finite-budget statistic, not a failure of Gaussian exactness. For fixed M, P-IMLA is exactly IMLA in whitened coordinates: Theorem 3 transports the factor of Klatzer et al. [2024], Thm. 3.3, while relaxing C^2 to $C^{1,1}$ regularity. We differ in the M-norm Moreau construction, stationary-bias analysis, and treatment of the joint-proximal implementation. Coupling methods [Eberle, 2016, Klatzer et al., 2024] control convergence distances; building on weak-error analysis [Talay and Tubaro, 1990, Abdulle et al., 2014], we identify the leading invariant coefficient and its drift-only defect, under explicit regularity and Poisson assumptions. The generalized Hessian therefore determines both the rate and the leading bias, which vanishes on Gaussians while remaining first order in the step size for general smooth targets. The non-smooth construction extends proximal Langevin methods [Pereyra, 2016, Durmus et al., 2018, 2022, Habring et al., 2026] using an M-norm envelope whose smoothness constant adds to L_M in the same metric. Recent Bayesian imaging work [Laumont et al., 2022, Holden et al., 2022, Ehrhardt et al., 2024, Renaud et al., 2025] provides a broader modelling and algorithmic context for this construction. Our JP bounds separately quantify the change of proximal model and the finite inner-solver error.

Table 1: Notation. The preconditioner M is fixed throughout.

Symbol	Meaning
$\pi,\ \pi_\lambda$	Original and Moreau-smoothed targets
$\delta,\ \lambda,\ \tau = \delta/2$	Langevin step, Moreau parameter, proximal step
$\ell,\ m_M,\ L_M$	Spectral value and curvature bounds in the M-metric
$L_\lambda = L_M + \lambda^{-1}$	Smoothness bound for $U_\lambda = f + g^{\lambda,M}$
$P,\ Q,\ \widehat{Q}$	Exact smoothed, exact joint-proximal, implemented JP updates
$K_{\mathrm{P}},\ K_{\mathrm{in}}$	Picard and inner ADMM iteration budgets

2 Background

2.1 Problem setup and preconditioned Langevin

We sample from $\pi(x) \propto \exp(-f(x) - g(x))$ on $\mathbb{R}^d$ as introduced in §1, under the following structural hypothesis throughout:

Assumption 1 (Target regularity)**.** *$f \in C^2(\mathbb{R}^d)$ is m-strongly convex and L-smooth, i.e. $m\,I \preceq \nabla^2 f(x) \preceq L\,I$ for every x; and $g\colon \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is proper, convex, and lower semicontinuous, but possibly non-smooth.*

A fixed, position-independent symmetric positive-definite *preconditioner* $M \succ 0$ reshapes the Langevin diffusion:

$$\mathrm{d}X_t \;=\; -M^{-1}\nabla f(X_t)\,\mathrm{d}t \;+\; \sqrt{2}\,M^{-1/2}\,\mathrm{d}W_t, \tag{4}$$

whose invariant distribution is still $\pi \propto \exp(-f)$ (restricting momentarily to $g \equiv 0$): the preconditioner accelerates mixing without distorting the target. The effective strong-convexity and smoothness

constants of f in the induced M-norm $\|v\|_M^2 := \langle v, Mv\rangle$ are m_M and L_M, defined by

$$m_M\, I \;\preceq\; M^{-1/2}\,\nabla^2 f(x)\,M^{-1/2} \;\preceq\; L_M\, I \quad \forall\, x, \qquad \kappa_{\text{eff}} \;:=\; \frac{L_M}{m_M}. \tag{5}$$

When $M \approx \nabla^2 f$ (the optimization-motivated "Newton-like" choice), $\kappa_{\text{eff}} = O(1)$ and the preconditioner has, in principle, fully absorbed the conditioning of the target. The M-induced Wasserstein distance W_2^M on $\mathcal{P}_2(\mathbb{R}^d)$—defined by replacing Euclidean cost with $\|\cdot\|_M^2$ in the standard optimal-transport formulation—is the natural metric for quantifying convergence of preconditioned chains.

Whitening. Let $T = M^{1/2}$ and $\widetilde{f}(z) = f(T^{-1}z)$, $\widetilde{g}(z) = g(T^{-1}z)$. Then $T\,\mathrm{prox}_{\tau f}^{M}(w) = \mathrm{prox}_{\tau \widetilde{f}}(Tw)$, $g^{\lambda,M}(T^{-1}z) = \widetilde{g}^{\lambda}(z)$, and $W_2^M(\mu,\nu) = W_2(T_{\#}\mu, T_{\#}\nu)$. Thus fixed-$M$ P-IMLA is canonically conjugate to IMLA; the following calculation isolates Euler's discretisation bias in these coordinates.

Proposition 1 (P-ULA variance inflation at preconditioned step sizes)**.** *Let $f(x) = \frac{1}{2}x^\top Hx$ with $H \succ 0$, $g \equiv 0$, preconditioner $M \succ 0$, and write $\ell_1, \dots, \ell_d$ for the eigenvalues of $M^{-1/2}HM^{-1/2}$. For any $\delta > 0$ with $\delta\ell_i < 2$ for all i (the forward-Euler stability boundary), the preconditioned Euler–Maruyama iteration*

$$X_{n+1} \;=\; X_n \;-\; \delta\, M^{-1}\nabla f(X_n) \;+\; \sqrt{2\delta}\, M^{-1/2}\xi_n, \qquad \xi_n \sim \mathcal{N}(0, I),$$

admits a unique stationary distribution $\pi_\delta^{\text{Euler}}$ whose variance in generalized-Hessian eigencomponent $z_i = q_i^\top M^{1/2}x$ (with q_i a unit eigenvector of $M^{-1/2}HM^{-1/2}$) is

$$\mathrm{Var}_{\pi_\delta^{\text{Euler}}}(z_i) \;=\; \ell_i^{-1} \,\cdot\, \frac{1}{1 \;-\; \delta\ell_i/2}, \tag{6}$$

against the target's $\mathrm{Var}_\pi(z_i) = \ell_i^{-1}$. The inflation factor $(1-\delta\ell_i/2)^{-1}$ diverges as $\delta\ell_i \to 2^-$, Euler therefore trades mixing speed against stationary accuracy even after preconditioning; instability occurs at $\delta\ell_i \geq 2$.

Proof. In whitened coordinates $z = Q^\top M^{1/2}x$, where Q diagonalises $M^{-1/2}HM^{-1/2} = Q\,\mathrm{diag}(\ell)Q^\top$, the iteration decouples into scalar AR(1) processes $z_{n+1,i} = (1-\delta\ell_i)\,z_{n,i} + \sqrt{2\delta}\,\eta_{n,i}$ with $\eta_{n,i} \sim \mathcal{N}(0,1)$ independent. The stationary variance of an AR(1) $Z_{n+1} = rZ_n + \sigma\eta_n$ is $\sigma^2/(1-r^2)$; substituting $r = 1-\delta\ell_i$ and $\sigma^2 = 2\delta$ and simplifying via $1-r^2 = \delta\ell_i(2-\delta\ell_i)$ yields (6). □

2.2 The M-norm Moreau envelope

To handle a non-smooth g, we do not smooth it in the Euclidean metric—as MYULA [Durmus et al., 2018] does—but in the *preconditioned* M-metric. For a parameter $\lambda > 0$, define

$$g^{\lambda,M}(x) \;:=\; \min_{z\in\mathbb{R}^d} \Big\{\, g(z) + \tfrac{1}{2\lambda}\|z-x\|_M^2 \,\Big\}, \tag{7}$$

the *M-norm Moreau envelope* of g with parameter λ. This envelope is a $C^{1,1}$ convex surrogate for g with the key gradient-Lipschitz property

$$\big\|\nabla g^{\lambda,M}(x) - \nabla g^{\lambda,M}(y)\big\|_{M^{-1}} \;\leq\; \tfrac{1}{\lambda}\,\|x-y\|_M, \tag{8}$$

i.e. $\nabla g^{\lambda,M}$ is $(1/\lambda)$-Lipschitz *in the M-metric* (Bauschke and Combettes 2011, Prop. 12.27, applied to the M-Hilbert space $(\mathbb{R}^d, \langle\cdot,\cdot\rangle_M)$). Thus the smoothed potential $U_\lambda := f + g^{\lambda,M}$ is m_M-strongly convex and $(L_M + 1/\lambda)$-smooth in the M-norm, with effective condition number $\kappa_{\text{eff}}^{\lambda} = (L_M + 1/\lambda)/m_M$. The surrogate target is $\pi_\lambda \propto \exp(-U_\lambda)$; the Moreau bias $W_2^M(\pi_\lambda, \pi)$ is controlled as Term III of Theorem 6.

The explicit preconditioned comparator SP-MYULA uses the same surrogate:

$$X_{n+1} = X_n - \delta M^{-1}\nabla U_\lambda(X_n) + \sqrt{2\delta}\, M^{-1/2}\xi_n.$$

Its smoothing, discretisation and inner-solve errors are distinct.

2.3 Where P-IMLA sits in the landscape

Table 2 plots the proximal-Langevin landscape along the two axes of Figure 1: discretisation of the underlying Langevin diffusion, and geometry imposed on it by a preconditioner. On the explicit-Euler side, MYULA [Durmus et al., 2018] and SKROCK [Pereyra et al., 2020] achieve κ_f-dependent rates with no preconditioning, and the sole prior method at the "non-trivial M + non-smooth g" intersection, Corbineau P-ULA [Corbineau et al., 2019], retains explicit Euler—so inherits (1)'s variance inflation—and comes with no mixing-rate theory. On the implicit side, Klatzer et al. [2024] obtain the sharp $\tilde{O}(\sqrt{\kappa_f})$ rate but uniformly fix $M = I$; the semi-implicit variant pCNL [Cotter et al., 2013] treats only the prior covariance implicitly, is bound to Gaussian-prior function-space formulations, and is always Metropolis-corrected. P-IMLA, to our knowledge, is the first to combine a matrix preconditioner, implicit-midpoint discretisation, and M-metric Moreau treatment of nonsmooth priors; its rate $\tilde{O}(\sqrt{\kappa_{\text{eff}}})$ (Theorem 3) improves on Klatzer et al. [2024]'s $\tilde{O}(\sqrt{\kappa_f})$ whenever $\kappa_{\text{eff}} < \kappa_f$.

Table 2: **The proximal-Langevin landscape.** Rates reproduce the cited settings and are not uniform comparisons of total work. For IMLA and P-IMLA, the displayed $\widetilde{O}(\sqrt{\kappa})$ is contraction to the discretised invariant law; accuracy relative to the target also requires bias control (Rem. 7). "None" means no corresponding rate stated here; I-Moreau denotes Euclidean smoothing.

Method	Discretisation	M	Non-smooth g	Rate / setting
ULA	Expl. Euler	I	—	$\tilde{O}(\kappa_f/\varepsilon^2)$
MYULA	Expl. Euler	I	I-Moreau	$\tilde{O}(\kappa_\lambda/\varepsilon^2)$
SKROCK	Expl. stabilised	I	I-Moreau	$\tilde{O}(\sqrt{\kappa_\lambda})$, Gaussian
P-ULA	Expl. Euler	f only	I-Moreau	*none*
RMMALA	Euler + MH	position-dep.	—	*none*
Fisher MALA	Euler + MH	adaptive	—	*none*
pCNL	Semi-impl. + MH	prior cov.	—	dim-free, Gaussian prior
IMLA	**Impl. midpoint**	I	I-Moreau	$\tilde{O}(\sqrt{\kappa_f})$
P-IMLA (ours)	**Impl. midpoint**	**general** $\succ 0$	M**-Moreau**	$\tilde{\boldsymbol{O}}(\sqrt{\boldsymbol{\kappa_{\text{eff}}}})$

3 The P-IMLA Algorithm

Smooth case. Fix $\delta > 0$ and a constant SPD preconditioner $M \succ 0$. For smooth targets ($g \equiv 0$), the *Preconditioned Implicit Midpoint Langevin Algorithm* is the one-step kernel

$$X_{n+1} \;=\; -X_n \;+\; 2\,\mathrm{prox}^{M}_{(\delta/2)f}\Big(X_n + \tfrac{\sqrt{2\delta}}{2}\,M^{-1/2}\xi_n\Big), \qquad \xi_n \overset{\text{iid}}{\sim} \mathcal{N}(0, I), \tag{9}$$

with $\mathrm{prox}^{M}_{\tau f}(y) \coloneqq \arg\min_x\{f(x) + \frac{1}{2\tau}\|x - y\|^2_M\}$. At $M = I$ the update recovers IMLA [Klatzer et al., 2024]; for quadratic f, at $M = H = \nabla^2 f$ (or a fixed spectrally equivalent surrogate) it realises the matrix-preconditioned variant introduced here. We use a single fixed M, as in Titsias [2023], Hird and Livingstone [2025], rather than the position-dependent metric of Girolami and Calderhead [2011]—a simplification essential for the M-norm Moreau construction (§2.2) and the spectral analysis of §4.

Why this is implicit midpoint. Let $p \coloneqq \mathrm{prox}^{M}_{(\delta/2)f}(w)$ with $w \coloneqq X_n + \frac{\sqrt{2\delta}}{2}M^{-1/2}\xi_n$. The proximal first-order condition

$$p \;=\; w - \tfrac{\delta}{2}M^{-1}\nabla f(p) \tag{10}$$

combined with $X_{n+1} = 2p - X_n$ gives

$$X_{n+1} \;=\; X_n \;-\; \delta\, M^{-1}\nabla f\Big(\tfrac{X_n + X_{n+1}}{2}\Big) \;+\; \sqrt{2\delta}\, M^{-1/2}\xi_n, \tag{11}$$

since $p = (X_n + X_{n+1})/2$: the proximal operator *is* the solution of the midpoint equation (11), repackaged as a single off-the-shelf call. The proximal is strongly convex in x for every $\delta > 0$, so no step-size restriction $\delta < 2/L_M$ arises. When f is quadratic, $f(x) = \frac{1}{2}x^\top Hx - h_{\text{data}}^\top x$

Algorithm 1 Picard inner solver for $p = \mathrm{prox}^M_{(\delta/2)U_\lambda}(w)$ with $U_\lambda = f + g^{\lambda,M}$.

Require: $w \in \mathbb{R}^d$, parameters δ, λ (set $\tau \coloneqq \delta/2$), Fourier-diagonal $\hat{M}, \hat{H}, \hat{h}_{\mathrm{data}}$, periodic forward-difference operator D, ADMM penalty ρ_a, Picard steps K_{P}, ADMM steps K_{in}

Cache $\hat{D}_{\mathrm{denom}} \leftarrow \tau\hat{H} + (1+\tau/\lambda)\hat{M}$, $\hat{w} \leftarrow \mathcal{F}(w)$; warm-start x

for $k = 1, \dots, K_{\mathrm{P}}$ **do**

  $u \leftarrow \mathrm{prox}^M_{\lambda g}(x)$ via 2-block ADMM (warm-started; K_{in} FFTs + soft-threshold)

  $\hat{u} \leftarrow \mathcal{F}(u)$; $x \leftarrow \mathcal{F}^{-1}\Big(\hat{D}^{-1}_{\mathrm{denom}}\big(\tau\,\hat{h}_{\mathrm{data}} + \hat{M}\hat{w} + (\tau/\lambda)\hat{M}\hat{u}\big)\Big)$

end for

return $p \leftarrow x$

with $h_{\mathrm{data}} = A^\top y/\sigma^2_{\mathrm{noise}}$ (Gaussian likelihood, forward operator A); when additionally H, M are Fourier-diagonal, the proximal admits a closed form costing $O(d \log d)$ per iteration.

Non-smooth case: P-IMLA$_\lambda$. For $\pi \propto \exp(-(f+g))$ with g proper convex lsc, run P-IMLA on the Moreau-smoothed $U_\lambda \coloneqq f + g^{\lambda,M}$ from (7), in the spirit of Pereyra [2016], Durmus et al. [2018]'s MYULA but with M-metric Moreau and implicit-midpoint discretisation, yielding

$$X_{n+1} \;=\; -X_n \;+\; 2\,\mathrm{prox}^M_{(\delta/2)U_\lambda}\Big(X_n + \tfrac{\sqrt{2\delta}}{2}\,M^{-1/2}\xi_n\Big). \tag{12}$$

U_λ is $C^{1,1}$ with M-metric Lipschitz constant $L_M + 1/\lambda$ (8), so Theorem 3 applies; Theorem 5 requires additional smoothness and Poisson regularity; the gap $W_2^M(\pi_\lambda, \pi)$ is controlled by Theorem 6 Term III. The inner problem is the proximal of the sum U_λ. With $\tau \coloneqq \delta/2$, differentiating (7) at its unique minimizer gives $\nabla g^{\lambda,M}(p) = (M/\lambda)(p-u)$, where $u = \mathrm{prox}^M_{\lambda g}(p)$. The first-order condition $0 = \nabla f(p) + (M/\lambda)(p-u) + (M/\tau)(p-w)$ for $p = \mathrm{prox}^M_{\tau U_\lambda}(w)$ gives

$$\big(\tau\hat{H} + (1+\tau/\lambda)\hat{M}\big)\,\hat{p} \;=\; \tau\,\hat{h}_{\mathrm{data}} + \hat{M}\hat{w} + (\tau/\lambda)\,\hat{M}\,\widehat{\mathrm{prox}^M_{\lambda g}(p)}. \tag{13}$$

The LHS is a closed-form solve when H, M are circulant (so $\hat{H}, \hat{M}$ are diagonal); for anisotropic TV, $g(x) = \rho_{\mathrm{TV}}\|Dx\|_1$, the inner M-norm proximal is computed by a 2-block ADMM with the constraint $z = Dx$ and soft-thresholding in z [Boyd et al., 2011]. We solve (13) by Picard iteration on p, giving Algorithm 1.

With exact inner proximals, the Picard map is a contraction with factor $(\tau/\lambda)/(1 + \tau m_M + \tau/\lambda) = 0.929$ at the imaging parameters ($m_M = 1$, $\delta = 0.141$, $\lambda = 5 \times 10^{-3}$). Warm starts reduce the initial residual; a fixed sweep count alone is not an accuracy certificate. Per-outer cost is $O(K_{\mathrm{P}} K_{\mathrm{in}} d \log d)$, with the TV proximal computed inexactly in practice.

Engineering optimization (JP-IMLA). A practical variant *JP-IMLA* replaces the nested Picard solve with a single 2-block ADMM evaluating $\mathrm{prox}^M_{\tau(f+g)}$ (App. B); its reported finite-budget SSE differs from the Picard result by at most 0.7% across imaging phantoms (Table 8) at $\sim 60\times$ per-iter speedup. The main experiments use Picard; the ideal theorems concern exact proximal updates. Appendix B separates exact-kernel and finite-solver guarantees; Table 7 reports separate budget timings and pointwise certificates, which do not certify every production update.

4 The Theory of P-IMLA

The theoretical analysis chains four results, all proved in Appendix A. Theorem 2 establishes Gaussian exactness at any step size. Theorem 3 gives the corresponding reflected-proximal contraction for strongly convex $C^{1,1}$ potentials; for fixed M, this is the sharp IMLA factor in whitened coordinates [Klatzer et al., 2024]. Theorem 5 characterises the leading stationary bias via backward error analysis (BEA): its defect is a *first-order differential operator*, whereas P-ULA's also contains a second-order term. Theorem 6 gives an error budget for non-smooth targets. Write $r(\ell) = (1 - \delta\ell/2)/(1 + \delta\ell/2)$ for a spectral value ℓ; λ denotes the Moreau parameter.

4.1 Gaussian exactness and Cayley contraction

Theorem 2 (Gaussian exactness)**.** *Let $f(x) = \frac{1}{2}x^\top Hx - h^\top x$ with $H \succ 0$, $g \equiv 0$, $\pi = \mathcal{N}(H^{-1}h,\, H^{-1})$, and preconditioner $M \succ 0$ with $\kappa_{\mathrm{eff}} = L_M/m_M$ as in* (5)*. For every*

$\delta > 0$, the P-IMLA iteration (9) *has π as its unique stationary distribution, and contracts as $W_2^M(\mathrm{Law}(X_N), \pi) \le \rho(\delta)^N W_2^M(\mu_0, \pi)$ with $\rho(\delta) = \max\{|r(m_M)|, |r(L_M)|\}$. The minimax-optimal step size $\delta^\star = 2/\sqrt{m_M L_M}$ achieves $\rho^\star = (\sqrt{\kappa_{\mathrm{eff}}}-1)/(\sqrt{\kappa_{\mathrm{eff}}}+1)$, giving mixing complexity $N = O(\sqrt{\kappa_{\mathrm{eff}}}\,\log(1/\varepsilon))$.*

4.2 Contraction of the reflected proximal map

Theorem 3 (M-norm Wasserstein contraction)**.** *Let $f \in C^{1,1}(\mathbb{R}^d)$ be m_M-strongly convex and L_M-smooth in the M-norm, $g \equiv 0$. For every $\delta > 0$, two P-IMLA chains X_n, Y_n driven by shared Gaussian noise satisfy $\|X_{n+1} - Y_{n+1}\|_M \le \rho(\delta)\,\|X_n - Y_n\|_M$ almost surely, with $\rho(\delta)$ as in Theorem 2; the bound is attained by quadratic f (Gaussian worst case). Consequently P-IMLA admits a unique stationary distribution π_δ and $W_2^M(\mathrm{Law}(X_N), \pi_\delta) \le \rho(\delta)^N\, W_2^M(\mu_0, \pi_\delta)$.*

Remark 4 (Contraction and empirical convergence)**.** *The bound controls distance to the discrete invariant law, not a threshold for $\widehat{R}$ or reconstruction error. The diagnostics in §5 measure empirical convergence separately. Gaussian exactness also holds at $M = I$; preconditioning improves the finite-budget contraction factor through κ_{eff}.*

4.3 Sharp drift-only bias via backward error analysis

Theorem 5 (Drift-only BEA bias)**.** *Let $g \equiv 0$, and let $f \in C^6(\mathbb{R}^d)$ be globally m_M-strongly convex and L_M-smooth in the fixed M-norm, with derivatives of orders 3–6 of at most polynomial growth. For an observable φ, assume that $\mathcal{L}u_\varphi = -(\varphi - \pi(\varphi))$ has a solution $u_\varphi \in C^6_{\mathrm{pol}}$, where $\mathcal{L}u = -M^{-1}\nabla f \cdot \nabla u + M^{-1} : \nabla^2 u$ and C^6_{pol} means polynomial growth of the function and its derivatives through order six. Define $h(x) := \mathrm{tr}(M^{-1}\nabla^2 f(x))$. Then, as $\delta \downarrow 0$,*

$$\pi_\delta(\varphi) - \pi(\varphi) \;=\; \tfrac{\delta}{4}\,\mathrm{Cov}_\pi\big(h,\,\varphi\big) \;+\; o(\delta). \tag{14}$$

The required δ-uniform polynomial moments are proved in Lemma 10. On Gaussian targets, h is constant and the coefficient vanishes, consistently with Theorem 2.

Why drift-only. The weighted expansion gives $P_\delta u = u + \delta\mathcal{L}u + \delta^2(\mathcal{L}^2 u/2 + B_2 u) + o(\delta^2)$. The proximal Jacobian identity $SM^{-1}S^\top = M^{-1} - \delta\Gamma + O(\delta^2)$, with $S = (I + (\delta/2)M^{-1}\nabla^2 f)^{-1}$ and $\Gamma = M^{-1}\nabla^2 f M^{-1}$, cancels the second-order modification, leaving $B_2 u = \frac{1}{4}(M^{-1}\nabla h)\cdot\nabla u$. P-ULA retains a diffusion modification (Appendix A.4); both schemes generally have $O(\delta)$ weak invariant bias. Theorem 5 does not assume a density expansion: the leading W_2^M coefficient is conditional on the additional $L^3(\pi)$ condition in Proposition 13. Under stronger C^8 and two-level Poisson regularity, Corollary 12 gives second-order Richardson cancellation. A TV Moreau envelope is generally only $C^{1,1}$, so its guarantee below does not invoke this BEA theorem.

4.4 End-to-end error budget for non-smooth targets

Theorem 6 (End-to-end W_2^M bound)**.** *Let $\pi \propto \exp(-f - g)$ with f as in Assumption 1 and g convex and globally L_g^M-Lipschitz in the M-norm. Set $U_\lambda = f + g^{\lambda,M}$ and $L_\lambda = L_M + 1/\lambda$. For $\lambda > 0$, $0 < \delta \le 2/L_\lambda$ and $N \ge 1$, the exact P-IMLA$_\lambda$ kernel satisfies*

$$W_2^M\big(\mathrm{Law}(X_N),\,\pi\big) \;\le\; \underbrace{\rho(\delta,\lambda)^N\, W_0}_{\text{(I) mixing}} \;+\; \underbrace{\frac{C_1(\lambda)\sqrt{d}\,\delta^{3/2} + \bar{C}_2(\lambda)\delta^2}{1 - \rho(\delta,\lambda)}}_{\text{(II) discretisation}} \;+\; \underbrace{L_g^M\,\sqrt{\lambda/m_M}}_{\text{(III) Moreau}}, \tag{15}$$

where $W_0 = W_2^M(\mu_0, \pi_{\delta,\lambda})$, $\rho(\delta,\lambda) = \max\{|r(m_M)|, |r(L_\lambda)|\}$, $C_1(\lambda) = L_\lambda/\sqrt{2}$, and one may take

$$\bar{C}_2(\lambda) = L_\lambda^2\left(\tfrac{3}{2}\sqrt{10d/m_M} + \tfrac{1}{2}\sqrt{d/L_\lambda}\right).$$

The terms are coupled through ρ and their λ-dependent constants.

Remark 7 (Precision protocol)**.** *At fixed λ, the all-step contraction theorem gives $N = O(\sqrt{\kappa^\lambda_{\mathrm{eff}}}\log(W_0/\varepsilon))$ to the* discrete invariant law *at $\delta^\star_\lambda = 2/\sqrt{m_M L_\lambda}$, where $\kappa^\lambda_{\mathrm{eff}} = L_\lambda/m_M$. This optimum generally exceeds the step range of Theorem 6. Full ε-accuracy to π requires controlling the discretisation and Moreau terms as well; the displayed contraction complexity alone does*

not provide that guarantee. At fixed λ, the two accumulated discretisation terms are $O(\sqrt{\delta})$ and $O(\delta)$ as $\delta \downarrow 0$.

Remark 8 (Gaussian duality and polyhedral tightness)**.** *Quadratic targets attain the contraction bound and have zero stationary bias; at $M = H$ and $\delta = 2$ the update produces an independent exact Gaussian sample. For Term III, the generic $\sqrt{\lambda}$ bound can be conservative: 1-D quantile integration on $f(x) = x^2/2$ with the polyhedral penalties $|x|$ and $\max(2x, -x)$ gives empirical W_2^M-vs-λ slopes 1.95 and 0.97 (Appendix C.9). These examples do not establish a uniform improved rate for all polyhedral penalties.*

4.5 Guarantees for JP-IMLA

Let P denote the canonical P-IMLA$_\lambda$ kernel and Q the exact JP-IMLA kernel, obtained by replacing $\mathrm{prox}^M_{\tau U_\lambda}$ with $\mathrm{prox}^M_{\tau(f+g)}$, $\tau = \delta/2$.

Proposition 9 (JP model and solver perturbations)**.** *Under the hypotheses on f, g of Theorem 6, fix $\delta, \lambda > 0$ and write $\rho = \rho(\delta, \lambda)$. For joint and smoothed proximal points p, q at a common input, and $\mu, \nu \in \mathcal{P}_2$,*

$$\begin{aligned} \varepsilon_{\mathrm{model}} = 2\lambda L_g^M \frac{\tau^{-1} + L_M}{\tau^{-1} + m_M}, \qquad \|p - q\|_M \le \varepsilon_{\mathrm{model}}/2, \\ W_2^M(Q\mu, P\nu) \le \rho W_2^M(\mu, \nu) + \varepsilon_{\mathrm{model}}. \end{aligned} \tag{16}$$

For a finite-ADMM chain $\widehat{X}_j$ satisfying Proposition 16, with fresh noise independent of the past and possibly history-dependent warm starts, let $U_{K,j}$ denote its certificate at the actual step input. Set $e_{K,j} = (8\mathbb{E}U_{K,j}/(m_M + \tau^{-1}))^{1/2}$ and $D_j = W_2^M(\mathcal{L}(\widehat{X}_j), \pi_{\delta,\lambda})$, assuming these quantities are finite. Then

$$\begin{aligned} D_{j+1} &\le \rho D_j + \varepsilon_{\mathrm{model}} + e_{K,j}, \\ D_N &\le \rho^N D_0 + \sum_{j=0}^{N-1} \rho^{N-1-j}(\varepsilon_{\mathrm{model}} + e_{K,j}). \end{aligned} \tag{17}$$

Every invariant law $\nu_Q \in \mathcal{P}_2$ of Q satisfies $W_2^M(\nu_Q, \pi_{\delta,\lambda}) \le \varepsilon_{\mathrm{model}}/(1-\rho)$.

Proof sketch. Monotonicity of ∂g gives the proximal gap; reflection and shared-noise coupling combine it with canonical contraction. The certificate and Minkowski's inequality give the solver recursion; complete proofs are in Appendix B.1.

Adding the discretisation and Moreau terms of Theorem 6 gives the original-target bound (43) when $0 < \delta \le 2/L_\lambda$; the perturbation and solver bounds above hold for every $\delta > 0$. The exact Q has an invariant law, unique in $\mathcal{P}_2$ for quadratic f and finite $g = \sum_e \rho_e |d_e^\top x|$, $\rho_e > 0$, including our imaging model (Theorem 18). This supplies no mixing rate or uniqueness claim for finite ADMM. Appendix B.2 reports the corresponding numerical audits.

5 Experiments

We compare samplers on 256×256 TV deconvolution using geometric, Shepp–Logan and cameraman phantoms. Appendix C reports Gaussian variance and contraction checks, nonquadratic bias stress tests, Moreau-bias measurements and step-size sweeps. A separate 64×64 random fan-beam experiment tests a non-Fourier forward operator (Appendix C.6). Appendix D adds a $d = 256$ bias-coefficient experiment satisfying the potential assumptions of Theorem 5, and separates convergence diagnostics from reconstruction quality and posterior texture. For TV deconvolution we use $M = \hat{H}$, giving $\kappa_{\mathrm{eff}}^\lambda = 1 + 1/\lambda$ for the Moreau-smoothed composite (12).

Setup. A 256×256 phantom under Gaussian blur ($\sigma_{\mathrm{blur}} = 2.0$, $\sigma_{\mathrm{noise}} = 0.05$); f quadratic with $\kappa_f = 4\times10^5$, $g = \rho_{\mathrm{TV}}\|D\cdot\|_1$ [anisotropic TV in the spirit of Rudin et al., 1992] with $\rho_{\mathrm{TV}} = 0.3$, $\lambda = 5\times10^{-3}$, giving $\kappa_{\mathrm{eff}}^\lambda = 201$. We compare four samplers under a fixed $100\times$ iteration allocation: ours $N_{\mathrm{ours}} = 2{,}000$ via Algorithm 1 ($K_{\mathrm{P}} = 30$, $K_{\mathrm{in}} = 50$, $\rho_a = 200$, $\delta = 0.141$); SP-MYULA, MYULA, SK-ROCK [$s = 10$, Pereyra et al., 2020] each $N_{\mathrm{base}} = 2\times10^5$ at largest stable step (App. C.10). Eight chains per sampler start from $X_0 = 0$, with averages accumulated without burn-in,

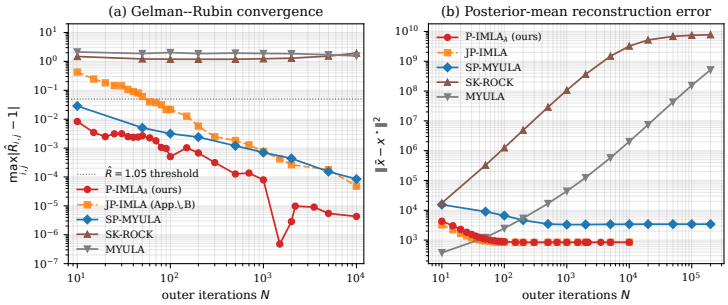

Figure 2: **Classical per-pixel Gelman–Rubin trajectory at 256×256.** Eight chains from the same zero start; (a) $\max_{i,j}|\hat{R}_{i,j}-1|$ on log scale to $N=10{,}000$; (b) extends to $N=2\times10^5$. P-IMLA$_\lambda$ reaches $\hat{R}<1.05$ at $N=10$ and its per-chain SSE changes from 855 to 846 by $N=10{,}000$. SP-MYULA also reaches $\hat{R}\le1.001$, despite larger reconstruction error; MYULA and SK-ROCK end at 2.41 and 11.07. These same-start diagnostics [Gelman and Rubin, 1992] are complemented by overdispersed, rank-normalized split diagnostics in Appendix D.2.

on an NVIDIA RTX PRO 6000 Blackwell (96 GB) with 4-process parallelisation. Actual per-chain wall-clocks are reported in Table 3; the iteration allocation does not make all wall-clocks equal.

Diagnostics and matched-budget comparison. The classical per-pixel $\hat{R}$ in Figure 2 reaches 1.008 at $N=10$ and ≤1.001 at $N=2{,}000$ for P-IMLA$_\lambda$; an $N=10{,}000$ extension (ours and JP-IMLA, on the same hardware) shows stable reconstruction SSE. Because these chains share $X_0=0$, this diagnostic alone does not establish convergence; Theorem 3 bounds Wasserstein contraction, not a threshold-crossing time for $\hat{R}$. Table 3 separately reports reconstruction SSE $=\|\bar{x}-x^\star\|^2$ ($\pm$ SE): P-IMLA$_\lambda$ 855.22 ± 0.66 (per-pixel RMSE ≈0.11), versus 3,432 for SP-MYULA. This approximately fourfold difference is a cold-start, finite-budget result, not a measurement of stationary bias. MYULA / SK-ROCK have SSE $\sim10^9, 10^{10}$ and $\hat{R}=2.41, 11.07$.

Table 3: 256×256 **TV deconvolution, geometric phantom, fixed iteration budgets.** $\rho_{\mathrm{TV}}=0.3$, $\sigma_{\mathrm{noise}}=0.05$, $\lambda=5\times10^{-3}$; RTX PRO 6000 Blackwell, 4-way parallel. SSE is per-chain $\|\bar{x}_c-x^\star\|^2$ averaged across K chains $\pm$ SE, including the initial transient. $\hat{R}$ is classical per-pixel, from common zero starts. These are reconstruction and implementation-budget comparisons; joint-kernel guarantees are stated separately in the appendix. Cross-phantom results appear in Table 4.

Method	SSE at final N	K	N_{final}	Wall-clock/chain	$\hat{R}$ at N_{final}
MYULA	$\sim10^9$ (no mix)	8	2×10^5	43 min	2.41
SK-ROCK ($s=10$)	$\sim10^{10}$ (no mix)	8	2×10^5	425 min	11.07
SP-MYULA	$3{,}432.0\pm1.0$	8	2×10^5	74 min	≤1.001
P-IMLA$_\lambda$ (ours)	$\mathbf{855.22\pm0.66}$	8	**2,000**	**70 min**	$\mathbf{\le1.001}$
JP-IMLA (variant, §3)	$\mathbf{852.72\pm0.65}$	8	**2,000**	**1.2 min**	$\mathbf{\le1.001}$

Cross-phantom robustness (Shepp–Logan, cameraman). We extend the same finite-budget protocol to *Shepp–Logan* and *cameraman*, four chains each. Across all three phantoms ours achieves a 3.5 ×–4.0× reconstruction-SSE reduction over SP-MYULA with the 100× iteration allocation (Table 4). Changing burn-in, initialization or solver accuracy defines a different comparison; these ratios do not isolate invariant-measure error.

Additional bias and roughness checks. On a 16×16 log-cosh lattice satisfying Theorem 5's potential assumptions, all nine coefficient estimates across three preconditioners and three observables agree with the prediction within their 95% intervals. Separate overdispersed imaging runs distinguish high-frequency draw variability controlled by λ from mid-frequency posterior-mean texture reduced by ρ_{TV}; doubling ρ_{TV} improves PSNR by 3.5 dB (geometric) and 2.7 dB (cameraman). Independent exact-target Gibbs experiments at 32^2 and 48^2 retain the same texture without Moreau smoothing or time discretization. Full protocols, observable-specific diagnostics and their remaining slow modes are in Appendix D.

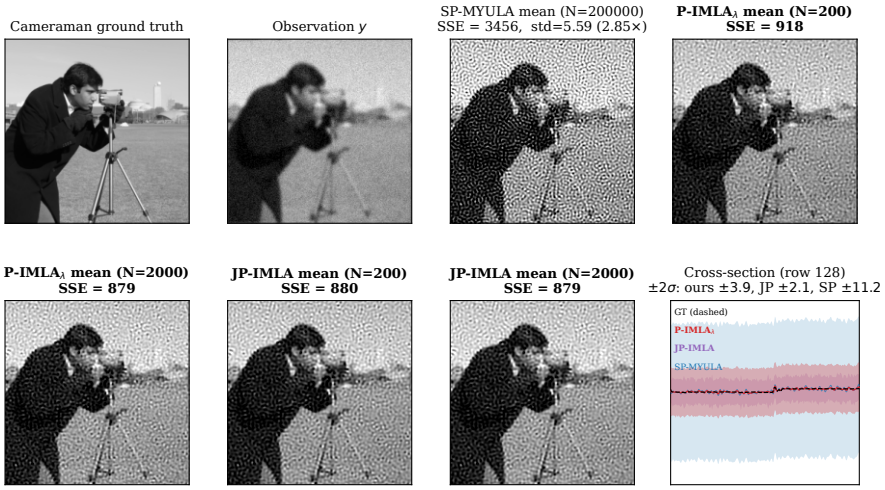

Figure 3: 256×256 **cameraman posterior-mean reconstructions at fixed budgets.** Images are averages across chains, not individual posterior draws. R1: GT, observation, SP-MYULA at $N = 2 \times 10^5$ (R1C3; pooled-mean SSE 3,456, mean pixel std $\bar{s} = 5.59$), and P-IMLA$_\lambda$ at $N = 200$ (R1C4; SSE 918). R2C1: P-IMLA$_\lambda$ at $N = 2{,}000$ (SSE 879, $\bar{s} = 1.96$). R2C2–C3: JP-IMLA (§3) at $N = 200$ (SSE 880) and $N = 2{,}000$ (SSE 879), with the historical $\sim 60\times$ per-iteration implementation speedup. R2C4: row-128 cross-section with ± 2 posterior-standard-deviation bands. Pooled-mean SSE differs from the per-chain average in Table 4. Appendix D.3 shows that stronger TV regularization reduces the means' visible texture.

Table 4: **Cross-phantom matched-budget SSE** (geometric: 8 chains; Shepp–Logan, cameraman: 4 chains; $\pm$ SE; ours $N = 2{,}000$, baselines $N = 2 \times 10^5$). MYULA SSE $\sim 10^9$ on all three.

Phantom	ours SSE	SP-MYULA SSE	SP-MYULA / ours
geometric	855.22 ± 0.66	$3{,}432.0 \pm 1.0$	$4.01\times$
Shepp–Logan	$1{,}013.99 \pm 0.81$	$3{,}525.6 \pm 1.0$	$3.48\times$
cameraman	890.24 ± 1.19	$3{,}460.4 \pm 2.0$	$3.89\times$

6 Conclusion

Implicit midpoint with fixed matrix preconditioning combines Gaussian exactness with the sharp IMLA contraction factor in the effective condition number. The variance-budget identity (2) explains the Gaussian cancellation; the drift-only BEA coefficient identifies the leading weak bias beyond Gaussians under explicit smoothness assumptions. At imaging scale, P-IMLA$_\lambda$ improves finite-budget reconstruction and relaxation at the reported settings, while the diagnostic sweeps separate smoothing, solver and prior effects. The joint-proximal variant retains a substantial per-iteration speed advantage, with model and solver errors controlled separately by the kernel comparison and primal–dual certificate.

Limitations. The BEA expansion does not apply directly to a $C^{1,1}$ TV envelope, and the end-to-end accuracy bound uses a smaller step range than the all-step contraction theorem. Exact-JP uniqueness is proved for the imaging penalty class, not for arbitrary finite ADMM implementations. Small-image exact-posterior checks and overdispersed full-image diagnostics complement the original experiments but do not certify all full-image posterior moments. Position-dependent preconditioning and image-scale non-Fourier validation remain open; Appendix E gives the full scope.

Acknowledgments

We thank the anonymous reviewers for their careful reading and constructive feedback, which substantially improved the paper. This work is supported by Guangdong Province (No. 2023QN10X215), Shenzhen Science and Technology Program (No. JCYJ20241202130548062), the Natural Science Foundation of Shenzhen (No. JCYJ20230807142703006), and the Key Research Platforms and Projects of the Guangdong Provincial Department of Education (No. 2023ZDZX1034). The authors declare no competing interests.

References

Assyr Abdulle, Gilles Vilmart, and Konstantinos C. Zygalakis. High order numerical approximation of the invariant measure of ergodic SDEs. *SIAM Journal on Numerical Analysis*, 52(4):1600–1622, 2014.

Heinz H. Bauschke and Patrick L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. CMS Books in Mathematics. Springer, 2011.

Nawaf Bou-Rabee and J. M. Sanz-Serna. Geometric integrators and the Hamiltonian Monte Carlo method. *Acta Numerica*, 27:113–206, 2018.

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.

Xiang Cheng and Peter L. Bartlett. Convergence of Langevin MCMC in KL-divergence. In *Proceedings of Algorithmic Learning Theory (ALT)*, volume 83 of *PMLR*, pages 186–211, 2018.

Xiang Cheng, Niladri S. Chatterji, Peter L. Bartlett, and Michael I. Jordan. Underdamped Langevin MCMC: a non-asymptotic analysis. In *Proceedings of the Conference on Learning Theory (COLT)*, volume 75 of *PMLR*, pages 300–323, 2018.

Sinho Chewi. Log-concave sampling. Book draft, March 2026 version. Available at `https://chewisinho.github.io/main.pdf`, 2026.

Marie-Caroline Corbineau, Denis Kouamé, Emilie Chouzenoux, Jean-Yves Tourneret, and Jean-Christophe Pesquet. Preconditioned P-ULA for joint deconvolution-segmentation of ultrasound images. *IEEE Signal Processing Letters*, 26(10):1456–1460, 2019.

Simon L. Cotter, Gareth O. Roberts, Andrew M. Stuart, and David White. MCMC methods for functions: modifying old algorithms to make them faster. *Statistical Science*, 28(3):424–446, 2013.

Arnak S. Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B*, 79(3):651–676, 2017.

Alain Durmus and Éric Moulines. High-dimensional Bayesian inference via the unadjusted Langevin algorithm. *Bernoulli*, 25(4A):2854–2882, 2019.

Alain Durmus, Éric Moulines, and Marcelo Pereyra. Efficient Bayesian computation by proximal Markov chain Monte Carlo: when Langevin meets Moreau. *SIAM Journal on Imaging Sciences*, 11(1):473–506, 2018.

Alain Durmus, Éric Moulines, and Marcelo Pereyra. A proximal Markov chain Monte Carlo method for Bayesian inference in imaging inverse problems: when Langevin meets Moreau. *SIAM Review*, 64(4):991–1028, 2022.

Andreas Eberle. Reflection couplings and contraction rates for diffusions. *Probability Theory and Related Fields*, 166:851–886, 2016.

Matthias J. Ehrhardt, Lorenz Kuger, and Carola-Bibiane Schönlieb. Proximal Langevin sampling with inexact proximal mapping. *SIAM Journal on Imaging Sciences*, 17(3):1729–1760, 2024.

Zachary Frangella, Joel A. Tropp, and Madeleine Udell. Randomized Nyström preconditioning. *SIAM Journal on Matrix Analysis and Applications*, 44(2):718–752, 2023.

Andrew Gelman and Donald B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472, 1992.

Mark Girolami and Ben Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B*, 73(2):123–214, 2011.

Andreas Habring, Martin Holler, and Thomas Pock. Subgradient Langevin methods for sampling from nonsmooth potentials. *SIAM Journal on Mathematics of Data Science*, 6:897–925, 2024.

Andreas Habring, Alexander Falk, Martin Zach, and Thomas Pock. Diffusion at absolute zero: Langevin sampling using successive Moreau envelopes. *SIAM Journal on Imaging Sciences*, 19 (1):35–77, 2026.

Max Hird and Samuel Livingstone. Quantifying the effectiveness of linear preconditioning in Markov chain Monte Carlo. *Journal of Machine Learning Research*, 26(119):1–51, 2025.

Liam Hodgkinson, Robert Salomone, and Fred Roosta. Implicit Langevin algorithms for sampling from log-concave densities. *Journal of Machine Learning Research*, 22(136):1–30, 2021.

Matthew D. Hoffman and Andrew Gelman. The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(47):1593–1623, 2014.

Matthew Holden, Marcelo Pereyra, and Konstantinos C. Zygalakis. Bayesian imaging with data-driven priors encoded by neural networks: theory, methods, and algorithms. *SIAM Journal on Imaging Sciences*, 15(2):892–924, 2022.

Teresa Klatzer, Paul Dobson, Yoann Altmann, Marcelo Pereyra, Jesús María Sanz-Serna, and Konstantinos C. Zygalakis. Accelerated Bayesian imaging by relaxed proximal-point Langevin sampling. *SIAM Journal on Imaging Sciences*, 17(2):1078–1117, 2024.

Rémi Laumont, Valentin De Bortoli, Andrés Almansa, Julie Delon, Alain Durmus, and Marcelo Pereyra. Bayesian imaging using plug-and-play priors: when Langevin meets Tweedie. *SIAM Journal on Imaging Sciences*, 15(2):701–737, 2022.

Benedict Leimkuhler, Daniel Paulin, and Peter A. Whalley. Contraction and convergence rates for discretized kinetic Langevin dynamics. *SIAM Journal on Numerical Analysis*, 62(3):1226–1258, 2024.

Radford M. Neal. MCMC using Hamiltonian dynamics. In Steve Brooks, Andrew Gelman, Galin L. Jones, and Xiao-Li Meng, editors, *Handbook of Markov Chain Monte Carlo*, pages 113–162. Chapman & Hall/CRC, 2011.

Marcelo Pereyra. Proximal Markov chain Monte Carlo algorithms. *Statistics and Computing*, 26(4): 745–760, 2016.

Marcelo Pereyra, Luis A. Vargas Mieles, and Konstantinos C. Zygalakis. Accelerating proximal Markov chain Monte Carlo by using an explicit stabilized method. *SIAM Journal on Imaging Sciences*, 13(2):905–935, 2020.

Marien Renaud, Valentin De Bortoli, Arthur Leclaire, and Nicolas Papadakis. From stability of Langevin diffusion to convergence of proximal MCMC for non-log-concave sampling. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Gareth O. Roberts and Richard L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996.

Leonid I. Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1–4):259–268, 1992.

Andrew M. Stuart. Inverse problems: a Bayesian perspective. *Acta Numerica*, 19:451–559, 2010.

Michel Talagrand. Transportation cost for Gaussian and other product measures. *Geometric and Functional Analysis*, 6(3):587–600, 1996.

Denis Talay and Luciano Tubaro. Expansion of the global error for numerical schemes solving stochastic differential equations. *Stochastic Analysis and Applications*, 8(4):483–509, 1990.

Michalis K. Titsias. Optimal preconditioning and Fisher adaptive Langevin sampling. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

Aki Vehtari, Andrew Gelman, Daniel Simpson, Bob Carpenter, and Paul-Christian Bürkner. Rank-normalization, folding, and localization: an improved $\widehat{R}$ for assessing convergence of MCMC. *Bayesian Analysis*, 16(2):667–718, 2021.

Santosh S. Vempala and Andre Wibisono. Rapid convergence of the unadjusted Langevin algorithm: isoperimetry suffices. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Andre Wibisono. Sampling as optimization in the space of measures: the Langevin dynamics as a composite optimization problem. In *Proceedings of the Conference on Learning Theory (COLT)*, volume 75 of *PMLR*, pages 2093–3027, 2018.

A Full Proofs from Section 4

This appendix provides the full proofs of Theorems 2–6 and the supporting technical lemmas referenced in Section 4. The proofs are self-contained given the definitions and labeled equations of Sections 2–4; no additional notation is introduced beyond the specifications below.

Standing conventions for the appendix. We write $H = \nabla^2 f$, $A = M^{-1/2} H M^{-1/2}$, $\Gamma = M^{-1} H M^{-1}$, and $b = -M^{-1}\nabla f$; these are functions of the current point unless f is quadratic. We use the Einstein summation convention on repeated indices and write $M^{-1}_{ij} = (M^{-1})_{ij}$. The spectral variable is ℓ; λ denotes Moreau smoothing.

A.1 Proof of Theorem 2 (Gaussian exactness)

Recall the setting: $f(x) = \frac{1}{2} x^\top H x - h^\top x$ with $H \succ 0$, $g \equiv 0$, $\pi = \mathcal{N}(\mu, H^{-1})$ with $\mu = H^{-1}h$, and $M \succ 0$. P-IMLA reads

$$X_{n+1} = -X_n + 2\,\mathrm{prox}^M_{(\delta/2)f}\Big(X_n + \tfrac{\sqrt{2\delta}}{2}\, M^{-1/2}\xi_n\Big), \qquad \xi_n \overset{\text{iid}}{\sim} \mathcal{N}(0, I). \tag{18}$$

Step 1 (Joint diagonalization and whitening). Let $A = M^{-1/2} H M^{-1/2} = Q \Lambda Q^\top$ be the spectral decomposition with $\Lambda = \mathrm{diag}(\ell_1, \dots, \ell_d)$ and Q orthogonal. Introduce whitened coordinates

$$Z_n \;\coloneqq\; Q^\top M^{1/2}(X_n - \mu).$$

Under this change of variables: the M-norm becomes Euclidean, $\|X - \mu\|_M = \|Z\|$; the target becomes $\pi_Z = \bigotimes_{i=1}^d \mathcal{N}(0, \ell_i^{-1})$; and W_2^M in X-coordinates equals Euclidean W_2 in Z-coordinates.

We claim that (18) maps to d independent scalar recurrences

$$Z_{n+1,i} \;=\; r_i\, Z_{n,i} \;+\; \frac{\sqrt{2\delta}}{1 + a_i}\,\eta_{n,i}, \qquad r_i \coloneqq \tfrac{1-a_i}{1+a_i},\;\; a_i \coloneqq \tfrac{\delta}{2}\ell_i,\;\; \eta_{n,i} \sim \mathcal{N}(0,1). \tag{19}$$

Indeed, the proximal of the quadratic f at y satisfies the first-order condition $Hp - h + \frac{2}{\delta} M(p - y) = 0$, which gives

$$p \;=\; \big(\tfrac{\delta}{2}H + M\big)^{-1}\big(\tfrac{\delta}{2}(H\mu + (h - H\mu)) + My\big) \;=\; \big(\tfrac{\delta}{2}H + M\big)^{-1}\big(\tfrac{\delta}{2}H\mu + My\big),$$

using $h = H\mu$. Subtracting $\mu = (\frac{\delta}{2}H + M)^{-1}(\frac{\delta}{2}H + M)\mu$ gives $p - \mu = (\frac{\delta}{2}H + M)^{-1} M(y - \mu)$. Applying $X_{n+1} - \mu = -(X_n - \mu) + 2(p - \mu)$ with $y = X_n + \frac{\sqrt{2\delta}}{2} M^{-1/2}\xi_n$ yields

$$X_{n+1} - \mu \;=\; \big(-I + 2(\tfrac{\delta}{2}H + M)^{-1}M\big)\,(X_n - \mu) \;+\; \sqrt{2\delta}\,(\tfrac{\delta}{2}H + M)^{-1} M^{1/2}\xi_n.$$

Whitening gives $Z_{n+1} = R\, Z_n + \sqrt{2\delta}\, S\, \eta_n$ where $\eta_n \coloneqq Q^\top \xi_n \sim \mathcal{N}(0, I)$, $R = (I + \frac{\delta}{2}\Lambda)^{-1}(I - \frac{\delta}{2}\Lambda)$, and $S = (I + \frac{\delta}{2}\Lambda)^{-1}$. Since R, S are diagonal with entries $r_i, 1/(1 + a_i)$, the recurrence (19) follows coordinate-wise.

Step 2 (Zero bias via variance budget). The stationary variance of the Cayley AR(1) (19) is

$$\mathrm{Var}(Z_{\infty,i}) \;=\; \frac{2\delta/(1 + a_i)^2}{1 - r_i^2}.$$

Since $1 - r_i^2 = 1 - \big(\frac{1-a_i}{1+a_i}\big)^2 = \frac{(1+a_i)^2 - (1-a_i)^2}{(1+a_i)^2} = \frac{4a_i}{(1+a_i)^2}$, the Cayley denominators cancel:

$$\mathrm{Var}(Z_{\infty,i}) \;=\; \frac{2\delta/(1 + a_i)^2}{4a_i/(1 + a_i)^2} \;=\; \frac{2\delta}{4 \cdot (\delta\ell_i/2)} \;=\; \frac{1}{\ell_i} \quad \text{for every } \delta > 0.$$

This is the variance-budget identity (2). Combined with $\mathbb{E}[Z_{\infty,i}] = 0$ (from $|r_i| < 1$) and the independence of components, $Z_\infty \sim \pi_Z$, hence $X_\infty \sim \pi$. The synchronous contraction below gives uniqueness, as in Appendix A.2. This establishes Gaussian exactness.

Step 3 (Wasserstein contraction via synchronous coupling). For two chains driven by the same noise $\{\xi_n\}$, whitening preserves the shared-noise structure and (19) gives

$$Z_{n+1,i} - Z'_{n+1,i} \;=\; r_i\,(Z_{n,i} - Z'_{n,i}) \qquad \text{(deterministically, for each } i\text{)}.$$

Iterating, $\|Z_N - Z'_N\| \le \big(\max_i |r_i|\big)^N \|Z_0 - Z'_0\|$, i.e.,

$$\|X_N - X'_N\|_M \;\le\; \rho(\delta)^N \|X_0 - X'_0\|_M, \qquad \rho(\delta) = \max_i |r_i| = \max\big(|r(m_M)|, |r(L_M)|\big),$$

since $r'(\ell) = -\delta/(1+\delta\ell/2)^2 < 0$, so $|r|$ attains its maximum on $[m_M, L_M]$ at an endpoint. Taking an optimal W_2^M-coupling at time 0 and using synchronous noise proves the Wasserstein contraction bound.

Step 4 (Minimax optimum and mixing complexity). For $\delta \in [2/L_M, 2/m_M]$, the endpoint magnitudes are $r(m_M) = (1-\delta m_M/2)/(1+\delta m_M/2)$ and $|r(L_M)| = (\delta L_M/2-1)/(\delta L_M/2+1)$, respectively decreasing and increasing in δ. Outside this interval, moving δ toward it decreases both magnitudes. Equating $r(m_M) = |r(L_M)|$ yields

$$\tfrac{1-\delta m_M/2}{1+\delta m_M/2} \;=\; \tfrac{\delta L_M/2-1}{\delta L_M/2+1} \iff \delta^\star \;=\; \tfrac{2}{\sqrt{m_M L_M}},$$

and substituting $\delta^\star$ gives the common value

$$\rho^\star \;=\; \tfrac{1-\sqrt{m_M/L_M}}{1+\sqrt{m_M/L_M}} \;=\; \tfrac{\sqrt{\kappa_{\text{eff}}}-1}{\sqrt{\kappa_{\text{eff}}}+1},$$

establishing the optimal step and factor. For mixing: solving $\rho^{\star N} \le \varepsilon$ gives $N \ge \log(1/\varepsilon)/\log(1/\rho^\star) = O(\sqrt{\kappa_{\text{eff}}}\log(1/\varepsilon))$ using $\log(1/\rho^\star) = 2/\sqrt{\kappa_{\text{eff}}} + O(\kappa_{\text{eff}}^{-3/2})$. This proves the stated mixing complexity. □

A.2 Proof of Theorem 3 (reflected-proximal contraction)

Write $\tau = \delta/2$ and $\widetilde{f}(z) = f(M^{-1/2}z)$. Whitening turns the M-norm into the Euclidean norm and the preconditioned update into IMLA for $\widetilde{f}$. Thus it suffices to prove the result for $M = I$, with strong-convexity and smoothness constants m_M, L_M.

Let $p = \operatorname{prox}_{\tau\widetilde{f}}(w)$ and $q = \operatorname{prox}_{\tau\widetilde{f}}(v)$. There is a symmetric matrix $\bar{H}$ satisfying

$$m_M I \preceq \bar{H} \preceq L_M I, \qquad \nabla\widetilde{f}(p) - \nabla\widetilde{f}(q) = \bar{H}(p-q).$$

For a C^2 potential, take the Hessian integrated along the segment from q to p. For $C^{1,1}$, mollify $\widetilde{f}$: the mollifications retain the same Hessian bounds, their gradients converge locally uniformly, and a convergent subsequence of the bounded symmetric secant matrices gives the displayed identity. The proximal optimality conditions now give

$$w - v = (I + \tau\bar{H})(p-q).$$

Consequently, for $R_\tau = 2\operatorname{prox}_{\tau\widetilde{f}} - I$,

$$R_\tau(w) - R_\tau(v) = (I - \tau\bar{H})(I+\tau\bar{H})^{-1}(w-v). \tag{20}$$

The eigenvalues of the matrix on the right are $(1-\tau\ell)/(1+\tau\ell)$ with $\ell \in [m_M, L_M]$. This scalar function is decreasing, so

$$\|R_\tau(w) - R_\tau(v)\| \le \max_{\ell\in\{m_M, L_M\}} \left|\frac{1-\tau\ell}{1+\tau\ell}\right| \|w-v\| = \rho(\delta)\|w-v\|.$$

For shared noise, $w = X_n + \sqrt{\delta/2}\,\xi_n$ and $v = Y_n + \sqrt{\delta/2}\,\xi_n$; the extra noise translation in the reflected-proximal representation cancels. Whitening back proves the pathwise inequality. A quadratic potential with an endpoint eigenvalue attains this factor. Attainment does not require that every potential attaining it be quadratic.

The one-step map sends $\mathcal{P}_2$ into itself, since the proximal map is Lipschitz and the noise has a finite second moment. Applying the pathwise inequality to an optimal coupling gives

$$W_2^M(\mu P_\delta, \nu P_\delta) \le \rho(\delta) W_2^M(\mu,\nu).$$

Here $\rho(\delta) < 1$ for every $\delta > 0$. The space $(\mathcal{P}_2, W_2^M)$ is complete, so Banach's fixed-point theorem gives a unique invariant law $\pi_\delta \in \mathcal{P}_2$ and the stated iterated bound. This is also the only invariant probability law: convergence from each deterministic initial state, tested against bounded Lipschitz functions and integrated against any invariant probability law, gives equality with π_δ by dominated convergence. This proves Theorem 3.

A.3 Proof of Theorem 5 and its refinements

Throughout this subsection M is fixed, $g = 0$, $C = M^{-1}$, $b = -C\nabla f$, $h = \operatorname{tr}(C\nabla^2 f)$ and $\mathcal{L}u = b \cdot \nabla u + C : \nabla^2 u$. We write C^k_{pol} for functions whose derivatives through order k, including the function itself, have at most polynomial growth. A remainder $O_{V_q}(\delta^a)$ is bounded in absolute value by $K\delta^a V_q(x)$, with $V_q(x) = 1 + \|x - x_\star\|_M^q$, uniformly for sufficiently small δ; o_{V_q} has the analogous meaning. Constants may depend on the potential and the test function, but not on δ.

Uniform moments. The following lemma supplies the integrability needed for the weighted expansions; it requires only the assumptions of the contraction theorem.

Lemma 10 (δ-uniform invariant moments)**.** *Let $V \in C^{1,1}$ be globally m_M-strongly convex and L_V-smooth in the M-norm, and let ν_δ be the invariant law of its exact midpoint kernel. If $0 < \delta \le 2/L_V$ and $x_\star$ minimises V, then*

$$\mathbb{E}_{\nu_\delta}\|X - x_\star\|_M^2 \le \frac{10d}{m_M}, \qquad (\mathbb{E}_{\nu_\delta}\|X - x_\star\|_M^p)^{1/p} \le K\sqrt{\frac{d+p}{m_M}} \quad (p \ge 1), \tag{21}$$

where K is a numerical constant.

Proof. Whiten and translate so that $M = I$ and $x_\star = 0$, and retain the notation V. Put $\tau = \delta/2$, $\sigma = \sqrt{\tau}$, take $X \sim \nu_\delta$ independently of $\xi \sim \mathcal{N}(0, I)$, and set

$$p = \operatorname{prox}_{\tau V}(X + \sigma\xi), \qquad X' = -X + 2p.$$

The invariant law has a finite second moment by Theorem 3. Since $p = (X + X')/2$ and $X' - X = -\delta\nabla V(p) + \sqrt{2\delta}\,\xi$, stationarity gives

$$0 = -2\delta\,\mathbb{E}\langle\nabla V(p), p\rangle + 2\sqrt{2\delta}\,\mathbb{E}\langle\xi, p\rangle.$$

Conditioning on X, Gaussian integration by parts and the a.e. Jacobian of the Lipschitz proximal map imply

$$\mathbb{E}\langle\xi, p\rangle = \sigma\,\mathbb{E}\operatorname{tr}(I + \tau\nabla^2 V(p))^{-1} \le \sigma d.$$

For $C^{1,1}$ potentials this identity follows either by the weak derivative version of Gaussian integration by parts or by mollification. Thus $m_M\mathbb{E}\|p\|^2 \le \mathbb{E}\langle\nabla V(p), p\rangle \le d$. The optimality condition also gives $X = p + \tau\nabla V(p) - \sigma\xi$, hence

$$\mathbb{E}\|X\|^2 \le 2(1 + \tau L_V)^2\, d/m_M + 2\tau d \le 10d/m_M,$$

using $\tau L_V \le 1$ and $m_M \le L_V$.

For higher moments, the map $\Phi(x, \xi) = -x + 2\operatorname{prox}_{\tau V}(x + \sigma\xi)$ has state Lipschitz factor $\rho = (1 - \tau m_M)/(1 + \tau m_M)$ in this step range, and noise Lipschitz factor at most $2\sigma/(1 + \tau m_M)$. Construct its stationary solution by starting at 0 in the remote past and using independent Gaussian noises. Contraction makes these finite-past solutions converge in L^2 to the unique invariant law. The influence of the noise k steps in the past is at most

$$a_k = \frac{2\sqrt{\tau}}{1 + \tau m_M}\rho^k, \qquad \sum_{k=0}^{\infty} a_k^2 = \frac{4\tau}{(1 + \tau m_M)^2(1 - \rho^2)} = \frac{1}{m_M}.$$

Each finite-past map is therefore $1/\sqrt{m_M}$-Lipschitz as a function of its Gaussian inputs. Gaussian concentration, followed by the L^2 limit, gives the same concentration bound for $\|X\|$ about its mean. Integrating its Gaussian tail yields $(\mathbb{E}|\|X\| - \mathbb{E}\|X\||^p)^{1/p} \le K_0\sqrt{p/m_M}$. Combining this with $\mathbb{E}\|X\| \le \sqrt{10d/m_M}$ proves the second assertion. □

Step 1 (Finite-order implicit expansion). Assume the hypotheses of Theorem 5. Write $\varepsilon = \sqrt{\delta}$, $\eta = \sqrt{2}M^{-1/2}\xi$ and $\Delta = X_1 - x$. The implicit equation is

$$\Delta = \varepsilon\eta + \varepsilon^2 b(x + \Delta/2).$$

For $\delta L_M \le 1$, the global Lipschitz bound on b gives

$$\|\Delta\|_M \le 2\varepsilon\|\eta\|_M + 2\varepsilon^2 L_M\|x - x_\star\|_M. \tag{22}$$

All derivatives of b needed below have polynomial growth. Define

$$A_4 = \varepsilon\eta + \varepsilon^2 b + \frac{\varepsilon^3}{2} Db\,\eta + \varepsilon^4 \left(\frac{1}{2} Db\,b + \frac{1}{8} D^2 b[\eta,\eta] \right), \tag{23}$$

where the coefficients are evaluated at x. To bound the remainder, set $G(v) = v - \varepsilon\eta - \varepsilon^2 b(x + v/2)$. Taylor's formula with integral remainder, applied to $b(x + A_4/2)$ through order two, cancels all powers through ε^4 and gives

$$\|G(A_4)\|_M \le K\varepsilon^5 (1 + \|x - x_\star\|_M^q + \|\eta\|_M^q)$$

for some finite q. The estimate is global: coefficients of A_4 are polynomially bounded, and the integral remainder is bounded by the corresponding polynomial weight along the segment. Moreover, $D_v G = I + (\varepsilon^2/2) M^{-1} \nabla^2 f(x + v/2)$ is at least 1-strongly monotone in the M-inner product. Since $G(\Delta) = 0$, Cauchy–Schwarz applied to its monotonicity inequality gives

$$\|\Delta - A_4\|_M \le K\varepsilon^5 (1 + \|x - x_\star\|_M^q + \|\eta\|_M^q). \tag{24}$$

No analyticity of the proximal map or symmetry of the full increment is assumed.

Step 2 (Conditional moments and the local weak operator). The Gaussian identities $\mathbb{E}\eta = 0$ and $\mathbb{E}\eta_i\eta_j = 2C_{ij}$, together with (23)–(24), give

$$\mathbb{E}[\Delta_i \mid x] = \delta b_i + \delta^2 \left(\tfrac{1}{2}(Db\,b)_i - \tfrac{1}{4}(C\nabla h)_i \right) + O_{V_q}(\delta^{5/2}), \tag{25}$$

$$\mathbb{E}[\Delta_i \Delta_j \mid x] = 2\delta C_{ij} + \delta^2 (b_i b_j - 2\Gamma_{ij}) + O_{V_q}(\delta^{5/2}), \tag{26}$$

$$\mathbb{E}[\Delta_i \Delta_j \Delta_k \mid x] = 2\delta^2 (b_i C_{jk} + b_j C_{ik} + b_k C_{ij}) + O_{V_q}(\delta^{5/2}), \tag{27}$$

$$\mathbb{E}[\Delta_i \Delta_j \Delta_k \Delta_l \mid x] = 4\delta^2 (C_{ij}C_{kl} + C_{ik}C_{jl} + C_{il}C_{jk}) + O_{V_q}(\delta^{5/2}), \tag{28}$$

where $\Gamma = C\nabla^2 f\, C$. For example, $\mathbb{E} D^2 b[\eta,\eta] = 2C : \nabla^2 b = -2C\nabla h$; the second-moment cross terms use $Db\,C = C(Db)^\top = -\Gamma$. Odd Gaussian monomials vanish after expectation, term by term.

Taylor expansion of $u(x + \Delta)$ through order four, with its fifth-order remainder bounded using (22), is valid for $u \in C^6_{\mathrm{pol}}$. Substituting these moments yields

$$P_\delta u = u + \delta \mathcal{L} u + \delta^2 A_2 u + O_{V_q}(\delta^{5/2}), \qquad A_2 = \tfrac{1}{2}\mathcal{L}^2 + B_2, \qquad B_2 u = \tfrac{1}{4}(C\nabla h)\cdot\nabla u. \tag{29}$$

For completeness, the four derivative orders in $A_2 u$ are

$$\left(\tfrac{1}{2} Db\,b - \tfrac{1}{4} C\nabla h\right)\cdot\nabla u + \left(\tfrac{1}{2} bb^\top - \Gamma\right) : \nabla^2 u + b_i C_{jk}\partial_{ijk} u + \tfrac{1}{2} C_{ij} C_{kl} \partial_{ijkl} u.$$

Direct differentiation of $\mathcal{L}u$ gives the same expression for $\mathcal{L}^2 u/2$ except that its first-derivative coefficient is $Db\,b/2 - C\nabla h/2$. This verifies the displayed B_2. The second-moment cancellation can equivalently be read from the proximal Jacobian:

$$SCS^\top = C - \delta\Gamma + O(\delta^2), \qquad S = (I + (\delta/2) C\nabla^2 f)^{-1}. \tag{30}$$

It does not require M and $\nabla^2 f$ to commute.

Step 3 (Stationarity and the Poisson equation). The one-step estimate and invariant perturbation argument proved independently in Appendix A.5, with $V = f$, imply $W_2^M(\pi_\delta, \pi) = O(\sqrt{\delta})$. Thus $\pi_\delta \Rightarrow \pi$. Lemma 10 upgrades this to convergence of expectations of continuous polynomially bounded functions, by truncation and uniform integrability. Apply (29) to the assumed Poisson solution u_φ and integrate against π_δ. Stationarity and the uniform moments give

$$\pi_\delta(\varphi) - \pi(\varphi) = \delta\, \pi_\delta(A_2 u_\varphi) + O(\delta^{3/2}) = \delta\, \pi(A_2 u_\varphi) + o(\delta).$$

Strong log-concavity gives Gaussian tails for π; polynomial derivative bounds allow integration by parts with smooth cutoffs and passage to the limit. In particular $\pi(\mathcal{L}^2 u_\varphi) = 0$ and

$$\pi\big((C\nabla h)\cdot\nabla u_\varphi\big) = -\pi(h\,\mathcal{L} u_\varphi) = \pi\big(h(\varphi - \pi(\varphi))\big).$$

Therefore $\pi(A_2 u_\varphi) = \mathrm{Cov}_\pi(h,\varphi)/4$, proving Theorem 5. This argument uses only zeroth-order convergence of π_δ; it never assumes a first-order invariant-density expansion.

Second-order weak expansion and Richardson extrapolation. We record the stronger assumptions needed for an actual second-order guarantee. Assume now that $f \in C^8$, that $D^j f$ has polynomial growth for $3 \le j \le 8$, and that the same global strong-convexity and smoothness bounds hold. The finite-order expansion above extends to $A_6 = \sum_{k=1}^{6} \varepsilon^k a_k$, where

$$\begin{aligned}
a_1 &= \eta, \qquad a_2 = b, \qquad a_3 = \tfrac{1}{2} Db\,\eta, \qquad a_4 = \tfrac{1}{2} Db\,b + \tfrac{1}{8} D^2 b[\eta,\eta], \\
a_5 &= \tfrac{1}{4}(Db)^2\eta + \tfrac{1}{4} D^2 b[\eta, b] + \tfrac{1}{48} D^3 b[\eta,\eta,\eta], \\
a_6 &= \tfrac{1}{4}(Db)^2 b + \tfrac{1}{16} Db\, D^2 b[\eta,\eta] + \tfrac{1}{8} D^2 b[\eta, Db\,\eta] + \tfrac{1}{8} D^2 b[b,b] \\
&\quad + \tfrac{1}{16} D^3 b[\eta,\eta,b] + \tfrac{1}{384} D^4 b[\eta,\eta,\eta,\eta].
\end{aligned}$$

Substitution into G cancels powers through ε^6. Taylor's integral remainder and the same strong-monotonicity argument give

$$\|\Delta - A_6\|_M \le K\varepsilon^7 (1 + \|x - x_\star\|_M^q + \|\eta\|_M^q).$$

For $u \in C^8_{\mathrm{pol}}$, its sixth-order Taylor expansion consequently yields

$$P_\delta u = u + \delta \mathcal{L} u + \delta^2 A_2 u + \delta^3 A_3 u + O_{V_q}(\delta^{7/2}). \tag{31}$$

Here the differential operator A_3, of order at most six, is defined explicitly by the finite sum

$$A_3 u = \mathbb{E}_\eta \left[\sum_{j=1}^{6} \frac{1}{j!} \sum_{\substack{k_1 + \cdots + k_j = 6 \\ k_i \ge 1}} D^j u(x)[a_{k_1}, \ldots, a_{k_j}] \right].$$

Its coefficients have polynomial growth. Every coefficient of an odd power of ε is odd in η, as is seen from the displayed a_k; its Gaussian expectation is zero. This accounts in particular for the absence of a $\delta^{5/2}$ term in (31).

Theorem 11 (Second-order weak invariant expansion)**.** *Under the C^8 assumptions above, suppose $\mathcal{L} u_\varphi = -(\varphi - \pi(\varphi))$ has a solution $u_\varphi \in C^8_{\mathrm{pol}}$. Set $\psi = A_2 u_\varphi$ and $\bar\psi = \psi - \pi(\psi)$, and suppose $\mathcal{L} u_{\bar\psi} = -\bar\psi$ has a solution $u_{\bar\psi} \in C^6_{\mathrm{pol}}$. Then*

$$\pi_\delta(\varphi) - \pi(\varphi) = \delta c_1(\varphi) + \delta^2 c_2(\varphi) + o(\delta^2),$$

where $c_1(\varphi) = \mathrm{Cov}_\pi(h, \varphi)/4$ and $c_2(\varphi) = \pi(B_2 u_{\bar\psi}) + \pi(A_3 u_\varphi)$.

Proof. Integrate (31) against π_δ, use stationarity and the Poisson equation, and divide by δ:

$$\pi_\delta(\varphi) - \pi(\varphi) = \delta\, \pi_\delta(\psi) + \delta^2 \pi_\delta(A_3 u_\varphi) + o(\delta^2).$$

The uniform moments and $\pi_\delta \Rightarrow \pi$ give $\pi_\delta(A_3 u_\varphi) = \pi(A_3 u_\varphi) + o(1)$. Apply the already proved first-order theorem to the observable ψ, using its assumed Poisson solution: $\pi_\delta(\psi) = \pi(\psi) + \delta\pi(B_2 u_{\bar\psi}) + o(\delta)$. Finally $\pi(\psi) = \pi(A_2 u_\varphi) = c_1(\varphi)$. Substitution proves the result without a density expansion. □

Corollary 12 (Richardson extrapolation)**.** *Under Theorem 5, $2\pi_{\delta/2}(\varphi) - \pi_\delta(\varphi) = \pi(\varphi) + o(\delta)$. Under the stronger hypotheses of Theorem 11,*

$$2\pi_{\delta/2}(\varphi) - \pi_\delta(\varphi) - \pi(\varphi) = -\tfrac{1}{2} c_2(\varphi)\delta^2 + o(\delta^2).$$

Proof. Insert the corresponding invariant expansion at δ and $\delta/2$. The terms linear in δ cancel; the quadratic coefficient is $2(\delta/2)^2 - \delta^2 = -\delta^2/2$. □

These are invariant-expectation statements, not finite-sample MSE bounds. The weights $(2, -1)$ can increase Monte Carlo variance; coupling coarse and fine chains with nested Brownian increments and reporting cost-normalised MSE addresses that separate issue. The combination is not a probability distribution. The $C^{1,1}$ TV Moreau potential does not automatically satisfy the stronger regularity assumptions.

Deterministic numerical check of Richardson extrapolation. We test Corollary 12 on the smooth one-dimensional target $f(x) = x^2/2 + 0.7 \log\cosh x$, with $M = 1$. Here $1 \le f'' \le 1.7$ and all derivatives of order two and higher are bounded, so the potential satisfies the strengthened regularity assumptions. For $\varphi \in \{h = f'', x^2, \log\cosh x\}$, we compute the midpoint invariant law deterministically: 128-node Gauss–Hermite quadrature integrates the injected Gaussian noise, four-point cubic Lagrange deposition represents the transition operator on 3,201 points in $[-8, 8]$, and power iteration stops at an L^1 change below 10^{-14}. Proximal Newton solves use tolerance 10^{-15}. Target expectations use grid quadrature; the theoretical coefficient $c_1(\varphi) = \mathrm{Cov}_\pi(h, \varphi)/4$ is evaluated independently on 200,001 points in $[-10, 10]$.

Let $b_\varphi(\delta) = \pi_\delta(\varphi) - \pi(\varphi)$. We fit $b_\varphi(\delta)$ to a cubic polynomial without an intercept using seven step sizes from 0.04 to 0.32. Table 5 reports the fitted leading coefficient and the extrapolated residuals on three exact halving pairs. The Gaussian control, obtained by setting the $\log\cosh$ coefficient to zero, has maximum numerical bias 3.1×10^{-11} at $\delta \in \{0.04, 0.32\}$. Thus the tabulated residuals are well above the calibrated numerical error; there is no Monte Carlo sampling error.

Table 5: **Richardson check on a smooth nonquadratic target.** Residual columns are $|2b_\varphi(\delta/2) - b_\varphi(\delta)|$. The last column gives the two successive base-two log slopes; Corollary 12 predicts asymptotic order two.

φ	$\widehat{c}_1/c_1$	$\delta = 0.32$	$\delta = 0.16$	$\delta = 0.08$	Slopes
h	1.0030	3.16×10^{-4}	8.18×10^{-5}	2.01×10^{-5}	1.95, 2.02
x^2	1.0001	7.46×10^{-4}	1.96×10^{-4}	5.00×10^{-5}	1.93, 1.97
$\log\cosh x$	1.0007	3.35×10^{-4}	8.68×10^{-5}	2.18×10^{-5}	1.95, 1.99

The leading coefficient agrees within 0.3%, and the extrapolated slopes 1.93–2.02 support the second-order invariant-bias prediction. The verification script is `rebuttal/verify_items_7_8.py`, Part 7.

Conditional density, KL and Wasserstein consequence. The weak expansions above do not imply the following density condition. Let $c(x) = \frac{1}{4}(h(x) - \pi(h))$ and assume

$$\pi_\delta \ll \pi, \qquad \left\| \frac{d\pi_\delta}{d\pi} - 1 - \delta c \right\|_{L^3(\pi)} = o(\delta). \tag{32}$$

The function c belongs to $L^3(\pi)$ under our assumptions.

Proposition 13 (Conditional KL and W_2^M coefficient)**.** *If* (32) *holds, then*

$$\mathrm{KL}(\pi_\delta \| \pi) = \tfrac{\delta^2}{32} \mathrm{Var}_\pi(h) + o(\delta^2), \qquad W_2^M(\pi_\delta, \pi) \le \frac{\delta}{4\sqrt{m_M}} \mathrm{Std}_\pi(h) + o(\delta).$$

Proof. Write $s = d\pi_\delta/d\pi - 1$, so $s \ge -1$ and $\pi(s) = 0$. For $\Phi(s) = (1+s)\log(1+s) - s$, with its continuous value at $s = -1$, there is a numerical constant K such that $|\Phi(s) - s^2/2| \le K|s|^3$ for all $s \ge -1$. On $|s| \le 1/2$ this follows from Taylor's integral remainder. On $s > 1/2$ use $0 \le \Phi(s) \le s^2$, and on $[-1, -1/2]$ use boundedness of Φ and $|s|^3 \ge 1/8$. Thus

$$\left| \mathrm{KL}(\pi_\delta \| \pi) - \tfrac{1}{2}\|s\|^2_{L^2(\pi)} \right| \le K \|s\|^3_{L^3(\pi)}.$$

Condition (32) gives $\|s\|_3 = O(\delta)$ and $\|s\|_2^2 = \delta^2 \|c\|_2^2 + o(\delta^2)$, proving the KL expansion. Whitening and the Talagrand inequality for an m_M-strongly log-concave measure [Talagrand, 1996] give $(W_2^M)^2 \le 2\,\mathrm{KL}/m_M$ and hence the stated bound. □

Condition (32) is an additional hypothesis; neither the moment lemma nor the weak observable expansion proves it.

A.4 P-ULA backward error analysis

For comparison, P-ULA has increment $\Delta = \delta b + \sqrt{2\delta}\, M^{-1/2}\xi$. Its first two conditional moments are

$$\mathbb{E}\Delta_i = \delta b_i, \qquad \mathbb{E}\Delta_i\Delta_j = 2\delta C_{ij} + \delta^2 b_i b_j;$$

its third and fourth moments agree with (27)–(28) at order δ^2. Taylor expansion and comparison with $\mathcal{L}^2/2$ therefore yield

$$B_2^{\mathrm{ULA}} u = -\tfrac{1}{2}(Db\, b)\cdot\nabla u + \tfrac{1}{2}(C\nabla h)\cdot\nabla u + \Gamma_{ij}\partial_{ij} u. \tag{33}$$

The last term is a second-order differential operator, absent from the midpoint defect. This is an operator-level distinction: both methods generally have first-order weak invariant bias in δ. On Gaussian targets the P-ULA stationary variance is inflated by $(1-\delta\ell/2)^{-1}$ in eigendirection ℓ, as proved in Proposition 1; midpoint preserves the variance exactly. This comparison does not assert that preconditioning requires a particular integrator.

A.5 Proof of Theorem 6 (end-to-end bound)

Part I (The smoothed kernel). For convex g, the metric Moreau envelope satisfies $\nabla g^{\lambda,M} = \lambda^{-1}M(I - \mathrm{prox}_{\lambda g}^{M})$. Firm nonexpansiveness of the proximal map in the M-Hilbert space gives

$$\|\nabla g^{\lambda,M}(x) - \nabla g^{\lambda,M}(y)\|_{M^{-1}} \le \lambda^{-1}\|x-y\|_M.$$

Thus $U_\lambda = f + g^{\lambda,M}$ is $C^{1,1}$, m_M-strongly convex and $L_\lambda = L_M + 1/\lambda$-smooth in the M-norm. Theorem 3 gives its exact midpoint kernel $P_{\delta,\lambda}$ a unique invariant law $\pi_{\delta,\lambda}$ and contraction factor $\rho(\delta,\lambda) = \max\{|r(m_M)|, |r(L_\lambda)|\}$ for all $\delta > 0$.

Part II (A one-step estimate retaining both terms). We use the estimate with the continuous target as initial law, so that the discrete contraction factor can be used directly in the invariant perturbation argument.

Lemma 14 (One-step strong error under $C^{1,1}$)**.** *Let $V \in C^{1,1}$ be globally m_M-strongly convex and L_V-smooth in the M-norm, and let $\pi_V \propto e^{-V}$. Start its Langevin diffusion and one exact midpoint step from the same $X_0 \sim \pi_V$, independently of a Brownian motion W. Couple their noises by $\xi = W_\delta/\sqrt{\delta}$. For $0 < \delta \le 2/L_V$,*

$$\left(\mathbb{E}\|X_1^{\mathrm{disc}} - X_\delta\|_M^2\right)^{1/2} \le \frac{L_V}{\sqrt{2}}\sqrt{d}\,\delta^{3/2} + L_V^2\left(\frac{3}{2}\sqrt{\frac{10d}{m_M}} + \frac{1}{2}\sqrt{\frac{d}{L_V}}\right)\delta^2. \tag{34}$$

Proof. Write $b = -M^{-1}\nabla V$, $p = (X_0 + X_1^{\mathrm{disc}})/2$, and let $x_\star$ minimise V. The coupled updates give

$$X_1^{\mathrm{disc}} = X_0 + \sqrt{2}M^{-1/2}W_\delta + \delta b(p), \qquad X_\delta = X_0 + \sqrt{2}M^{-1/2}W_\delta + \int_0^\delta b(X_t)\,dt,$$

so $e = X_1^{\mathrm{disc}} - X_\delta = \int_0^\delta [b(p) - b(X_t)]\,dt$. Integration by parts under π_V, followed by strong convexity, yields $\mathbb{E}_{\pi_V}\|X - x_\star\|_M^2 \le d/m_M$. This identity is justified by Gaussian tails and approximation for $C^{1,1}$ potentials. The diffusion is stationary, hence

$$\sup_{t\le\delta}(\mathbb{E}\|X_t - x_\star\|_M^2)^{1/2} \le R, \qquad R = \sqrt{10d/m_M}.$$

Since $\mathrm{prox}_{(\delta/2)V}^{M}(x_\star) = x_\star$, its nonexpansiveness and $\delta/2 \le 1/L_V$ give

$$R_p := (\mathbb{E}\|p - x_\star\|_M^2)^{1/2} \le R + \sqrt{d/L_V}.$$

The Brownian part of $p - X_t$ is $M^{-1/2}(W_\delta/\sqrt{2} - \sqrt{2}W_t)$, whose squared M-norm has expectation

$$\tfrac{1}{2}d\delta + 2dt - 2d\min(t,\delta) = d\delta/2 \qquad (0 \le t \le \delta).$$

The drift part is $(\delta/2)b(p) - \int_0^t b(X_s)\,ds$. Minkowski's inequality and $\|b(x)\|_M \le L_V\|x - x_\star\|_M$ imply

$$(\mathbb{E}\|p - X_t\|_M^2)^{1/2} \le \sqrt{d\delta/2} + \delta L_V(R_p/2 + R).$$

Finally Minkowski's inequality in the integral defining e and the L_V-Lipschitz property of b give

$$(\mathbb{E}\|e\|_M^2)^{1/2} \le L_V\sqrt{d/2}\,\delta^{3/2} + L_V^2(R_p/2 + R)\delta^2.$$

Inserting the bound on R_p proves (34). □

Apply the lemma with $V = U_\lambda$, and denote its right-hand side by $\eta_{\delta,\lambda} = C_1(\lambda)\sqrt{d}\,\delta^{3/2} + \bar{C}_2(\lambda)\delta^2$. Since the exact diffusion preserves $\pi_\lambda \propto e^{-U_\lambda}$, the lemma supplies $W_2^M(\pi_\lambda P_{\delta,\lambda}, \pi_\lambda) \le \eta_{\delta,\lambda}$. Invariance of $\pi_{\delta,\lambda}$, discrete contraction and the triangle inequality give

$$W_2^M(\pi_{\delta,\lambda}, \pi_\lambda) \le \rho(\delta,\lambda) W_2^M(\pi_{\delta,\lambda}, \pi_\lambda) + \eta_{\delta,\lambda}.$$

Rearranging proves

$$W_2^M(\pi_{\delta,\lambda}, \pi_\lambda) \le E_{\mathrm{disc}}(\delta,\lambda) := \frac{C_1(\lambda)\sqrt{d}\,\delta^{3/2} + \bar{C}_2(\lambda)\delta^2}{1-\rho(\delta,\lambda)}, \qquad 0 < \delta \le 2/L_\lambda. \tag{35}$$

The same argument with $V = f$ proves the zeroth-order convergence used in Appendix A.3, independently of BEA.

Part III (Moreau approximation). Global M-Lipschitzness of g is equivalent to $\|\zeta\|_{M^{-1}} \le L_g^M$ for every $\zeta \in \partial g(x)$. For $z = \mathrm{prox}_{\lambda g}^M(x)$ and such a subgradient at x, convexity and the definition of the envelope give

$$g(x) - g^{\lambda,M}(x) \le \langle \zeta, x - z\rangle - \frac{\|x-z\|_M^2}{2\lambda} \le L_g^M \|x-z\|_M - \frac{\|x-z\|_M^2}{2\lambda}.$$

Maximising the final quadratic over $\|x-z\|_M$ yields

$$0 \le g(x) - g^{\lambda,M}(x) \le \tfrac{\lambda}{2}(L_g^M)^2. \tag{36}$$

Writing Z and Z_λ for the normalising constants, $Z_\lambda \ge Z$ and therefore

$$\mathrm{KL}(\pi_\lambda \| \pi) = \mathbb{E}_{\pi_\lambda}(g - g^{\lambda,M}) + \log(Z/Z_\lambda) \le \tfrac{\lambda}{2}(L_g^M)^2.$$

The potential $f+g$ is m_M-strongly convex in the M-norm. Whitening and the Talagrand inequality (also valid for this nonsmooth strongly convex potential by convex approximation) give

$$W_2^M(\pi_\lambda, \pi) \le L_g^M \sqrt{\lambda/m_M}.$$

Part IV (Triangulation). Theorem 3 gives $W_2^M(\mu_0 P_{\delta,\lambda}^N, \pi_{\delta,\lambda}) \le \rho(\delta,\lambda)^N W_0$. Add this to (35) and the Moreau bound to obtain (15). Only the contraction statement is valid for arbitrary step size; the displayed discretisation estimate retains the condition $\delta \le 2/L_\lambda$.

For the finite-dimensional TV penalty $g(x) = \rho_{\mathrm{TV}}\|Dx\|_1$, one may take $L_g^M = \rho_{\mathrm{TV}} \sup_{\|s\|_\infty \le 1} \|D^\top s\|_{M^{-1}} < \infty$. This makes explicit the metric dependence of the smoothing term.

B JP-IMLA: kernel guarantees and empirical comparisons

B.1 Kernel guarantees for JP-IMLA

Throughout this subsection $M \succ 0$ is constant, f is globally m_M-strongly convex and L_M-smooth in the M-norm, and g is finite, convex and Lipschitz in that norm. Write $L_g^M = \sup_{(x,s)\in \mathrm{gra}\,\partial g} \|s\|_{M^{-1}} < \infty$, $U_\lambda = f + g^{\lambda,M}$ and $\tau = \delta/2$. At input $w = x + \sqrt{\delta/2}\,M^{-1/2}\xi$, $\xi \sim \mathcal{N}(0,I)$, define

$$\begin{aligned} p(w) &= \mathrm{prox}_{\tau(f+g)}^M(w), \qquad q(w) = \mathrm{prox}_{\tau U_\lambda}^M(w), \\ Q &: x \mapsto -x + 2p(w), \quad P : x \mapsto -x + 2q(w). \end{aligned} \tag{37}$$

Q is the exact joint-proximal kernel; P is the canonical smoothed kernel analysed in Theorem 6. The parameter λ indexes the comparison with P; the exact joint proximal map itself does not depend on λ. The finite-ADMM implementation also carries inner variables between steps and will be treated separately below.

Proposition 15 (Linear proximal and kernel gap)**.** *For every w and $\delta, \lambda > 0$,*

$$\|p(w) - q(w)\|_M \le \lambda L_g^M \frac{\tau^{-1} + L_M}{\tau^{-1} + m_M}, \qquad \varepsilon_{\mathrm{model}} := 2\lambda L_g^M \frac{\tau^{-1} + L_M}{\tau^{-1} + m_M}. \tag{38}$$

With $L_\lambda = L_M + \lambda^{-1}$ and $\rho = \max_{a\in\{m_M, L_\lambda\}} |(1-\tau a)/(1+\tau a)| < 1$, the same-noise coupling yields, for $\mu, \nu \in \mathcal{P}_2$,

$$W_2^M(Q\mu, P\nu) \le \varepsilon_{\mathrm{model}} + \rho W_2^M(\mu,\nu). \tag{39}$$

Proof. Whiten by $z = M^{1/2}x$; the following argument is Euclidean with constants m_M, L_M, L_g^M, and transforms back without loss. Let $r = \operatorname{prox}_{\lambda g}(q)$, $s_r = (q-r)/\lambda \in \partial g(r)$, and choose $s_p \in \partial g(p)$ from the joint optimality condition. With $d = q - p$ and $h = q - r$, subtraction of the two optimality conditions gives

$$Ad + s_r - s_p = 0, \qquad Ad := \tau^{-1}(q-p) + \nabla f(q) - \nabla f(p).$$

Monotonicity between (r, s_r) and (p, s_p) gives $\langle Ad, d - h\rangle \le 0$. Strong monotonicity and Lipschitzness of $\tau^{-1}I + \nabla f$ imply

$$(\tau^{-1} + m_M)\|d\|^2 \le \langle Ad, d\rangle \le \langle Ad, h\rangle \le (\tau^{-1} + L_M)\|d\|\, \lambda L_g^M.$$

This proves (38), including $d = 0$. Reflection doubles the proximal difference. At a common state and noise, the Q and P outputs thus differ by at most $\varepsilon_{\mathrm{model}}$; the P outputs from two states differ by at most ρ times the state distance by Theorem 3. Apply the triangle inequality in L^2 to an optimal initial coupling to obtain (39). □

The linear bound strengthens the submitted $L_g^M\sqrt{\tau\lambda/2}$ proximal bound; a smaller observed gap does not contradict that earlier upper bound. At the spectrally balanced $\tau_* = (m_M L_\lambda)^{-1/2}$, direct substitution gives

$$\frac{\varepsilon_{\mathrm{model}}}{1-\rho_*} = L_g^M\sqrt{\frac{\lambda}{m_M}}\, C_{\mathrm{JP}}(\lambda), \quad C_{\mathrm{JP}}(\lambda) = \frac{\sqrt{m_M L_\lambda} + L_M}{\sqrt{m_M L_\lambda} + m_M}\left(\sqrt{\lambda L_\lambda} + \sqrt{\lambda m_M}\right) \longrightarrow 1. \quad (40)$$

Thus the accumulated joint-model gap has the same order as the Moreau term. This identity concerns the kernel comparison, which holds for every $\delta > 0$; it does not enlarge the step range of the discretisation estimate in Theorem 6. For TV, $L_g^M = O(\rho_{\mathrm{TV}}\sqrt{d})$ with a constant depending on D and M; the dimension dependence is not removed by the comparison.

A certificate for the finite ADMM solve. Suppose $g = \phi \circ D$ and set $F_w(x) = f(x) + \|x - w\|_M^2/(2\tau)$, $J_w(x) = F_w(x) + \phi(Dx)$ and $\mu_{\mathrm{prox}} = m_M + \tau^{-1}$. Let (x_K, z_K, y_K) be the primal, split and unscaled dual variables after a complete ADMM iterate with an exact z update. Then $y_K \in \partial\phi(z_K)$: this follows directly from the z optimality condition and the subsequent dual update. Define

$$\begin{aligned} a_K &= \nabla F_w(x_K) + D^\top y_K, \\ c_K &= \phi(Dx_K) - \phi(z_K) - \langle y_K, Dx_K - z_K\rangle, \\ U_K &= c_K + \frac{\|a_K\|_{M^{-1}}^2}{2\mu_{\mathrm{prox}}}. \end{aligned} \quad (41)$$

Proposition 16 (Primal–dual inner-error certificate)**.** *Under the preceding conditions,*

$$0 \le J_w(x_K) - J_w(p) \le U_K, \qquad \|x_K - p\|_M \le \sqrt{2U_K/\mu_{\mathrm{prox}}}. \quad (42)$$

Proof. Fenchel weak duality gives $-F_w^*(-D^\top y_K) - \phi^*(y_K) \le J_w(p)$. Put $s = -D^\top y_K$. The primal–dual gap is the sum of

$$F_w(x_K) + F_w^*(s) - \langle s, x_K\rangle \quad \text{and} \quad \phi(Dx_K) + \phi^*(y_K) - \langle y_K, Dx_K\rangle.$$

The first term is at most $\|\nabla F_w(x_K) - s\|_{M^{-1}}^2/(2\mu_{\mathrm{prox}})$ because F_w^* is $1/\mu_{\mathrm{prox}}$-smooth in the dual norm. Fenchel equality at (z_K, y_K) makes the second term exactly $c_K \ge 0$. Finally, μ_{prox}-strong convexity of J_w gives $J_w(x_K) - J_w(p) \ge \mu_{\mathrm{prox}}\|x_K - p\|_M^2/2$. □

For anisotropic TV, $\phi(z) = \rho_{\mathrm{TV}}\|z\|_1$ and the exact soft-threshold update ensures $\|y_K\|_\infty \le \rho_{\mathrm{TV}}$. The equivalent box-feasible expression $c_K = \rho_{\mathrm{TV}}\|Dx_K\|_1 - \langle y_K, Dx_K\rangle$ can be used after projecting a floating-point dual onto this box; it only requires dual feasibility. A split residual alone omits this complementarity term and is not the certificate in (41).

Implemented process and end-to-end bound. Let $\widehat{X}_j$ be the image iterate and S_j all warm-started ADMM variables. At each step fresh Gaussian noise is independent of the past; the inner budget K can be fixed or chosen from the history. Apply the certificate at that step's actual input and let

$$e_{K,j} = \left(\frac{8\mathbb{E}U_{K,j}}{\mu_{\mathrm{prox}}}\right)^{1/2}, \qquad D_0 = W_2^M(\mathcal{L}(\widehat{X}_0), \pi_{\delta,\lambda}),$$

where $\pi_{\delta,\lambda}$ is the invariant law of P. Whenever the displayed expectations are finite,

$$\begin{aligned} W_2^M(\mathcal{L}(\widehat{X}_N), \pi) \le \rho^N D_0 + \sum_{j=0}^{N-1} \rho^{N-1-j}(\varepsilon_{\mathrm{model}} + e_{K,j}) \\ + \frac{C_1(\lambda)\sqrt{d}\,\delta^{3/2} + \overline{C}_2(\lambda)\delta^2}{1-\rho} + L_g^M\sqrt{\lambda/m_M}. \end{aligned} \tag{43}$$

Here the constants and step range are those of Theorem 6; in particular its displayed small-step estimate applies for $0 < \delta \le 2/L_\lambda$. The kernel and solver terms themselves have no such step restriction.

To prove the marginal recursion, at each step optimally couple $\widehat{X}_j$ with $Y_j \sim \pi_{\delta,\lambda}$, lift this coupling to $(\widehat{X}_j, S_j, Y_j)$ by conditioning on $\widehat{X}_j$, and draw fresh shared noise independently. Evolve Y_j with P. By (42), reflection gives $\|\widehat{X}_{j+1} - Q(\widehat{X}_j, \xi_j)\|_M \le 2\sqrt{2U_{K,j}/\mu_{\mathrm{prox}}}$. Writing $D_j = W_2^M(\mathcal{L}(\widehat{X}_j), \pi_{\delta,\lambda})$, Minkowski's inequality and (39) yield $D_{j+1} \le \rho D_j + \varepsilon_{\mathrm{model}} + e_{K,j}$. Iteration and the stationary discretisation and smoothing terms of Theorem 6 prove (43). This bounds the image marginal of the augmented process $(\widehat{X}_j, S_j)$; no Markov assumption on $\widehat{X}_j$ alone is used.

Existence and uniqueness for the exact joint kernel.

Proposition 17 (Existence and stationary comparison)**.** *For every fixed $\delta > 0$, Q has an invariant law in $\mathcal{P}_2$. Every such law ν_Q satisfies $W_2^M(\nu_Q, \pi_{\delta,\lambda}) \le \varepsilon_{\mathrm{model}}/(1-\rho)$.*

Proof. Iterate (39) against the stationary P chain. For any $\mu_0 \in \mathcal{P}_2$, the distances $W_2^M(Q^n\mu_0, \pi_{\delta,\lambda})$ are bounded uniformly in n. The triangle inequality in L^2 therefore bounds the second moments of $Q^n\mu_0$ uniformly. The kernel is Feller because the proximal map is Lipschitz; dominated convergence applies to bounded continuous test functions. Tightness of the Cesaro averages and the Krylov–Bogoliubov argument give an invariant law, with finite second moment by lower semicontinuity. Apply (39) to any invariant ν_Q and the invariant P law and rearrange to obtain the stated distance bound. □

For clarity, reflected proximal maps need not be strict contractions: for $f(x) = x^2/2$, $g(x) = a|x|$, the proximal map is zero on $[-\tau a, \tau a]$ and its reflection is $x \mapsto -x$ there. A synchronous coupling nonetheless satisfies the following useful bound. Write $D_n = \|X_n - Y_n\|_M$ for two exact Q chains.

$$D_{n+1}^2 \le D_n^2 - 4\tau m_M \left(\frac{(D_n - 2\tau L_g^M)_+}{1+\tau L_M}\right)^2. \tag{44}$$

Indeed, in whitened coordinates put $a = X_n - Y_n$ and $b = p(w_X) - p(w_Y)$. Optimality gives $a = b + \tau(u_X - u_Y)$, with $u_i \in \partial(f+g)(p(w_i))$. Thus $\|2b - a\|^2 \le \|a\|^2 - 4\tau m_M\|b\|^2$, while $\|a\| \le (1+\tau L_M)\|b\| + 2\tau L_g^M$. This proves (44). In particular D_n decreases and $\limsup_n D_n \le \delta L_g^M$. For invariant initial marginals, dominated convergence gives the unconditional comparison $W_2^M(\nu_1, \nu_2) \le \delta L_g^M$ between any two invariant laws.

Theorem 18 (Unique invariant law in the imaging setting)**.** *If f is strongly convex quadratic and $g(x) = \sum_{e=1}^E \rho_e |d_e^\top x|$, with finitely many nonzero d_e and $\rho_e > 0$, then Q has a unique invariant law ν_Q in $\mathcal{P}_2$. For every $\mu_0 \in \mathcal{P}_2$, $W_2^M(Q^n\mu_0, \nu_Q) \to 0$. This statement gives no convergence rate.*

Proof. Whiten and write $f(x) = \frac{1}{2}(x-\mu)^\top H(x-\mu) + \mathrm{const}$, with $m_M I \preceq H \preceq L_M I$; the transformed g is still a finite ℓ_1 composite. Set $A = I + \tau H$. Choose a unit vector v outside the finite

Table 6: Eight-chain kernel comparison at 64^2, $\lambda = 0.005$, $\delta = 0.141$. The gate predicts an initial RMS separation 0.00412567; its measured/predicted ratio is 1.0000226.

N	RMS M-distance	Across-chain min–max	Model envelope
1	0.004126	0.002429–0.005619	0.11647
10	0.007550	0.006160–0.008652	0.66887
60	0.007132	0.005937–0.008315	0.88406

union of hyperplanes $d_e^\top A^{-1} v = 0$, and put $\sigma_e = \operatorname{sign}(d_e^\top A^{-1} v)$. For this fixed sign vector define the affine branch

$$p_\sigma(w) = A^{-1}\left(w + \tau H\mu - \tau \sum_e \rho_e \sigma_e d_e\right).$$

For any compact set of pairs (x, y), a sufficiently large fixed t makes every $\sigma_e d_e^\top p_\sigma(w)$ positive for all $w = x + tv + z$, $w = y + tv + z$, $\|z\| \le 1$, and their connecting segments. This follows because the coefficient of t is the strictly positive $|d_e^\top A^{-1} v|$, and the remaining terms are uniformly bounded on the compact set. At every such w, the branch's sign pattern verifies the joint-proximal KKT condition, so strong convexity gives $p(w) = p_\sigma(w)$. On this region the reflected map has linear part $2A^{-1} - I$ and norm

$$\rho_f = \max_{a \in \{m_M, L_M\}} |(1 - \tau a)/(1 + \tau a)| < 1.$$

Consequently the Gaussian event $\|\sqrt{\delta/2}\,\xi - tv\| \le 1$ has a strictly positive probability p_C, uniform over pairs in the chosen compact set C, and on this event $D_{n+1} \le \rho_f D_n$. This argument uses a fixed affine branch and checks KKT at its endpoints; it makes no monotonicity assumption about a true prox path as it crosses active faces.

For two invariant initial marginals, or one arbitrary $\mu_0 \in \mathcal{P}_2$ and one invariant marginal, use synchronous noises. The reflected resolvent is nonexpansive, so $D_n \downarrow D_\infty$ almost surely and $D_n \le D_0 \in L^2$. The second-moment bound proved above makes the pair of marginals uniformly tight. If $q = \mathbb{P}(D_\infty \ge a) > 0$ for some $a > 0$, choose compact C with $\sup_n \mathbb{P}((X_n, Y_n) \notin C) < q/2$. The strict-contraction event then gives, at every n,

$$\mathbb{E}(D_n^2 - D_{n+1}^2) \ge (1 - \rho_f^2) p_C\, \mathbb{E}[D_n^2 \mathbf{1}_{\{(X_n, Y_n) \in C\}}] \ge (1 - \rho_f^2) p_C a^2 q/2 > 0.$$

Summing contradicts $\mathbb{E}D_0^2 < \infty$. Hence $D_\infty = 0$; dominated convergence gives $\mathbb{E}D_n^2 \to 0$, proving uniqueness and the stated W_2 convergence. □

Remark 19 (Scope of the JP guarantees)**.** *The exact joint kernel has the model-gap and stationary guarantees above, and the finite-ADMM process has* (43)*. Uniqueness of the implemented warm-start process is not asserted. The uniqueness theorem covers quadratic likelihoods with anisotropic TV or a finite wavelet-ℓ_1 penalty; it does not assert uniqueness for every convex nonsmooth prior or a quantitative mixing rate for Q. The measured per-iteration speedup reported below retains its original solver-budget and hardware conditions; it is not a consequence of a Wasserstein bound or a guarantee of equal accuracy at equal cost.*

B.2 Kernel and finite-solver audits

Exact-kernel comparison. The 64×64 audit solves the two proximal problems to a relative movement tolerance 10^{-8} across six values of λ spanning a factor of 40. The maximum gap over three probe points has log–log slope 0.985, consistent with (38). Figure 4 reports this audit; its right panel is one shared-noise realization. The independent eight-chain aggregation is reported in Table 6 so that a single realization is not confused with a multi-chain RMS.

Production-budget certificate. At 256^2, $\delta = 0.141$, $\rho_{\mathrm{TV}} = 0.3$, and $\rho_a = 3$, the three-probe reference model-gap scale is 7.16792 in the reflected M-distance. At an audited warm-start input after 200 preliminary updates, the production budget $K = 12$ gives a measured step error 2.92717, or 0.408 times that scale. Its converged reference solve is recorded in the supplied output. The computable certificate is 56.87016, or 7.934 times the same scale: the observed 0.4 ratio is a reference-solve

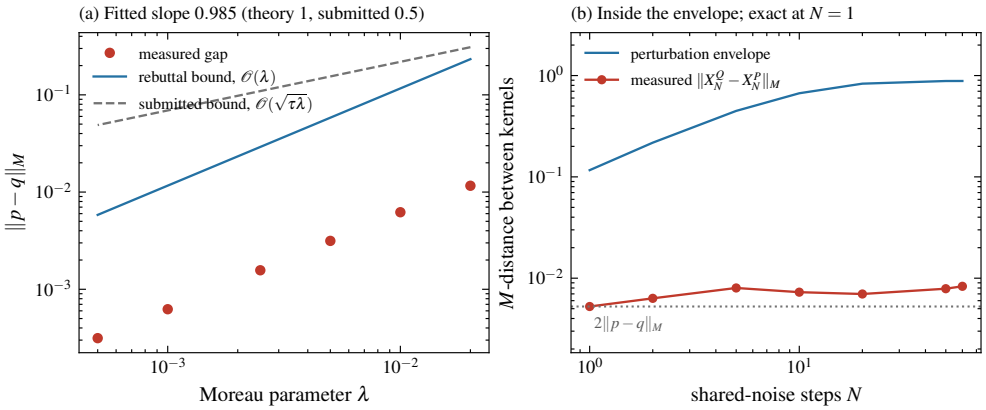

Figure 4: **Joint versus smoothed proximal kernels.** Left: measured proximal gap and the linear and earlier square-root upper bounds. The title's "submitted 0.5" denotes the exponent of the earlier upper bound, not a predicted measured slope. Right: the original single shared-noise realization and the model-gap envelope. The corrected eight-chain RMS values are in Table 6; neither panel is a bound on SSE.

Table 7: JP inner-budget audit at 256^2. SSE is a reconstruction statistic (mean $\pm$ SE over eight chains), not a kernel-distance estimate. "Certificate" is the separate warm-input upper bound in reflected M-distance; it does not use a converged reference solve.

Method / ADMM budget	Per-chain SSE	Time (min)	Certificate
JP, $K = 12$	851.80 ± 0.77	0.198	56.870
JP, $K = 25$	885.46 ± 0.80	0.386	–
JP, $K = 50$	899.52 ± 0.82	0.715	–
JP, $K = 100$	904.23 ± 0.83	1.348	2.842
JP, $K = 200$	905.40 ± 0.83	2.904	–
TRUE Picard	854.49 ± 0.76	25.240	–

comparison, not a certified upper bound. For $K = 100$ the certificate is 2.84233 (0.397 times that scale). These are pointwise solver audits, not estimates of the expectation $\mathbb{E}U_{K,j}$ along every step of a production chain. The latter $K = 100$ reference solve did not satisfy its requested stopping tolerance, so we use its certificate but do not report a "true error" from that reference. The audit establishes the usefulness of the certificate without declaring every finite reference solve converged.

Inner-budget sensitivity and timings. Table 7 uses the saved, paired geometric-phantom runs with 8 chains, $N = 2000$, $X_0 = 0$, and time averages accumulated from the first step. The TRUE Picard reference uses its submitted $(K_{\text{Picard}}, K_{\text{inner}}) = (30, 50)$; JP varies only its ADMM budget. These supplementary timings were measured on an RTX 5000 Ada and are kept separate from the original Blackwell timings below. The near agreement of JP at $K = 12$ with TRUE Picard changes as the inner budget increases: the SSE offset is -0.32% at $K = 12$ and $+5.96\%$ at $K = 200$. This is consistent with partial cancellation between solver and model errors at the production budget. It does not establish equality of the invariant laws, and the M-distance bound in Proposition 15 is not a relative-SSE bound. The original production setting is retained; the larger budgets serve to expose this sensitivity and to report a tighter certificate.

Reproducibility. The eight-chain output is `rebuttal/R4/exp_r4_jp_kernel_K8_20260727.json`. The certificate and budget-sweep scripts and JSON outputs are in `rebuttal/R4/`; `rebuttal/R4/README_camera_ready.md` gives the commands, fixed configurations, and reference-convergence limitations. The audit scripts preserve raw outputs and write new checks to separate paths.

B.3 Finite-budget comparison at early-N

We run JP-IMLA on the geometric phantom under the same protocol as the body's TRUE-Picard headline benchmark—same σ_{blur}, σ_{noise}, ρ_{TV}, $\delta = 0.141$, $\lambda = 5 \times 10^{-3}$, $K = 8$ chains from $X_0 = 0$, and *matched seed base*—so that JP and TRUE trajectories are directly comparable per chain. These are reconstruction and mixing diagnostics; Proposition 15 bounds reflected proximal distance, not relative SSE.

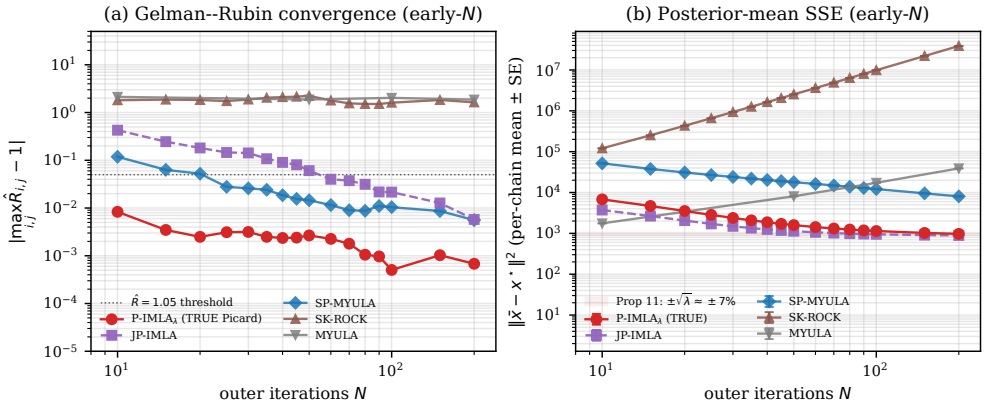

Figure 5: **Finite-budget JP/Picard comparison at early-N.** (a) Classical per-pixel $\hat{R}$ from the common zero start. (b) Per-chain reconstruction SSE. At $N = 200$, JP gives 888.28 ± 2.63 versus Picard's 972.09 ± 2.62, an 8.6% difference. The shaded percentage band is a visual guide from the original plot, not a consequence of a proximal or Wasserstein bound; it does not certify relative SSE.

Wall-clock per outer iteration. At 256×256 on an NVIDIA RTX PRO 6000 Blackwell, the recorded JP implementation averages 35.0 ms per iteration against Picard's 2.11 s, approximately $60\times$. The headline Picard budget is $(K_{\text{P}}, K_{\text{in}}) = (30, 50)$, while JP uses $K = 12$ ADMM steps. These are implementation timings under the recorded parallel workloads, not a complexity ratio at matched certified accuracy. Table 7 reports separate RTX 5000 Ada budget-sweep timings and certificates at audited warm inputs; those pointwise certificates do not certify every step of the plotted production runs.

B.4 Imaging-scale comparison

We extend the matched-budget benchmark of §5 to JP-IMLA on all three phantoms used in Table 4 (geometric, cameraman, Shepp–Logan), at $N = 2{,}000$ on RTX PRO 6000 Blackwell, with seed offsets matching the canonical TRUE-Picard runs.

Table 8: **Cross-phantom JP-IMLA vs TRUE Picard at $N = 2{,}000$.** Per-chain SSE means $\pm$ SE (JP: eight chains; Picard: eight geometric and four per other phantom). The kernels and finite inner solvers differ. Relative SSE is a reconstruction statistic, not a distributional error bound.

Phantom	TRUE Picard SSE	JP-IMLA SSE	Drift (JP/TRUE -1)
geometric	855.22 ± 0.66	852.72 ± 0.65	-0.29%
cameraman	890.24 ± 1.19	884.52 ± 0.62	-0.64%
Shepp–Logan	$1{,}013.99 \pm 0.81$	$1{,}012.27 \pm 0.63$	-0.17%

At these finite budgets, the reported JP and Picard SSEs differ by at most 0.7%. This agreement does not imply equality of invariant laws or uniform solver accuracy; the model and finite-ADMM terms are quantified separately in Appendix B.1.

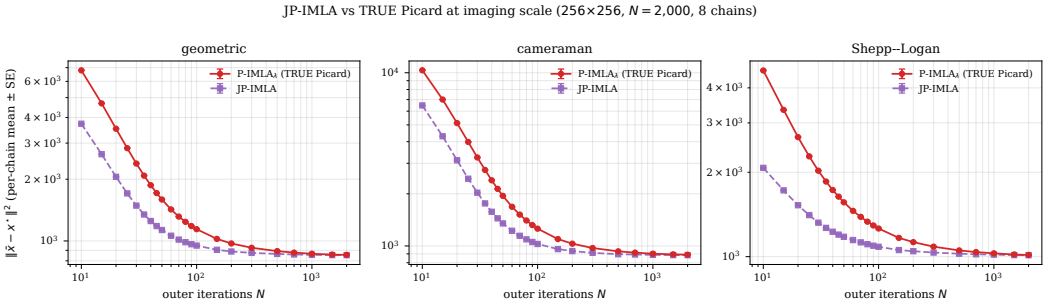

Figure 6: **JP-IMLA tracks TRUE Picard at imaging scale across three phantom morphologies.** Per-chain SSE means $\pm$ SE; both samplers run the same $N = 2{,}000$-iteration matched-budget protocol. The TRUE/JP traces are visually indistinguishable for $N \geq 200$ on all three phantoms. JP uses the original $K = 12$ budget; Table 7 separately reports budget sensitivity and pointwise certificates.

C Auxiliary experimental results

This appendix collects (a) validation experiments for Theorems 2–5 that the main-text §5 summarises in one sentence, and (b) intermediate-checkpoint reconstructions for the matched-budget benchmark of Figure 3.

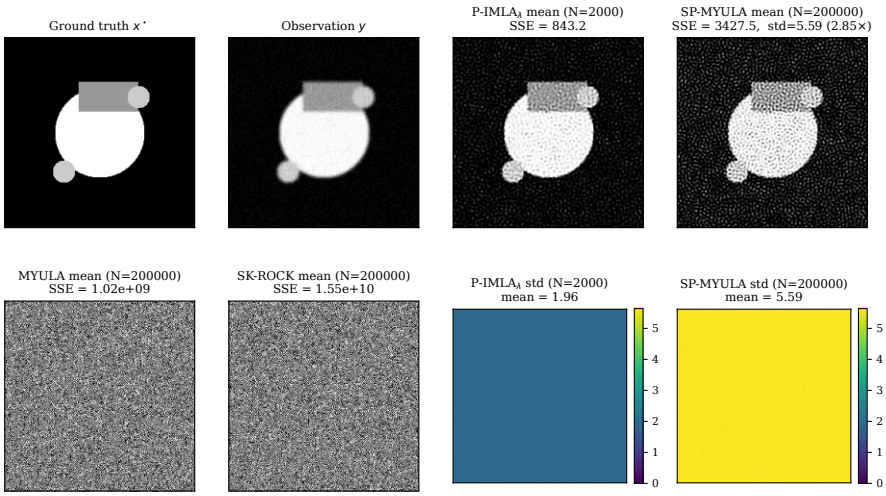

Figure 7: **Geometric phantom counterpart to main-text Figure 3**, matched-budget reconstructions: ours at $N = 2{,}000$ (top right; ensemble SSE 843.1, std mean 1.95); SP-MYULA / MYULA / SK-ROCK at $N = 2{\times}10^5$. Same structural conclusions as the cameraman panel apply: ours recovers the four phantom regions cleanly, SP-MYULA's mean is similar but std uniformly inflated $\sim$2.86×, MYULA and SK-ROCK mean fields are unrecognisable.

Geometric- and Shepp–Logan-phantom imaging panels.

P-IMLA$_\lambda$ recovery at intermediate N (geometric). Figure 7 shows the headline pair $N_{\text{ours}} = 2{,}000$ vs $N_{\text{base}} = 2{\times}10^5$. Identically rendered panels at $N_{\text{ours}} \in \{10, 50, 100, 200, 500, 1000\}$ (each paired with the corresponding 100× baseline budget) are provided in the supplementary material. Visual quality stabilises by $N \approx 200$ (where $\hat{R} = 1.0004$); $N = 10$ and $N = 50$ show partial mixing artefacts. The contraction timescale $N \approx 22$ does not predict a threshold for $\hat{R}$ or visual quality.

Uncertainty conventions are experiment-specific. For averages over K independent seeds, SE denotes the sample standard deviation divided by $\sqrt{K}$; single-chain batch-means errors and deterministic calculations are identified separately below.

Sampler naming conventions. Throughout this appendix we use *SP-IMLA* for the smooth-target P-IMLA of (9) with $M = H$; *SP-ULA* for the corresponding explicit-Euler scheme $X_{n+1} = X_n - \delta M^{-1}\nabla f(X_n) + \sqrt{2\delta}\, M^{-1/2}\xi_n$ at $M = H$; and *SP-MYULA* for the explicit-Euler counterpart of P-IMLA$_\lambda$ on the same M-norm-Moreau-smoothed potential $U_\lambda = f + g^{\lambda,M}$:

$$X_{n+1} \;=\; X_n - \delta M^{-1}\nabla U_\lambda(X_n) + \sqrt{2\delta}\, M^{-1/2}\xi_n, \qquad M = H, \tag{45}$$

identical to MYULA [Durmus et al., 2018] except (i) the Moreau envelope is taken in the M-norm matching the preconditioner geometry (matching the body's convention), and (ii) the drift and noise are preconditioned by M^{-1} and $M^{-1/2}$ respectively. SP-MYULA isolates the effect of midpoint vs. Euler integration with preconditioner choice, Moreau parameterisation, and noise injection all matched to P-IMLA$_\lambda$, the observed SSE gap also depends on finite-solver accuracy and initial transients.

C.1 Variance-budget identity (Theorem 2 and Proposition 1)

We test the variance-budget identity (2) and the Euler overfill (1) on two settings: **(a)** a 1-D Gaussian $f(x) = x^2/2$ at $M = 1$, swept across eight step sizes $\delta \in [0.1, 1.8]$; and **(b)** a $d = 100$ anisotropic Gaussian $H = \mathrm{diag}(1, \ldots, 100)$, $M = H$, $\kappa_{\mathrm{eff}} = 1$. Each sampler runs $N = 5\times10^6$ post-burn-in samples. Setup A uses batch-means SEs (batch size 10^4). Setup B computes per-mode variance ratios VR_j from one vector chain in whitened coordinates; the SE of their average is the sample standard deviation across the 100 independent Gaussian modes divided by $\sqrt{100}$.

Table 9: **Variance-budget verification (1-D Gaussian, Setup A).** SP-IMLA's VR lies in $[0.9978,\ 1.0009]$ across all eight δ ($|z|$-scores ≤ 1.34); Setup B per-mode average is 0.999937 ± 0.000087. SP-ULA tracks $1/(1-\delta/2)$ to within 0.12%.

δ	SP-IMLA VR	SP-ULA VR	Theory $1/(1-\delta/2)$
0.1	0.99885	1.0539	1.0526
0.5	1.00033	1.3324	1.3333
1.0	0.99949	2.0015	2.0000
1.5	0.99953	3.9986	4.0000
1.8	0.99941	9.9976	10.000

SP-IMLA's VR is statistically indistinguishable from 1 across an 18-fold step-size range, confirming the Cayley algebra cancellation (2) to four decimals even at $\delta = 1.8$, where SP-ULA's variance has inflated ten-fold. Theorem 2 makes no step-size assumption; explicit Euler breaks down precisely in the large-δ regime that preconditioning calls for.

C.2 Quantitative 2×2 landscape on a Gaussian target

Figure 1 reports a fixed-budget statistic on the Gaussian setup of C.3. Explicit Euler is either slow at stable steps or unstable beyond its spectral boundary. Both midpoint schemes have the exact Gaussian invariant law with exact proximal solves. For unpreconditioned IMLA, low-curvature modes retain their initial-mean error over the reported horizon: 98.7% of the displayed VR $= 48.6$ is transient, while the post-burn-in stationary ratio is 1.000. The quantity is a time-averaged normalised mean-square error, not stationary variance inflation. Preconditioning reduces the relevant condition number and shortens that transient. Appendix D.5 gives the exact AR(1) decomposition and its stationary-start check.

C.3 Contraction rate and κ_f-independence (Theorem 3)

The path-level panel verifies the deterministic contraction inequality of Theorem 3: at every iteration the M-norm separation contracts to within float precision, with Gaussians approaching the contraction

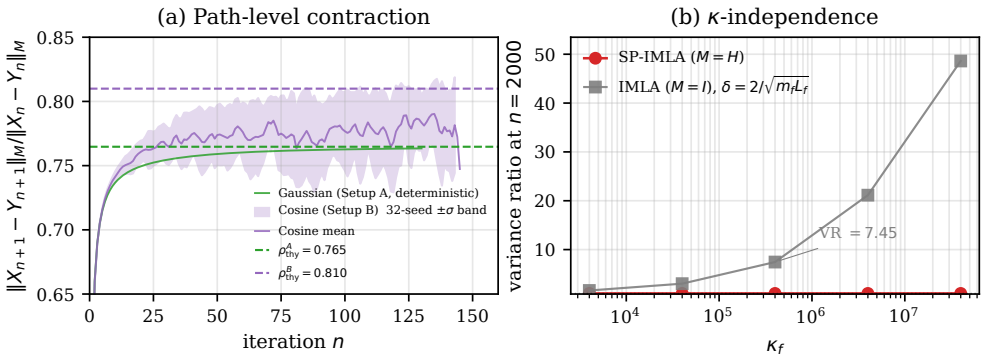

Figure 8: **(a)** Path-level contraction ratio $\|X_{n+1} - Y_{n+1}\|_M / \|X_n - Y_n\|_M$ for two chains driven by shared noise. Setup A (Gaussian, deterministic) attains mean ratio 0.7544, 98.65% of the bound $\rho_{\text{thy}} = 0.7647$. Setup B (32 independent seeds, $d = 50$ cosine-perturbed target $f + 0.3\sum_j(\cos x_j - 1)$) gives seed-mean max ratio 0.8074 ± 0.0015 vs the strict bound $\rho_B = 0.8100$ (1.73 SE below). Shading shows one standard deviation across seeds; quoted uncertainties are SEs. **(b)** Cold-start time-averaged normalized MSE over $n = 2000$ iterations as κ_f varies over four orders of magnitude (Exp 04, 32 independent seeds, $M = H$). SP-IMLA's $\text{VR}^{\text{mode}} = 1.0246 \pm 0.0001$ across $\kappa_f \in [4\times 10^3,\ 4\times 10^7]$; IMLA grows monotonically from $\text{VR} = 7.449$ to 48.6. These values include the initial-mean transient; both exact midpoint kernels preserve Gaussian stationary variance (Appendix D.5).

bound more tightly than the cosine target (98.65% vs 93.13%)—quantitative evidence that Gaussian targets attain the bound (Theorem 3). The κ-independence panel verifies the resulting $\sqrt{\kappa_{\text{eff}}}$ rate at the mixing-rate optimum (Remark 7): the $\sqrt{\kappa_f}$ dependence the IMLA curve exhibits is replaced under preconditioning by a constant.

C.4 Joint smoothing and step-size sweep at imaging scale

A λ-sweep at the same 256×256 setting over a 200-fold range $\lambda \in [10^{-3},\ 0.2]$ (Exp 10) exhibits two robust empirical features: (i) $\kappa_{\text{eff}}^{\lambda} = 1 + 1/\lambda$ matches the measured spectrum at every λ to four significant figures, while MYULA's κ_{eff} remains four orders of magnitude larger; (ii) ESS/iter saturates at 1.00 for $\lambda \geq 5\times10^{-3}$. The sweep traces a joint $(\delta_\lambda^\star, \rho_a^\star)$ schedule rather than independently varying λ in the inner solver; Term III's $\sqrt{\lambda}$ scaling is verified directly in 1-D in §C.9 (Remark 8).

C.5 Step-size robustness and the antithetic regime

Sweeping δ on a 64×64 non-smooth benchmark (same $M = \hat{H}$, $\lambda = 5\times10^{-3}$, $\delta^\star = 2/\sqrt{m_M L_\lambda} \approx 0.141$), we observe a regime not predicted by Theorem 3's worst-case bound. The Cayley factor $r(\ell)$ is negative for eigenvalues $\ell > 2/\delta$, allowing alternating-sign correlations. At $\delta > \delta^\star$, the reported effective-sample-size per iteration (Geyer initial-positive-sequence estimator, capped at one in our implementation) plateaus at $\text{ESS}/\text{iter} = 1.00$ for $\delta \in [0.2,\ 1.0]$ (Figure 9), an order-of-magnitude robustness range beyond $\delta^\star$. The Wasserstein contraction of Theorem 3 does not forbid this—$|r(L_\lambda)| \to 1$ as $\delta \to \infty$—but at finite δ alternating-sign autocorrelations can reduce variance without improving the worst-case path-level bound. A signed analysis of this regime is an open problem.

C.6 Beyond Fourier-diagonal forward operators (random fan-beam)

To verify that the framework is not specific to circulant forward operators, we re-run the matched-budget protocol of §5 with a non-Fourier A. **Setup**: 64×64 image ($d = 4{,}096$), $m = 4{,}096$ random fan-beam projection lines through the image domain (sparse $A \in \mathbb{R}^{m\times d}$ with $\sim 2.51\times10^5$ nonzero entries, ~ 64 pixels per ray); Gaussian noise $\sigma_{\text{noise}} = 0.05$; TV regulariser $\rho_{\text{TV}} = 0.3$, $\lambda = 5\times10^{-3}$, $\delta = 0.141$ as in the headline benchmark. The preconditioner is a randomised Nyström [Frangella et al., 2023] approximation $M = U\Lambda U^\top + \beta I$ of $A^\top A/\sigma_{\text{noise}}^2$ at rank $r = 128$; the inner Picard

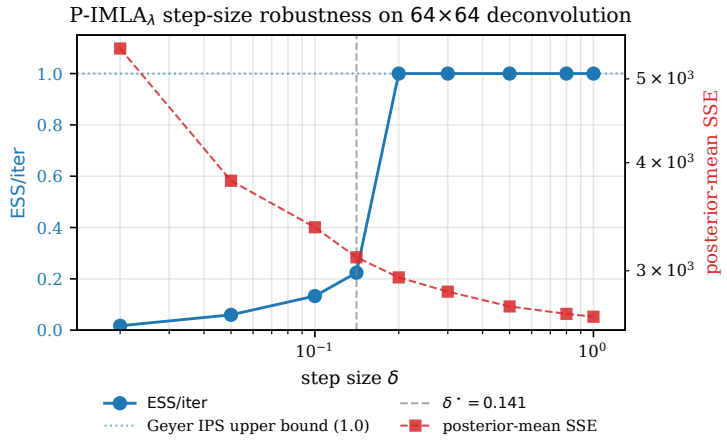

Figure 9: **Step-size robustness of P-IMLA$_\lambda$ on 64×64 deconvolution.** ESS/iter rises through the worst-case $\delta^\star = 0.141$ and *saturates at* 1.00 (implementation cap) across $\delta \in [0.2,\ 1.0]$. Posterior-mean SSE remains bounded across the same range. The negative-Cayley antithetic regime at $\delta > \delta^\star$ produces alternating-sign autocorrelations that the reflected-proximal contraction bound of Theorem 3 (which uses $|r(L_M)|$) does not capture; a signed analysis closing this gap is an open problem.

u-update (13) is no longer Fourier-diagonal, so we replace it by $K_{\text{cg}} = 12$ steps of conjugate gradient on the SPD operator $L = (1/\lambda)M + \rho_a D^\top D$ with relative tolerance 10^{-6}. This auxiliary run uses $(K_{\text{P}}, K_{\text{in}}) = (8, 15)$ and ADMM warm starts, a separate inner-solver budget from the headline $(30, 50)$ protocol.

Matched-budget result. P-IMLA$_\lambda$ runs $K = 8$ chains for $N = 300$ outer iterations; SP-MYULA runs $K = 4$ chains for $N = 30{,}000$ ($100\times$ iteration ratio). P-IMLA$_\lambda$ first crosses $\hat{R} < 1.05$ at the second checkpoint $N = 5$ ($\hat{R} = 1.0485$) and reaches $\hat{R} = 0.9990$ at $N = 300$, with per-chain SSE 36.99 ± 0.23; SP-MYULA crosses $\hat{R} < 1.05$ at $N = 10$ ($\hat{R} = 1.0190$) and reaches $\hat{R} = 1.0000$ at $N = 30{,}000$ with per-chain SSE 38.94 ± 0.14. Wall-clocks are 10.1–20.8 min/chain for ours and 137.4 min/chain for SP-MYULA, a $6.5-13.7\times$ difference under the stated iteration allocation. The SP-MYULA/P-IMLA SSE ratio is 1.05 (38.94/36.99). These finite-budget results demonstrate operation beyond Fourier-diagonal models; they do not measure stationary bias or establish a $100\times$ mixing-rate improvement. The TV envelope does not satisfy Theorem 5's smoothness assumptions. At rank $r = 128 \ll d = 4{,}096$, the Nyström M leaves the orthogonal-complement modes essentially un-preconditioned. Image-scale CT/MRI validation with rank matched to the spectrum remains open (Appendix E).

C.7 Formal drift-only coefficient: quartic stress tests

We illustrate the formal drift-only coefficient on the 1-D target $\pi \propto e^{-f(x)}$ with $f(x) = x^2/2 + \varepsilon\, x^4/24$, choosing the observable $\varphi(x) = x^2$. Then $h(x) = f''(x) = 1 + (\varepsilon/2)\, x^2$ and $\mathrm{Cov}_\pi(h, \varphi) = (\varepsilon/2)\, \mathrm{Var}_\pi(x^2)$, so the predicted leading bias coefficient is $c_1 = (\varepsilon/8)\, \mathrm{Var}_\pi(x^2)$ with $\mathrm{Var}_\pi(x^2)$ computed by quadrature. Each plotted point uses one P-IMLA chain with $N = 5\times10^6$ post-burn-in iterations; the figure displays point estimates without error bars. The implementation computes batch-means SEs with batch size 5,000; its separate Richardson coefficient check reports SEs across 32 independent seeds, without bootstrap resampling.

This quartic target is not globally L-smooth, so the experiment illustrates the coefficient outside the stated theorem assumptions; the globally smooth $d = 256$ test is in App. D. The empirical δ-slope of 1.00 (panel b) and the linear ε-fit (panel a) jointly confirm the leading-order coefficient predicted by the BEA expansion of §A.3.

2-D formal-coefficient stress test. The 1-D test illustrates the coefficient when $h(x) = f''(x)$ is non-constant. To examine the formula beyond one dimension, we repeat the test on $f(x_1, x_2) = (x_1^2 + x_2^2)/2 + \varepsilon\, x_1^2 x_2^2/4$, where $h(x) = \mathrm{tr}(\nabla^2 f) = 2 + \varepsilon(x_1^2 + x_2^2)/2$ and $\mathrm{Cov}_\pi(h, \varphi)$ is computed

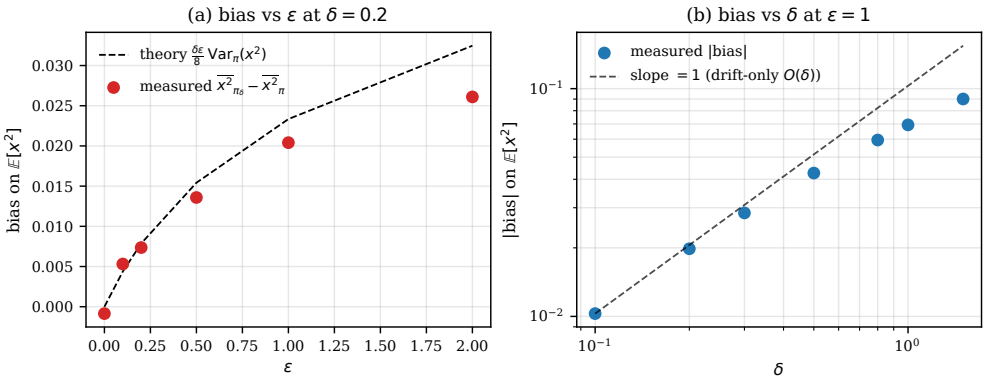

Figure 10: **Formal drift-only coefficient on a 1-D quartic test target. (a)** At fixed δ, measured bias estimates on $\mathbb{E}[x^2]$ are compared with the leading-order prediction $\delta\varepsilon/8 \cdot \mathrm{Var}_\pi(x^2)$; no uncertainty bars are displayed. **(b)** At fixed ε, the bias scales as δ^1 on log–log axes (reference dashed line), consistent with a leading $O(\delta)$ weak bias. Drift-only refers to the differential order of the defect; P-ULA also has a leading $O(\delta)$ weak bias (Appendix A.4).

by 2-D quadrature for $\varphi(x) = x_1^2 + x_2^2$. P-IMLA samples are run for $N = 5 \times 10^6$ post-burn-in iterations.

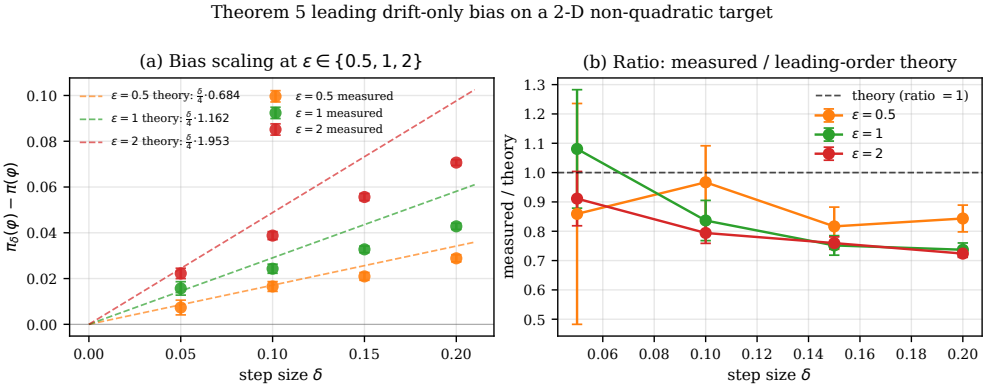

Figure 11: **2-D formal-coefficient check. (a)** Measured bias on $\mathbb{E}[\varphi]$ scales linearly with δ at each ε, with theory line $\delta\ \mathrm{Cov}_\pi(h, \varphi)/4$ from Theorem 5 (dashed). **(b)** Measured/theory ratio across all (ε, δ) tested, clustering near 1; the $\sim 0.72 - 0.84$ values at the largest δ are consistent with a negative $O(\delta^2)$ next-order BEA correction.

The control case $\varepsilon = 0$ yields measured bias indistinguishable from zero (consistent with $h \equiv 2$ giving $\mathrm{Cov}_\pi(h, \varphi) = 0$), and across $\varepsilon \in \{0.5, 1, 2\}$ the measured bias matches the leading-order prediction within ratio $0.72 - 1.08$ across all $\delta \in [0.05, 0.20]$. The polynomial interaction is not globally smooth or globally convex, so this is a formal-coefficient check rather than a verification under Theorem 5's assumptions.

C.8 SK-ROCK δ-stability sweep

We run SK-ROCK [$s = 10$, $\eta = 0.05$, Pereyra et al., 2020] on the same 64×64 Gaussian-deconvolution benchmark used in §C.5 ($\sigma_{\mathrm{blur}} = 2.0$, $\sigma_{\mathrm{noise}} = 0.05$, $\rho_{\mathrm{TV}} = 0.02$, $\lambda = 5 \times 10^{-3}$, $M = I$), and sweep δ over an order-of-magnitude range straddling the theoretical Chebyshev stability boundary $\delta_{\max}^{\mathrm{thy}} = 2s^2(1 + \eta/s^2)/L_M = 0.9955$ (with $L_M = 1 + 1/\lambda = 201$). For each δ the chain is run for 5,500 iterations (500 burn-in, 5,000 posterior); a sample is flagged as diverged if $\|X\|_\infty > 10^6$ or non-finite at any iteration.

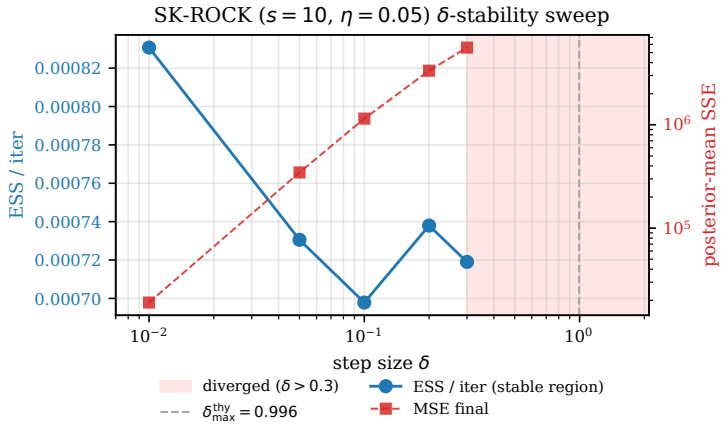

Figure 12: **SK-ROCK δ-stability sweep on a 64×64 Moreau-smoothed posterior.** Stable region (left of red shading): $\delta \leq 0.30$. Diverged region (red): $\delta \geq 0.50$. The empirical stability boundary $\delta^{\text{emp}}_{\text{max}} \approx 0.30$ is roughly $3\times$ tighter than the smooth-case Chebyshev prediction $\delta^{\text{thy}}_{\text{max}} = 0.9955$, indicating that SK-ROCK's published stability margin is *optimistic* (over-predicts the stable region) once the Moreau-smoothed sub-gradient enters the iteration. Even within the stable region, ESS/iter $\in [7\times10^{-4}, 8\times10^{-4}]$ remains essentially flat, an order 10^3 below P-IMLA$_\lambda$'s ESS/iter ≈ 1 at the same λ (cf. §C.5); MSE rises monotonically with δ within the stable region as the chain has not finished a single integrated-autocorrelation time within the 5,000-sample budget.

The takeaway is two-fold: (i) SK-ROCK's stability margin contracts substantially under Moreau-smoothed proximal handling, restricting the practical δ range to roughly $[10^{-2}, 0.3]$ at this λ; (ii) within that range, ESS/iter is bounded by $\sim 10^{-3}$, so the iteration-budget gap to P-IMLA$_\lambda$ identified in the main-text matched-budget benchmark (Table 3) is intrinsic to the explicit-Euler + Chebyshev recursion's mixing rate, not to a sub-optimal step-size choice.

C.9 Term III $\sqrt{\lambda}$ scaling on polyhedral g

The Talagrand-based derivation of Term III is generic in that it uses only the pointwise bound $(g - g^{\lambda,M})(x) \leq \frac{\lambda}{2}(L_g^M)^2$ together with strong convexity of f in the M-norm; it makes no structural assumption on the sub-level geometry of g beyond Lipschitzness. For the polyhedral regularizers that motivate P-IMLA$_\lambda$ in practice (TV, ℓ_1), we expect the bound to be loose because the M-norm subgradient bound collapses across a codimension-1 switching set. We verify this directly on the 1-D target $\pi \propto e^{-f-g}$, $f(x) = x^2/2$, in two contrasting setups: a *symmetric* polyhedral $g_{\text{sym}}(x) = |x|$ and an *asymmetric* polyhedral $g_{\text{asym}}(x) = \max(2x, -x)$. The target distributions π and π_λ are approximated by deterministic quadrature on a log-concentrated grid, and $W_2^M(\pi_\lambda, \pi)$ is evaluated via the 1-D quantile identity $W_2^2(P,Q) = \int_0^1 (F_P^{-1}(u) - F_Q^{-1}(u))^2\, du$ at 17 log-spaced values of $\lambda \in [10^{-4}, 1]$. This calculation has numerical integration and interpolation error, but no Monte Carlo sampling error bars.

Both slopes exceed the generic exponent $1/2$. The difference between the two examples is consistent with cancellation in the symmetric case, but a uniform polyhedral improvement or a universal symmetry-based $O(\lambda^2)$ rate is not proved here.

C.10 Baseline tuning sweep (fairness)

To pre-empt the question of whether the explicit-Euler baselines are fairly tuned, we sweep $\delta \in \{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 5\times10^{-3}, 10^{-2}, 5\times10^{-2}, 10^{-1}, 5\times10^{-1}\}$ for SP-MYULA and MYULA on the same 64×64 benchmark of §C.5. Per δ we run 5,500 iterations (500 burn-in) and report ESS/iter at the diagnostic pixel (Geyer initial-positive sequence) and the post-burn posterior-mean SSE. The body benchmark of §5 uses (rounded down from the largest stable for safety margin) $\delta_{\text{SP-MYULA}} = 9.45\times10^{-3}$, $\delta_{\text{MYULA}} = 3\times10^{-3}$, and $\delta_{\text{SK-ROCK}} = 0.25$ (cf. §C.8).

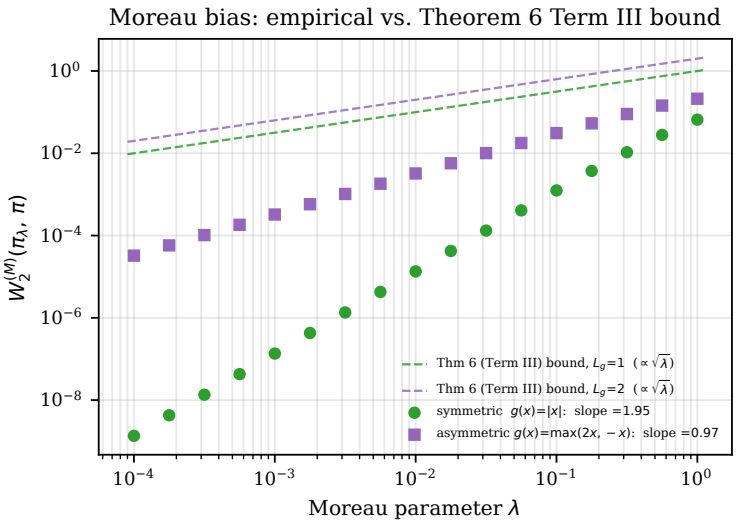

Figure 13: **Empirical verification of Theorem 6 Term III.** The dashed lines show the theoretical $L_g^M\sqrt{\lambda/m_M}$ bounds ($L_g^M = 1$ for g_{sym}, $L_g^M = 2$ for g_{asym}). The empirical $W_2^M(\pi_\lambda, \pi)$ lies *strictly below* the bound at every λ in both setups (maximum $W_2/$bound ratios 0.065 and 0.106). The log-log fitted slopes are 0.97 for the asymmetric case and 1.95 for the symmetric case, both tighter than the theorem's worst-case exponent of 0.5. These two examples show that the generic bound can be conservative; they do not establish an improved rate for all polyhedral penalties.

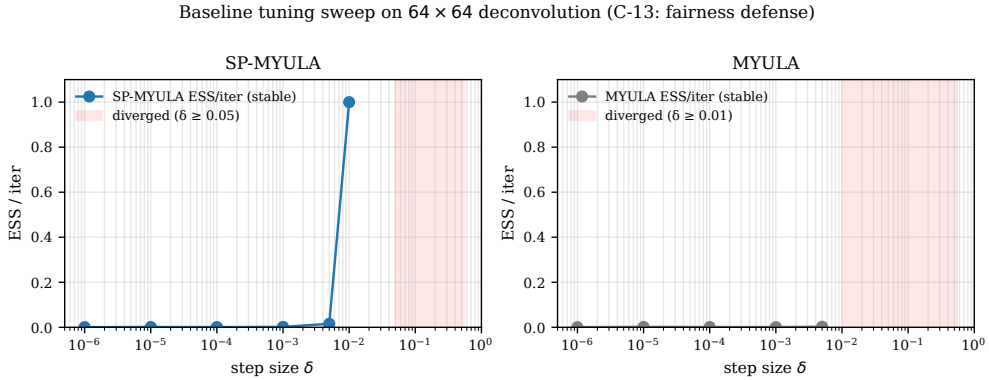

Figure 14: **Baseline tuning sweep on 64×64 deconvolution. Left:** SP-MYULA stable through $\delta = 10^{-2}$ (ESS/iter peaks at 1.0 at the largest stable δ); diverges from $\delta \geq 5\times10^{-2}$. **Right:** MYULA stable only through $\delta = 5\times10^{-3}$, with best ESS/iter $= 2.6\times10^{-3}$ ($\sim$380$\times$ lower than SP-MYULA's best); diverges from $\delta \geq 10^{-2}$. The body benchmark of §5 runs each baseline at its best stable step size from this sweep, controlling step-size choice within this sweep; it does not establish optimal baseline inner-solver tuning.

The headline narrative implication: at imaging-scale curvature (κ_f here is similar to the body κ_f ratio), MYULA's stability boundary forces δ so small that even at its *best* stable choice it requires $\sim400\times$ more iterations (best ESS/iter $\approx 2.6\times10^{-3}$ vs P-IMLA$_\lambda$'s ESS/iter ≈ 1 in §C.5) to extract a single effective sample.

C.11 Scope of empirical comparison: the Langevin family

The matched-budget benchmarks of §5 compare P-IMLA$_\lambda$ against three Langevin-family baselines—MYULA, SK-ROCK, SP-MYULA—under matched preconditioning, Moreau parameter, and noise injection. This isolates the effect we analyse and prove rate theorems for: the integration scheme

(explicit Euler vs. implicit midpoint) on the same underlying Langevin SDE with non-smooth proximal regularisation. The structural obstruction (1) and its midpoint cancellation (2) are properties of this dichotomy; SP-MYULA is the within-family explicit-Euler counterpart of P-IMLA$_\lambda$ on the identical M-norm-Moreau-smoothed potential (Sampler naming conventions, §C), but the finite-budget SSE gap also depends on solver accuracy and initial transients, and does not isolate the integration scheme alone.

Cross-family samplers such as Hamiltonian Monte Carlo [Neal, 2011] and NUTS [Hoffman and Gelman, 2014] accelerate via momentum and adaptive trajectory length on differentiable potentials; on non-smooth targets (TV imaging, ℓ_1 priors) they additionally require Moreau smoothing or specialised non-smooth variants, against which our SP-MYULA comparison is the structural reference. The matrix-preconditioning + implicit-midpoint architecture introduced here is complementary to momentum-based acceleration and can in principle be combined; that combination is follow-up work.

C.12 Independent reference-posterior verification of Proposition 1

To rule out the alternative explanation that SP-MYULA's $\sim 4\times$ larger SSE in Table 3 reflects a chain that has converged to the correct posterior but with looser concentration around the ground truth (rather than a biased stationary distribution), we build an independent MALA reference posterior at the same 256×256 geometric setup and directly compare second-moment statistics across samplers.

Setup. 4 chains of MALA on the smoothed target π_λ^M ($M = \hat{H}$, $\lambda = 5 \times 10^{-3}$) at $\delta_{\mathrm{MALA}} = 1.41 \times 10^{-5}$, warm-started from a P-IMLA$_\lambda$ stationary sample at $N = 2{,}000$ (a warm start reduces the initial transient but does not itself establish convergence), $N_{\mathrm{burn}} = 2 \times 10^5$, $N_{\mathrm{post}} = 10^6$ per chain on the same Pro 6000 hardware as Table 3 ($\sim$7.1 h/chain $\times$ 4-way parallel = $\sim$7.1 h total). Acceptance saturates at 1.00 at this δ; the chain therefore mixes slowly and the mean estimator is noisier than P-IMLA$_\lambda$'s, and local covariance stability does not by itself establish global mixing.

Headline (H1) readout. Pooled across all chains, comparing each sampler's second-moment statistics to the MALA reference:

Table 10: **Imaging-scale reference-posterior comparison** on the 256×256 geometric phantom (App. C.12). $\mathrm{tr}(\mathrm{Cov})$ is the trace of the per-pixel covariance; $\langle \hat{\sigma} \rangle$ is the average per-pixel posterior std. Pooled-sample counts in row 2.

Sampler	$\mathrm{tr}(\mathrm{Cov})$	$\langle \hat{\sigma} \rangle$	ratio to MALA
MALA reference (4 ch. $\times 10^6$)	2.67×10^5	2.017	—
P-IMLA$_\lambda$ TRUE Picard (8 ch. $\times 10^4$)	2.53×10^5	1.965	**0.95**
SP-MYULA (8 ch. $\times 2 \times 10^5$)	2.05×10^6	5.591	**7.68**

Reading. The reported covariance-trace ratios are 0.95 for P-IMLA$_\lambda$ and 7.68 for SP-MYULA relative to this MALA run; the per-pixel standard-deviation ratio is 2.77. These comparisons are consistent with an Euler discretisation effect, but the slowly mixing reference is not a certificate of exact full-image posterior moments. The overdispersed diagnostics and target-exact Gibbs checks in Appendix D provide complementary evidence and explicitly report the remaining limitations.

D Additional bias, convergence and posterior-model checks

D.1 High-dimensional nonquadratic bias coefficient

The quartic examples in Appendix C are stress tests outside the global assumptions of Theorem 5: the one-dimensional quartic has unbounded curvature, and the coupled two-dimensional quartic is not globally convex. Here we use a smooth, globally strongly convex target with bounded Hessian on the periodic 16×16 lattice ($d = 256$):

$$\begin{aligned} f(z) &= \tfrac{1}{2}\|z\|^2 + \alpha \sum_e \log\cosh((Bz)_e), \\ \alpha = 0.7, \qquad M &= I + \gamma B^\top B, \quad \gamma \in \{0, 0.15, 0.35\}, \end{aligned} \tag{46}$$

where B is the edge-incidence matrix of the periodic square grid. Since $\|B\|^2 = 8$, $I \preceq \nabla^2 f \preceq 6.6I$; all derivatives of order two and higher are bounded. The target remains fixed when γ changes. By translation and lattice symmetry,

$$h_M(z) = \operatorname{tr}(M^{-1}) + \alpha w(\gamma) \sum_e \operatorname{sech}^2((Bz)_e), \qquad w(\gamma) = b_e^\top M^{-1} b_e, \tag{47}$$

with a common leverage $w(\gamma)$ for every row $b_e^\top$ of B. For $\varphi = h_M$, the predicted coefficient $\frac{1}{4}\operatorname{Var}_\pi(h_M)$ therefore scales as $w(\gamma)^2$.

We use five step sizes $\delta \in \{0.02, 0.032, 0.05, 0.08, 0.128\}$, 64 chains per step size, 3,000 burn-in steps and 120,000 retained steps per chain. The inner Newton solve uses tolerance 10^{-11}. A Metropolis-corrected M-MALA reference uses 16 chains, 50,000 burn-in steps and 900,000 retained steps each. Reference rank-normalized split-$\hat{R}$ rounds to 1.0000; bulk ESS for h_M is about 10^6. The observables are h_M, $d^{-1}\|z\|^2$, and the normalized edge energy $\|Bz\|^2/(2d)$.

For each observable we fit the measured-to-predicted ratio of bias divided by δ as $c_0 + c_1\delta$, using inverse squared reported Monte Carlo standard errors as weights. Table 11 and Figure 15 report nominal 95% weighted-fit intervals, $c_0 \pm 1.96\,\mathrm{SE}(c_0)$; these are not one-standard-error intervals. All nine contain one. The h_M coefficient changes from 39.7727 to 13.6324 to 5.8883, a factor of 6.75, following the predicted leverage-squared law to within 0.2%. Thus the leading invariant defect depends quantitatively on M even for one fixed target.

Table 11: High-dimensional coefficient check. The slope fits $|\pi_\delta(\varphi) - \pi(\varphi)|$ against δ on logarithmic axes. The last column is the extrapolated measured/predicted ratio with a nominal 95% interval, computed with the same diagonal MCSE weights as Figure 15.

γ	Observable	Slope	Ratio $c_0 \pm 1.96\,\mathrm{SE}$
0	h_M	0.930	1.0005 ± 0.0179
0	$d^{-1}\|z\|^2$	0.928	1.0120 ± 0.0286
0	Edge energy	0.948	1.0024 ± 0.0202
0.15	h_M	0.993	0.9938 ± 0.0341
0.15	$d^{-1}\|z\|^2$	1.018	0.9980 ± 0.0500
0.15	Edge energy	1.007	0.9959 ± 0.0388
0.35	h_M	0.986	0.9989 ± 0.0590
0.35	$d^{-1}\|z\|^2$	1.052	0.9841 ± 0.0817
0.35	Edge energy	0.985	1.0066 ± 0.0674

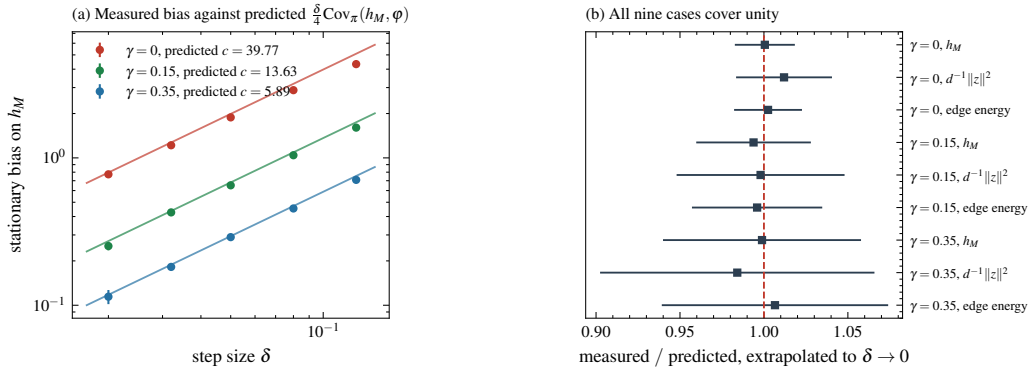

Figure 15: **Bias coefficient on a fixed $d = 256$ target.** Left: measured h_M bias (bars: one MCSE) and the predicted leading term for three preconditioners. Right: extrapolated measured/predicted ratios with nominal 95% weighted-fit intervals. The target is held fixed throughout.

D.2 Overdispersed chains and observable-specific diagnostics

The finite-budget benchmark in Section 5 uses a common zero start and no burn-in. We additionally run 13 geometric-phantom configurations and five cameraman configurations at 256^2, with eight

starts: zero, the observation, the minimizer of f, two noise-perturbed minimizers, two constant fields and a checkerboard. Ordinary runs have $5{,}000$ iterations, discarding 500; the three refined geometric runs have $10{,}000$, discarding $1{,}000$. The refined cameraman run has $5{,}000$, discarding 500. These experiments use RTX 4090 hardware and a baseline $(K_{\mathrm{P}}, K_{\mathrm{in}}) = (8, 15)$, not the $(30, 50)$ solver budget of the headline experiment. Hence their absolute SSEs and timings are not replacements for Table 3.

We compute rank-normalized split-$\hat{R}$, bulk/tail effective sample sizes (ESS), and Monte Carlo standard errors [Vehtari et al., 2021]. The diagnostic implementation is checked on independent draws, AR(1) processes and chains with shifted means. Table 12 reports sampling diagnostics alone; reconstruction SSE and PSNR are reported separately in Table 13.

Table 12: Diagnostics by observable and experimental group. Energy/TV/high-frequency ranges pool only those three scalar observables; they do not bound every observable. The spatial mean retains appreciable autocorrelation, especially in the refined cameraman run. Values are rounded from the archived summaries.

Group	Observables	Maximum $\hat{R}$	Bulk ESS range	Tail ESS range
Geometric (13 configurations)	$f + g$, TV, high-frequency	1.0076	768–18,938	855–33,086
Cameraman (5 configurations)	$f + g$, TV, high-frequency	1.0141	432–19,516	146–28,584
Geometric, $\lambda = 10^{-3}$	Spatial mean	1.0750	71	104
Refined cameraman	Spatial mean	1.1244	42	36

Across these runs classical per-pixel $\hat{R}$ is at most 1.002, but this does not remove the slow spatial-mean mode exposed by the split diagnostics. We therefore treat the refined small-step runs as sensitivity checks, not exact-posterior ground truth or a certificate of convergence for all functionals. In particular, convergence diagnostics measure mixing of the implemented chain; they do not bound its discretization, smoothing or inner-solve bias.

D.3 Separating draw roughness from posterior-mean texture

We vary δ, λ, ρ_{TV} and the inner solve. The baseline is $\delta = 0.141$, $\lambda = 0.005$, $\rho_{\mathrm{TV}} = 0.3$, $\rho_a = 200$ and $(K_{\mathrm{P}}, K_{\mathrm{in}}) = (8, 15)$. The inner-solve arms use $\rho_a = 0.5$ with budgets $(4, 8)$, $(8, 15)$ and $(16, 30)$. The refined runs use $\delta = 0.141/8$, $\lambda = 0.001$, $\rho_a = 0.5$, and $(12, 20)$ on geometric or $(8, 15)$ on cameraman. The Moreau/Picard inner solver and the joint-proximal solver need different ADMM penalties; these settings do not prescribe changing the joint solver's penalty.

Define high-frequency draw energy by summing $|\widehat{x}(k)|^2/d$ over $|k| > n/4$. Reducing λ from 0.005 to 0.001 lowers its chain-averaged value by 79.6% on both phantoms; dividing δ by four changes it by only $+0.42\%$ and $+0.41\%$. The geometric inner-solve arms change it by 1.5–3.3%. The large λ effect is accompanied by a covariance-trace decrease from approximately 2.45×10^5 to 4.99×10^4. The high-frequency modes occupy 0.8039 of the Fourier grid; across all 18 configurations, chain-averaged high-frequency energy divided by $0.8039\,\mathrm{tr}(\mathrm{Cov}\, X)$ lies between 1.185 and 1.242. Its nearly constant proportionality links draw roughness to posterior width, with some excess high-frequency covariance. These are controlled changes of the implemented smoothed chain, not a claim that either value equals the variance of the nonsmoothed posterior.

The visible texture in the posterior *mean* lies in a different band, $32 \le |k| < 58$ at $n = 256$ (approximately 5.6 pixels in wavelength). Let B denote mean-image power in this band divided by ground-truth power in the same band. Doubling ρ_{TV} reduces B from 5.47 to 2.09 on geometric and 3.58 to 1.43 on cameraman, improving PSNR by 3.46 and 2.70 dB, respectively. The cameraman change is visible in Figure 16. Reducing λ alone also reduces this mean-image band power (3.58 to 2.91); the two controls therefore do not act exclusively on disjoint effects. Decreasing δ alone increases it to 5.43, and the refined cameraman run has $B = 6.29$. Thus a less textured output need not be a less biased sampler. Changing ρ_{TV} changes the posterior model itself; the independent target check below distinguishes this effect from a P-IMLA implementation artifact.

Table 13: **Reconstruction quality, separate from convergence diagnostics.** These are pooled posterior-mean SSE, PSNR using the ground-truth intensity range, and mean-image band-power ratio B for the overdispersed $(8, 15)$ baseline protocol. All other parameters are fixed when ρ_{TV} is doubled.

Phantom	ρ_{TV}	SSE	PSNR (dB)	B
Geometric	0.3	544.94	20.80	5.47
Geometric	0.6	245.40	24.27	2.09
Cameraman	0.3	615.73	20.19	3.58
Cameraman	0.6	330.71	22.89	1.43

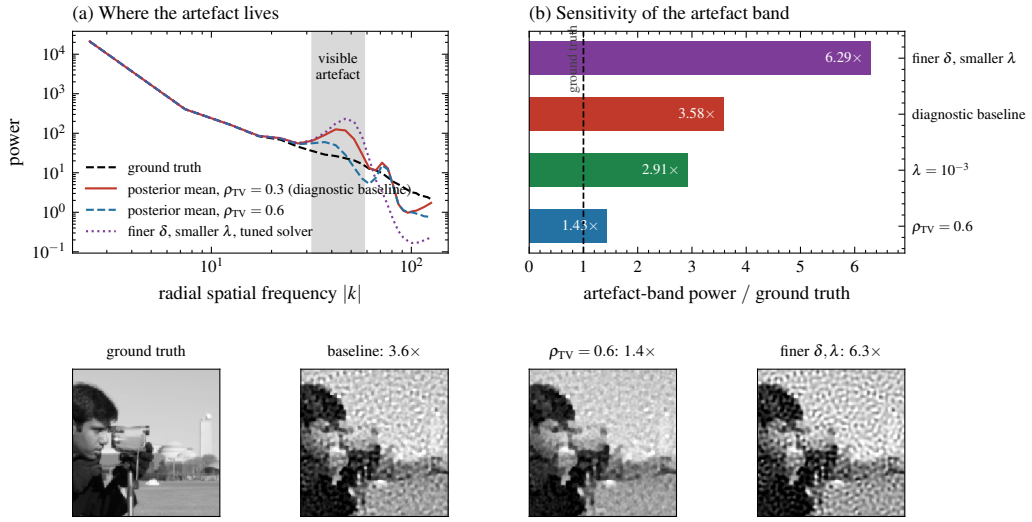

Figure 16: **Cameraman posterior-mean texture and regularization sensitivity.** The shaded spectral band is $32 \leq |k| < 58$. Bottom panels use identical crops and grayscale limits $[0, 1]$. The diagnostic baseline uses $(K_{\mathrm{P}}, K_{\mathrm{in}}) = (8, 15)$ and is distinct from the headline reconstruction protocol. The refined run is a sensitivity check with the residual slow mode reported in Table 12.

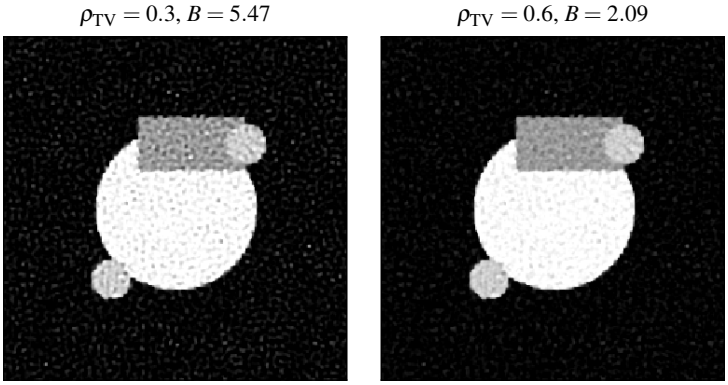

Figure 17: **Geometric posterior means with stronger TV regularization.** The two archived 256^2 reconstructions differ only in ρ_{TV}: eight chains, 5,000 iterations with 500 discarded, $(K_{\mathrm{P}}, K_{\mathrm{in}}) = (8, 15)$, and common grayscale limits $[0, 1]$. Stronger regularization reduces band-power ratio B from 5.47 to 2.09 and improves PSNR from 20.80 to 24.27 dB (Table 13). This controlled diagnostic comparison uses a different inner budget and averaging protocol from the original headline reconstructions.

D.4 Independent exact-target reference at 32^2 and 48^2

To test the nonsmoothed posterior independently of P-IMLA, we use data-augmentation Gibbs sampling for

$$\pi(x) \propto \exp\left[-\frac{\|Ax-y\|^2}{2\sigma^2} - \frac{\eta}{2}\|x\|^2 - \rho_{\mathrm{TV}}\|Dx\|_1\right], \qquad \sigma = 0.05, \quad \eta = 10^{-3}. \tag{48}$$

Each Laplace factor is a Gaussian scale mixture: $z_e \mid \tau_e \sim N(0,\tau_e)$, $\tau_e \sim \mathrm{Exp}(\rho_{\mathrm{TV}}^2/2)$. The conditional precision for x is $C + D^\top \mathrm{diag}(1/\tau_e)D$, where $C = A^\top A/\sigma^2 + \eta I$; its conditional mean solves this precision times the mean equal to $A^\top y/\sigma^2$. We sample this Gaussian by dense Cholesky and update $1/\tau_e$ from its inverse-Gaussian conditional. These transitions use neither Moreau smoothing nor time discretization nor P-IMLA's inner solver. "Exact-target" describes the invariant target and conditional draws; finite Monte Carlo error remains.

The blur is fixed at 2 pixels at both resolutions, and the artifact band is fixed in cycles per pixel, $|k|/n \in [32/256, 58/256)$. These are smaller members of the same model family, not continuum-coarsened copies of the 256^2 problem; their band-power ratios need not equal those at 256^2. All 32^2 runs and the 48^2, $\rho_{\mathrm{TV}} = 0.6$ run use four chains of $4{,}000$ steps, discarding $1{,}000$. The other two 48^2 runs use eight chains of $8{,}000$, discarding $2{,}000$.

Table 14: Exact-target Gibbs reference. B is posterior-mean band power divided by ground-truth band power at the same resolution. The MC floor estimates band power due to finite-chain mean uncertainty from between-chain variation, in the same normalized units.

Resolution	ρ_{TV}	B	MC floor	SSE	PSNR (dB)
32^2	0.15	12.394	0.00195	253.93	6.06
32^2	0.30	6.856	0.00123	89.66	10.58
32^2	0.60	3.028	0.00061	37.95	14.31
48^2	0.15	15.749	0.00040	627.30	5.65
48^2	0.30	9.403	0.00029	205.19	10.50
48^2	0.60	4.257	0.00060	76.35	14.80

The mean-image texture persists under this independently implemented exact-target sampler and decreases with stronger TV regularization (Figure 18, Table 14). We validate the Gaussian draw against the closed-form mean marginal variance (relative error 6.66×10^{-4} over $20{,}000$ draws), and cross-check a matrix-free perturb-and-MAP/CG implementation: at 32^2, the band ratios are 6.8662 and 6.8588, with mean pixel standard deviations differing by 0.064%. Independent Metropolis sampling of the exact potential at 48^2, $\rho_{\mathrm{TV}} = 0.3$, gives $B = 9.364$ versus Gibbs 9.403 and mean pixel standard deviation 1.3338 versus 1.3401. These checks concern distributional summaries, not pointwise identity of finite-run posterior means. They support a model-side source of the texture without asserting that P-IMLA has zero bias relative to this reference.

D.5 What the finite-budget Gaussian statistic measures

For the Gaussian experiment the reported statistic is

$$V_N = \frac{1}{Nd}\sum_{n=1}^{N}(X_n-\mu)^\top H(X_n-\mu), \qquad X_0 = 0.$$

It combines covariance filling and the initial mean offset. In orthonormal Hessian coordinates with eigenvalues h_i and Cayley factors $r_i = (1-\delta h_i/2)/(1+\delta h_i/2)$,

$$\mathbb{E}V_N = \frac{1}{d}\sum_i \left[1 - a_{i,N} + h_i\mu_i^2 a_{i,N}\right], \qquad a_{i,N} = \frac{1}{N}\sum_{n=1}^{N} r_i^{2n}.$$

For unpreconditioned IMLA at $\kappa_f \simeq 4\times10^7$ and $N = 2{,}000$, this gives $48.6329 = 0.6104 + 48.0225$: 98.7% is the unrelaxed initial-mean contribution. Direct simulation gives 48.5974. The stationary variance ratio is exactly one, as required by Gaussian exactness even without preconditioning. The slowest-mode relaxation time is $3{,}162$ iterations, consistent with the method's $\sqrt{\kappa_f}$ dependence. Figure 19 therefore explains a finite-budget mixing advantage, not failure of IMLA's Gaussian invariant covariance.

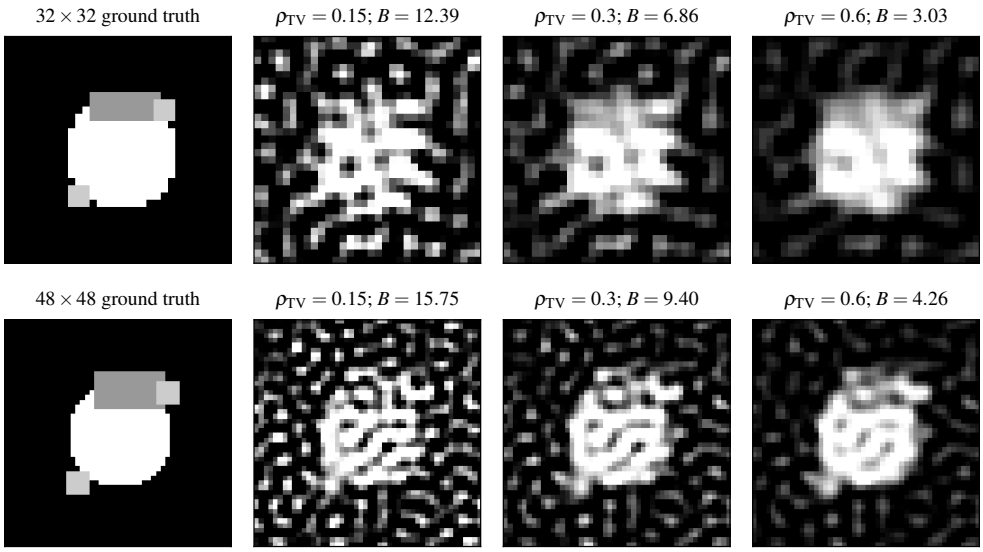

Figure 18: **Posterior means of the nonsmoothed TV model from independent Gibbs sampling.** Top: 32^2; bottom: 48^2. All panels use grayscale limits $[0, 1]$; numerical metrics use the unclipped arrays. The mean-image band-power ratio B decreases as ρ_{TV} increases. This figure reuses the archived reference samples and introduces no new sampling run.

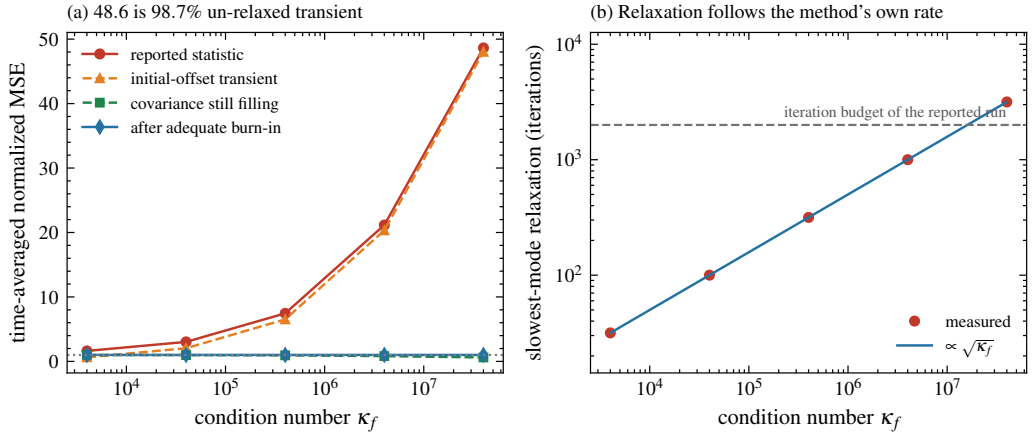

Figure 19: **Initial transients in the Gaussian comparison.** Left: exact decomposition of the cold-start time-averaged normalized MSE into mean-offset and covariance terms. Right: the slowest-mode relaxation time follows $\sqrt{\kappa_f}$. Adequate burn-in recovers a ratio close to one; the invariant ratio is exactly one.

E Limitations and open problems

Four limitations merit explicit acknowledgement. **(i)** For quadratic f with $M = \nabla^2 f$, the bound $\kappa_{\text{eff}}^{\lambda} = 1 + 1/\lambda$ retains the smoothing dependence. All results use fixed M; a position-dependent metric would require an additional drift correction and a separate analysis. **(ii)** The weak BEA result requires global smoothness, polynomial derivative growth and Poisson regularity. The $d = 256$ lattice experiment verifies its coefficient under these assumptions; the earlier quartic checks are only formal illustrations. The L^3 density expansion used for the sharp Wasserstein coefficient remains an additional condition, and is not proved for general targets. Richardson cancellation likewise requires the stronger assumptions of Corollary 12; it is not a TV-posterior guarantee. **(iii)** The headline benchmark uses a Fourier-diagonal forward operator. Appendix C.6 provides a 64×64 random fan-beam test with a rank-128 Nyström preconditioner and conjugate gradients, but image-scale CT/MRI validation remains open. **(iv)** Exact JP and the finite, warm-started ADMM implementation are different objects. Appendix B.1 proves exact-kernel uniqueness for quadratic likelihoods with finite ℓ_1-composite penalties and controls implementation error by certificates; it does not give a uniform certificate for every production $K = 12$ step. The observed small-budget SSE agreement can include error cancellation.

Empirical accuracy also depends on the target, observable and inner-solver budget. Common-start per-pixel diagnostics do not establish convergence of every image functional; the overdispersed checks retain slow mixing in some spatial-mean observables. The original $4\times$ SSE comparison holds at its stated settings, and is sensitive to baseline ADMM accuracy. Small-image Gibbs checks target the original anisotropic-TV posterior, whereas the full-image comparisons use a Moreau surrogate. A distributional error bound does not directly certify reconstruction SSE.

Open directions include a signed analysis of the antithetic regime, position-dependent metrics, and weakly log-concave targets. Our reported ESS implementation caps ESS per iteration at one; antithetic chains can have larger theoretical ESS, so this cap is an estimator convention rather than a universal upper bound.

Other regularisers. Algorithm 1 extends to proper convex lower-semicontinuous penalties with a computable M-prox, replacing the TV-specific inner solve. The quantitative Moreau and joint-kernel bounds additionally require $L_g^M = \sup_{x, s \in \partial g(x)} \|s\|_{M^{-1}} < \infty$. In finite dimensions this includes anisotropic and isotropic TV, wavelet ℓ_1, group and nuclear norms, Huber-TV, and TGV with auxiliary variables solved within the prox. Hard constraints have unbounded normal cones; nonconvex learned priors need a separate analysis. Changing g changes the posterior, and improved reconstruction does not follow from the sampler theory.

Full images and patch splitting. If the likelihood, prior and metric factorize over disjoint patches of size p, the update block-diagonalizes and the FFT cost becomes $O(d \log p)$ instead of $O(d \log d)$; the worst block determines contraction. Squared Wasserstein errors add across blocks, however, so the dimension factor in Theorem 6 still counts the full image. Overlapping or consensus-coupled patches alter the model and require a separate analysis.