Latent-IMH: Efficient Bayesian Inference for Inverse Problems with Approximate Operators

Youguang Chen **George Biros**
Oden Institute for Computational Engineering and Sciences
The University of Texas at Austin

Abstract

We study sampling from posterior distributions in Bayesian linear inverse problems where $\mathbf{A}$, the parameters to observables operator, is computationally expensive. In many applications $\mathbf{A}$ can be factored in a manner that facilitates the construction of a cost-effective approximation $\widetilde{\mathbf{A}}$. In this framework, we introduce Latent-IMH, a sampling method based on the Metropolis-Hastings independence (IMH) sampler. Latent-IMH first generates intermediate latent variables using the approximate $\widetilde{\mathbf{A}}$, and then refines them using the exact $\mathbf{A}$. Its primary benefit is that it shifts the computational cost to an offline phase. We theoretically analyze the performance of Latent-IMH using KL divergence and mixing time bounds. Using numerical experiments on several model problems, we show that, under reasonable assumptions, it outperforms state-of-the-art methods such as the No-U-Turn sampler (NUTS) in computational efficiency. In some cases Latent-IMH can be orders of magnitude faster than existing schemes.

1 INTRODUCTION

In linear inverse problems, we assume that $y = \mathbf{A}x$, where $x \in \mathbb{R}^{d_x}$ are the unknown parameters with prior $p(x)$, and $y \in \mathbb{R}^{d_y}$ are the observed variables corrupted by noise with a known distribution. In the Bayesian setting, we wish to sample from a posterior probability distribution $p(x|y)$. The main challenge is that, depending on the prior $p(x)$, the only practical way to sample from the posterior is by a numerical sampling scheme. Here we present such a scheme for problems with the following structure:

$$u = \mathbf{L}^{-1}\mathbf{B}x, \quad u \in \mathbb{R}^{d_u},\ x \in \mathbb{R}^{d_x}, \tag{1a}$$

$$y = \mathbf{O}u + e, \quad y, e \in \mathbb{R}^{d_y}, \tag{1b}$$

with e being noise with known distribution $q(e)$; and thus, $\pi(x|y) \propto q(y - \mathbf{O}\mathbf{L}^{-1}\mathbf{B}x)\, p(x)$. Here we introduce the *latent* variable u, which in applications is a real but unobserved variable. The linear *observation* operator $\mathbf{O} \in \mathbb{R}^{d_y \times d_u}$ is assumed to be full rank. $\mathbf{L} \in \mathbb{R}^{d_u \times d_u}$ is the *latent* operator, a linear, invertible operator related to the physics or dynamics of the underlying problem. For example, it can be a discretized partial differential operator, an integral operator, or a graph Laplacian. $\mathbf{B} \in \mathbb{R}^{d_u \times d_x}$, the *lifting operator*, is another linear operator that lifts the parameterization of x to the input space of $\mathbf{L}$. We also define the *forward operator* $\mathbf{F} = \mathbf{L}^{-1}\mathbf{B}$. We can easily see that $\mathbf{A} = \mathbf{O}\mathbf{F}$, and equivalently $p(x|y) \propto q(y - \mathbf{A}x)p(x)$.

Bayesian inverse problems that can be formulated using Eq. (1) include electrical impedance tomography (Kaipio et al., 2000), acoustic and electromagnetic scattering (Colton and Kress, 1998), exploration geophysics (Bunks et al., 1995; Sambridge and Mosegaard, 2002), photoacoustic tomography (Saratoon et al., 2013), optical tomography (Arridge and Schotland, 2009), image processing (Chan and Wong, 1998), and many others (Kaipio and Somersalo, 2006; Tarantola, 2005; Vogel, 2002; Ghattas and Willcox, 2021). Inverse problems also appear in graph inference tasks, in which we use partially known vertex features and seek to reconstruct unknown node values or edge weights (Ortega et al., 2018; Dong et al., 2020; Mateos et al., 2019).

As a specific example, consider Eq. (1) in the context of time-harmonic acoustics. In the forward problem, given x, we can compute the sound (pressure) u at every spatial point by solving $\mathbf{L}u = \mathbf{B}x$, where $\mathbf{L} = k^2 I + \Delta$ is the Helmholtz operator (k is the wavenumber) and $\mathbf{B}x = \sum_i s_i x_i$ with s_i being known spatial profiles of

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

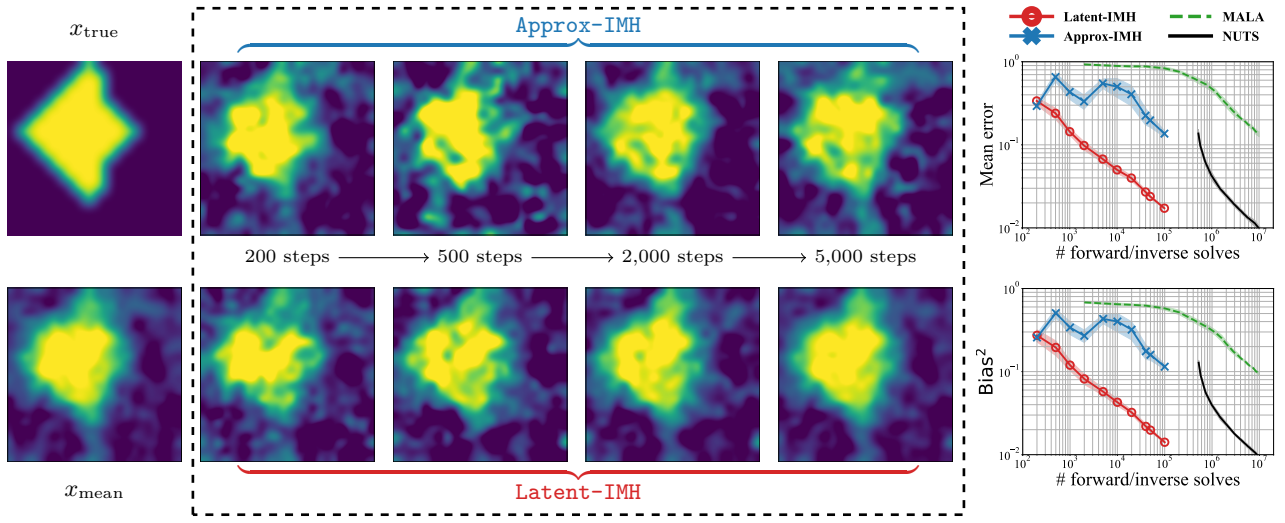

Figure 1: **Reconstructed acoustic source fields.** We reconstruct the sound source field x from partial observations under a non-Gaussian prior, where posterior inference is only possible via sampling. From left to right: x_{true} is the ground truth, and x_{mean} is an accurate posterior mean computed from thousands of samples. The **Approx-IMH** and **Latent-IMH** rows show posterior averages after 200, 500, 2000, and 5000 MCMC steps. The last column shows the relative mean error and squared bias of the second moment between the true posterior and sampled estimates as a function of the number of forward $\mathbf{F}$ and inverse $\mathbf{F}^{-1}$ solves. (MALA: Metropolis Adjusted Langevin sampler; NUTS: No-U-Turn Sampler.) Latent-IMH achieves high accuracy with substantial computational savings: for 10% relative mean error, it requires $\sim 10^3$ solves, Approx-IMH $\sim 10^5$, and MALA/NUTS millions of solves.

sound sources and x_i their unknown amplitudes. In the inverse problem, we wish to estimate x_i from noisy measurements $y = \mathbf{O}u + e$ of the sound field.

Our proposed sampler is based on the existence of an approximate operator $\widetilde{\mathbf{L}}$ such that $\widetilde{\mathbf{L}}^{-1}\mathbf{L} \approx \mathbf{I}$. Examples of $\widetilde{\mathbf{L}}$ include incomplete factorizations (Axelsson, 1994), approximate graph Laplacians (Cohen et al., 2014), truncated preconditioned Krylov solvers (Van der Vorst, 2003), multigrid solvers and coarse grid representations (Henson and Yang, 2002), approximate hierarchical matrices (Rouet et al., 2016), and many others. In these examples $\widetilde{\mathbf{L}}^{-1}$ can be $10\times$ or faster than $\mathbf{L}^{-1}$. Using $\widetilde{\mathbf{L}}$, we define $\widetilde{\mathbf{F}} = \widetilde{\mathbf{L}}^{-1}\mathbf{B}$ and $\widetilde{\mathbf{A}} = \mathbf{O}\widetilde{\mathbf{F}}$.

Now, let us return to our main goal, sampling from $\pi(x \mid y)$. Related work includes Markov chain Monte Carlo (MCMC) methods such as Langevin methods (Girolami and Calderhead, 2011), NUTS (Hoffman et al., 2014), and the Metropolis-adjusted Langevin algorithm (Roberts and Tweedie, 1996), as well as generative models (Song et al., 2021), normalizing flows (Papamakarios et al., 2021), low-rank approximations (Spantini et al., 2015), and many others (Biegler et al., 2010). These methods aim to generate samples of x by targeting the posterior density $q(y - \mathbf{A}x)\, p(x)$.

In Latent-IMH we generate samples from u and then, using the properties of the forward problem, we convert them into samples of x. This approach is motivated by the following observations. (1) We have an approximate operator $\widetilde{\mathbf{L}}$ at our disposal. (2) The observation operator $\mathbf{O}$ is computationally inexpensive. (3) Eq. (1a), can be used to define an implicit distribution $p(u)$ given $p(x)$. Then, using precomputation (that is, before "seeing" y), we can approximate $p(u)$ by $\widetilde{p}(u)$.

The use of approximate forward models has also been extensively studied in multilevel Monte Carlo and multilevel MCMC. For example, (Madrigal-Cianci et al., 2023) propose a multilevel MCMC method based on independent Metropolis–Hastings, where chains at different levels are coupled to construct low-variance estimators for the fine posterior. These approaches are complementary to ours: they focus on variance reduction via coupling, whereas we design improved proposal distributions for a single-level Metropolis–Hastings sampler. In particular, we use the approximate operator to construct an independence proposal targeting the exact posterior.

Another related class of methods is delayed-acceptance MCMC (e.g., two-stage or multi-fidelity MCMC (Peherstorfer et al., 2018)), where proposals from an approximate model are screened before evaluating the exact model. While effective, these methods are inherently local, as proposals depend on the current state. In contrast, our approach uses independent proposals,

enabling precomputation and parallel evaluation, and improving mixing in high-dimensional settings.

In this context our contributions are as follows: ❶ With Latent-IMH we introduce a proposal distribution in which we first sample $u \sim q(y - \mathbf{O}u)\widetilde{p}(u)$, and then we propose x such that $\mathbf{L}u = \mathbf{B}x$ (Section 2). ❷ We theoretically analyze Latent-IMH and compare it with a reference approximate IMH that uses $\widetilde{\mathbf{F}}$ to sample directly from $p(x)$ (Section 3 and Section 4); in particular, we introduce a new theoretical result for the mixing time (Theorem 4.3 and Theorem 4.4). ❸ We conduct numerical experiments to demonstrate the effectiveness of the method (Section 5).

2 PROBLEM SETTING AND METHODS

2.1 Problem setting

Our objective is to sample x from the `Exact posterior` $\pi(x|y)$, given y:

$$\pi(x|y) \propto q(y - \mathbf{A}x)p(x) = q(y - \mathbf{OF}x)p(x). \quad (2)$$

MCMC methods require repeated evaluations of the forward operator $\mathbf{F}$, which makes them costly. To address this, our aim is to leverage the approximate operator $\widetilde{\mathbf{F}}$ to facilitate sampling from $\pi(x|y)$.

As mentioned, we would like to sample an approximate distribution $\pi_a(u)$ and then use it to define a unique x. To define such a unique mapping, we first introduce **two assumptions**. ❶ $d_x \leq d_u$; and ❷ $d_y \leq d_x$. The first assumption is satisfied for nearly all applications that have the Eq. (1)'s structure. The second assumption is also common, but not always true. Under these assumptions, $\mathbf{F}$ is a thin, tall, full-rank matrix. That means that $p(u)$ is degenerate since it is supported in a lower dimensional subspace of $\mathbb{R}^{d_u}$. To address this, we introduce a **reparameterization trick** based on the SVD decomposition of $\mathbf{O}$ (which in general should be cheap to compute as it does not involve $\mathbf{F}$ of $\widetilde{\mathbf{F}}$. Since $d_y < d_x$, we introduce a pseudo-observation operator to accomplish this.

Let $\mathbf{O} = \mathbf{U}[\mathbf{S}\dot{:}\mathbf{0}]\mathbf{V}^\top$ be the full SVD of $\mathbf{O}$, so that $\mathbf{V}$ is unitary. Let $\mathbf{V}_y$ be the first d_y columns of $\mathbf{V}$, which define the rowspace of $\mathbf{O}$. Let $\mathbf{V} = [\mathbf{V}_y, \mathbf{V}_\perp]$. Let $\mathbf{V}_+$ be $d_x - d_y$ columns from $\mathbf{V}_\perp$. Define $\mathbf{V}_x = [\mathbf{V}_y, \mathbf{V}_+]$, $\mathbf{V}_x \in \mathbb{R}^{d_u \times d_x}$. Then we define a new observation operator, $\mathbf{Z} \in \mathbb{R}^{d_y \times d_x}$ by $\mathbf{Z} = \mathbf{OV}_x$. With these definitions, we can rewrite Eq. (1) as,

$$\begin{cases} u = \mathbf{V}_x^\top \mathbf{L}^{-1}\mathbf{B}x, \\ y = \mathbf{Z}u + e. \end{cases} \qquad \begin{cases} \mathbf{F} := \mathbf{V}_x^\top \mathbf{L}^{-1}\mathbf{B}, \\ \widetilde{\mathbf{F}} := \mathbf{V}_x^\top \widetilde{\mathbf{L}}^{-1}\mathbf{B}. \end{cases} \quad (3)$$

Notice that now the redefined operator $\mathbf{F}$ is square, and **assuming w.l.g.** that $\mathbf{B}$ is full rank, then both $\mathbf{F}$ and $\widetilde{\mathbf{F}}$ are full rank and thus invertible. The question is how to choose $\mathbf{V}_+$. We do to maximize conditioning of $\mathbf{V}_x^T \mathbf{L}^{-1}\mathbf{B}$. We give an algorithm to do so in the appendix (Section A).

Remark: Given this transformation, the analysis is done for $\mathbf{F}$ invertible square and square matrix with $\mathbf{O}$ being the modified observation operator. With this understanding, we set $d = d_u = d_x$ and denote the observation operator as $\mathbf{O} \in \mathbb{R}^{d_y \times d}$, so that the observation model is simplified to $y = \mathbf{OF}x + e$.

2.1.1 `Approx-IMH posterior` and `Latent-IMH posterior`

We consider two approximate posterior distributions based on $\widetilde{\mathbf{F}}$. These distributions will be used as importance sampling proposals in the IMH algorithm. Since we expect $\widetilde{\mathbf{A}}$ to be computationally cheaper than $\mathbf{A}$, a natural approach is to substitute $\mathbf{A}$ in Eq. (2) with $\widetilde{\mathbf{A}}$. This yields the following posterior, which we refer to as `Approx-IMH posterior`:

$$\pi_a(x|y) \propto q(y - \widetilde{\mathbf{A}}x)p(x) \quad (4)$$

We use this distribution as the baseline as we consider it represents of methods that use $\widetilde{\mathbf{A}}$ to accelerate sampling from $\pi(x|y)$ Martin et al. (2012).

Now, let us introduce Latent-IMH:

$$\begin{cases} u \sim \pi_a(u|y) \propto q(y - \mathbf{O}u)\widetilde{p}(u) \\ x \leftarrow \mathbf{F}^{-1}u, \end{cases} \quad (5)$$

where $\pi_a(u|y)$ is the posterior of u given y when u has the prior distribution $\widetilde{p}(u)$. It is easy to see that samples from this process follow a new posterior distribution, which we refer to as `Latent-IMH posterior`:

$$\pi_l(x|y) \propto q(y - \mathbf{A}x)\widetilde{p}(\mathbf{F}x). \quad (6)$$

Let us now briefly discuss the choice of $\widetilde{p}(u)$. One natural choice is to formally define $\widetilde{p}(u) = p(\widetilde{\mathbf{F}}^{-1}u)|\det \widetilde{\mathbf{F}}^{-1}|$ when $\widetilde{\mathbf{F}}$ is invertible. Alternatively, one can sample $x \sim p(x)$, set $u = \widetilde{\mathbf{F}}x$, and then construct $\widetilde{p}(u)$ using machine learning techniques such as normalizing flows or variational autoencoders.

2.2 Independence Metropolis-Hastings

In Algorithm 1, we recall the IMH algorithm with $g(x|y)$ being the proposal distribution. Given this template algorithm, we define two variants depending on $g(x|y)$: ❶ `Approx-IMH`, where $g = \pi_a$; or ❷ `Latent-IMH`, where $g = \pi_l$. For these variants, the acceptance ratio becomes

Algorithm 1 Independence Metropolis-Hastings

Input: observations y, number of MH steps k, proposal distribution $g(x|y)$

Output: samples $\{x_i\}_1^{k+1}$ from $p(x|y)$

$x_1 \sim g(x|y)$
for $t = 1$ to k **do**
  Propose $x' \sim g(x|y)$
  Compute acceptance ratio $a(x', x_t) = \min\left\{1, \frac{p(x'|y)}{p(x_t|y)} \frac{g(x_t|y)}{g(x'|y)}\right\}$
  Accept x' with probability $a(x', x_t)$; otherwise accept x_t
end for

- `Approx-IMH`:

$$a_a(x', x_t) = \min\left\{1, \frac{q(y - \mathbf{A}x')/q(y - \widetilde{\mathbf{A}}x')}{q(y - \mathbf{A}x_t)/q(y - \widetilde{\mathbf{A}}x_t)}\right\}. \tag{7}$$

- `Latent-IMH`:

$$a_l(x', x_t) = \min\left\{1, \frac{p(x')/\widetilde{p}(u')}{p(x_t)/\widetilde{p}(u_t)}\right\} \text{ or equivalently,} \tag{8a}$$

$$a_l(x', x_t) = \min\left\{1, \frac{p(x')/p(\widetilde{\mathbf{F}}^{-1}u')}{p(x_t)/p(\widetilde{\mathbf{F}}^{-1}u_t)}\right\}. \tag{8b}$$

Let us make a few remarks. Note that although both Eq. (7) and Eq. (8) are based on $\widetilde{\mathbf{F}}$, their acceptance ratios differ. Both use the exact forward operator. Eq. (7) uses $\mathbf{F}$ in the evaluation of $\mathbf{A}$. Eq. (8) uses $\mathbf{F}^{-1}$ to compute $x' = \mathbf{F}^{-1}u$ in the evaluation of $p(x')$. Regarding the acceptance ratio of `Latent-IMH` in Eq. (8), note that (8a) is derived directly from the `Latent-IMH posterior` expression in (6) and suitable when we have a black box $\widetilde{p}(u)$. Alternatively, (8b) assumes that $\widetilde{\mathbf{F}}$ is a linear invertible operator and the $\det \widetilde{\mathbf{F}}^{-1}$ term cancels out, which makes it more efficient when $\widetilde{\mathbf{F}}^{-1}u$ is inexpensive to compute. Second, notice that, remarkably, the likelihood term involving the noise distribution $q()$ does not appear in (8). It turns out that this property makes it quite accurate when the noise is small.

3 APPROX-IMH POSTERIOR VS. LATENT-IMH POSTERIOR

In this section, we compare `Approx-IMH posterior` and `Latent-IMH posterior` with `Exact posterior` in terms of the expected KL-divergence, where the expectation is taken over the observation y. We assume that both the prior and the noise distributions are Gaussian, as stated in Assumption 3.1. We derive exact expressions for the KL-divergence in a simplified setting (Proposition 3.3) and establish upper bounds for more general cases (Theorem B.2). First we define $\mathbb{D}_a$ and $\mathbb{D}_l$ by

$$\mathbb{D}_a := 2\,\mathbb{E}_y\left[\mathrm{KL}\left(\pi_a(x|y)||\pi(x|y)\right)\right],$$
$$\mathbb{D}_l := 2\,\mathbb{E}_y\left[\mathrm{KL}\left(\pi_l(x|y)||\pi(x|y)\right)\right]. \tag{9}$$

Assumption 3.1. *The prior $p(x) = \mathcal{N}(\mathbf{0}, \mathbf{I})$; the noise $q(e) = \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$.*

Given the Gaussian assumptions in Assumption 3.1, the posteriors for `Exact`, `Approx-IMH` and `Latent-IMH` are all multivariate Gaussian distributions, as summarized in Section 3. Then $\mathbb{D}_a$ and $\mathbb{D}_l$ can be expressed in a closed form as follows:

Proposition 3.2. *Assume that Assumption 3.1 holds. Define $\mathbf{\Delta}_a := \widetilde{\mathbf{A}}^\dagger - \mathbf{A}^\dagger$ and $\mathbf{\Delta}_l := \mathbf{A}_l^\dagger - \mathbf{A}^\dagger$, where $\mathbf{A}^\dagger$,$\widetilde{\mathbf{A}}^\dagger$, and $\mathbf{A}_l^\dagger$ are defined in Section 3. Let $\|\cdot\|_F$ be the Frobenius norm, then*

$$\mathbb{D}_a = \log\frac{|\mathbf{\Sigma}|}{|\mathbf{\Sigma}_a|} + \mathrm{Tr}(\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_a) - d + \frac{1}{\sigma^2}\|\mathbf{A}\mathbf{\Delta}_a\mathbf{A}\|_F^2 + \sigma^2\|\mathbf{\Delta}_a\|_F^2 + \|\mathbf{\Delta}_a\mathbf{A}\|_F^2 + \|\mathbf{A}\mathbf{\Delta}_a\|_F^2. \tag{10}$$

The corresponding expression for $\mathbb{D}_l$ is similar, with the substitutions $\mathbf{\Sigma}_a \to \mathbf{\Sigma}_l$ and $\mathbf{\Delta}_a \to \mathbf{\Delta}_l$.

Diagonal F, $\widetilde{\mathbf{F}}$, and O. To give a more intuitive understanding of the differences between π_a and π_l, consider the case where (1) $\mathbf{F}$ and $\widetilde{\mathbf{F}}$ are diagonal, with $\mathbf{F}_{ii} = s_i$ and $\widetilde{\mathbf{F}}_{ii} = \alpha_i s_i$, $i \in [d]$. (2) $\mathbf{O} = [\mathbf{I}\ \ \mathbf{0}]$. Then, the respective expressions for $\mathbb{D}_a$ and $\mathbb{D}_l$ are given in Proposition 3.3.

Proposition 3.3. *Under Assumption 3.1, diagonal $\mathbf{F}$, $\widetilde{\mathbf{F}}$, $\mathbf{O}$, and defining $\rho_i = {}^{\alpha_i^2 s_i^2 + \sigma^2}/_{s_i^2 + \sigma^2}$, $\zeta_i = {}^{1}/_{(\alpha_i^2 s_i^2 + \sigma^2)^2}$, we obtain the following expressions for expected KL-divergence:*

$$\mathbb{D}_a = -d_y + \sum_{i\in[d_y]} \log\rho_i + \frac{1}{\rho_i} + \sum_{i\in[d_y]} \zeta_i(\alpha_i - 1)^2(\alpha_i s_i^2 - \sigma^2)^2\frac{s_i^2}{\sigma^2}, \tag{11}$$

$$\mathbb{D}_l = -d + \sum_{i=d_y+1}^{d} \alpha_i^2 - \log\alpha_i^2 + \sum_{i\in[d_y]} \log\frac{\rho_i}{\alpha_i^2} + \frac{\alpha_i^2}{\rho_i} + \zeta_i(\alpha_i^2 - 1)^2 s_i^2\sigma^2. \tag{12}$$

It is easy to observe that both $\mathbb{D}_a$ and $\mathbb{D}_l$ increase with increasing spectrum perturbation $|\alpha_i|$ and the dimension d; $\mathbb{D}_a$ also increases as the number of observations d_y increases. If we denote the scale of the

Table 1: Distributions of `Exact`, `Approx` and `Latent` posteriors for normal prior, and normal noise with variance σ^2 (Assumption 3.1).

IMH	Distribution	Mean	Variance	Pseudo-inverse
`Exact`	$\pi(x\|y) = \mathcal{N}(\mu, \mathbf{\Sigma})$	$\mu = \mathbf{A}^\dagger y$	$\mathbf{\Sigma} = \mathbf{I} - \mathbf{A}^\dagger\mathbf{A}$	$\mathbf{A}^\dagger = \mathbf{A}^\top\left(\mathbf{A}\mathbf{A}^\top + \sigma^2\mathbf{I}\right)^{-1}$
`Approx`	$\pi_a(x\|y) = \mathcal{N}(\mu_a, \mathbf{\Sigma}_a)$	$\mu_a = \widetilde{\mathbf{A}}^\dagger y$	$\mathbf{\Sigma}_a = \mathbf{I} - \widetilde{\mathbf{A}}^\dagger\widetilde{\mathbf{A}}$	$\widetilde{\mathbf{A}}^\dagger = \widetilde{\mathbf{A}}^\top\left(\widetilde{\mathbf{A}}\widetilde{\mathbf{A}}^\top + \sigma^2\mathbf{I}\right)^{-1}$
`Latent`	$\pi_l(x\|y) = \mathcal{N}(\mu_l, \mathbf{\Sigma}_l)$	$\mu_l = \mathbf{A}_l^\dagger y$	$\mathbf{\Sigma}_l = \mathbf{F}^{-1}\widetilde{\mathbf{F}}\mathbf{\Sigma}_a\widetilde{\mathbf{F}}^\top\mathbf{F}^{-\top}$	$\mathbf{A}_l^\dagger = \mathbf{F}^{-1}\widetilde{\mathbf{F}}\widetilde{\mathbf{A}}^\dagger$

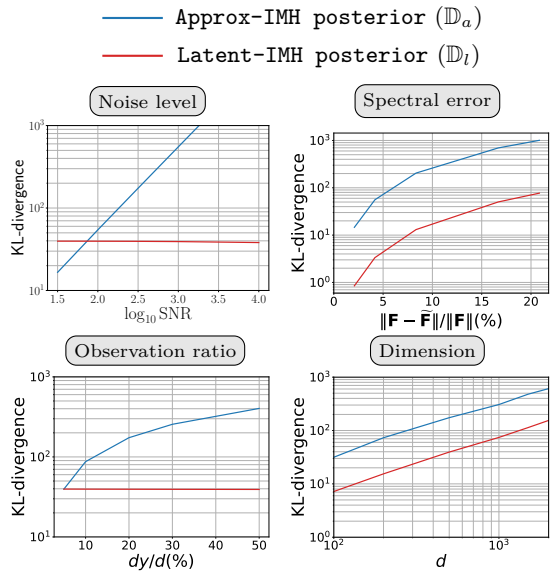

Figure 2: Sensitivity test results for the expected KL-divergence of `Approx-IMH posterior` and `Latent-IMH posterior` relative to `Exact posterior`.

perturbation by $\delta = |\alpha_i - 1|$, the last term of $\mathbb{D}_l$ in (12) scales as $\mathcal{O}(\delta\sigma^2/s_i^2)$, which is significantly smaller than the corresponding term in $\mathbb{D}_a$ (scaling as $\mathcal{O}(\delta^2 s_i^2/\sigma^2)$), especially when the signal strength s_i is large.

General case. In the more general setting where both $\mathbf{F}$ and $\widetilde{\mathbf{F}}$ are symmetric, we assume that $\widetilde{\mathbf{F}}$ closely approximates $\mathbf{F}$ in both its eigenvalues and eigenvectors. Under this assumption, we derive upper bounds for $\mathbb{D}_a$ and $\mathbb{D}_l$ in Theorem B.2 in the appendix. These upper bounds reveal similar conclusion as observed in the diagonal case.

Sensitivity to operator scalings. To further demonstrate the difference between $\mathbb{D}_a$ and $\mathbb{D}_l$, we perform numerical experiments with synthetic data. We adopt the distributions specified in Assumption 3.1 for prior and noise. Define $\mathbf{F} = \mathbf{V}\mathbf{S}\mathbf{V}^\top$ and $\widetilde{\mathbf{F}} = \mathbf{V}\widetilde{\mathbf{S}}\mathbf{V}^\top$, where $\mathbf{V}$ is unitary derived from eigendecomposition of some random matrix, $\mathbf{S}$ as a diagonal matrix with entries $\{1/i^2\}_{i\in[d]}$ and $\widetilde{\mathbf{S}}$ be the diagonal matrix with diagonals of $\{\alpha_i/i^2\}_{i\in[d]}$ (α_i is drawn uniformly from a range around 1). We report both the expected KL divergence and the value of this divergence over the KL divergence between the prior and `Exact posterior`. We compare $\mathbb{D}_a$ and $\mathbb{D}_l$ by examining the impact of four factors: noise level, spectral error (between $\widetilde{\mathbf{F}}$ and $\mathbf{F}$), observation ratio (d_y/d), and problem dimension (d). To quantify the noise level, we use the signal-to-noise ratio (SNR), defined as $\text{SNR} := \mathbb{E}\,\|y\|^2/\mathbb{E}\,\|e\|^2$.

The complete experimental design is summarized in Table 2 (see Section D), with the corresponding absolute KL-divergence results presented in Figure 2 and relative values shown in Figure 7 (see Section D). Our empirical findings align well with the theoretical analysis: both $\mathbb{D}_a$ and $\mathbb{D}_l$ increase with larger spectral errors and higher problem dimensions. However, $\mathbb{D}_a$ exhibits high sensitivity to noise levels and observation ratios, whereas $\mathbb{D}_l$ demonstrates greater robustness to these factors.

4 APPROX-IMH VS. LATENT-IMH

As discussed in the previous section, the KL divergence for `Approx-IMH posterior` and `Latent-IMH posterior` can vary depending on factors like noise level, observation ratio, spectral error, and problem dimensionality. These differences are also likely to have an impact on the performance of `Approx-IMH` and `Latent-IMH`. To explore this, we first establish general mixing time bounds for both methods in Theorem 4.3. We then apply these bounds to a simplified diagonal case in Theorem 4.4 to directly compare their mixing rates.

Our analysis focuses on the log-concave setting, as specified in Assumption 4.1. If the noise $q(e)$ is Gaussian, a strongly log-concave prior $p(x)$ is sufficient to satisfy this assumption.

Assumption 4.1. *The exact posterior $\pi(x|y)$ is strongly log-concave, i.e., $-\log\pi(x|y)$ is m-strongly convex on $\mathbb{R}^d$.*

We denote the modes of the exact and approximate

posteriors as follows:

$$x^* := \arg\max \pi(x|y), \quad x_a^* := \arg\max \pi_a(x|y),$$
$$x_l^* := \arg\max \pi_l(x|y).$$

We define the log-weight functions between the exact posterior with the approximate proposals by

$$w_a(x) := \log \frac{\pi(x|y)}{\pi_a(x|y)}, \qquad w_l(x) := \log \frac{\pi(x|y)}{\pi_l(x|y)}.$$

We assume the log-weight functions are locally Lipschitz, as formalized in Assumption 4.2.

Assumption 4.2. *Log-weight functions are locally Lipschitz. For any $R > 0$, there exist constants $C_a(R) \geq 0$ such that for all $(x, x') \in \mathsf{Ball}(x^*, R)$, $\|w_a(x) - w_a(x')\| \leq C_a(R)\|x - x'\|$. For the function $w_l(x)$, we assume that this condition holds with constant $C_l(R)$.*

The mixing time $\tau_{mix}^a(\epsilon)$ for `Approx-IMH` is defined as the minimum number of steps needed for the Markov chain to reach a total variation (TV) distance below ϵ from the target distribution $\pi(x|y)$. That is,

$$\tau_{mix}^a(\epsilon) := \inf\{n \in \mathbb{N} : \|gP_a^n - \pi\|_{\mathrm{TV}} \leq \epsilon\}, \qquad (13)$$

where P_a is the IMH transition kernel with $\pi_a(x|y)$ as proposal, and $\|f(x)\|_{\mathrm{TV}} := 1/2 \int_x |f|$. The mixing time for `Latent-IMH`, denoted by $\tau_{mix}^l(g, \epsilon)$, is defined similarly using P_l.

In Theorem 4.3, we establish mixing time bounds for both `Approx-IMH` and `Latent-IMH` under the log-concavity and local Lipschitz assumptions. Following Dwivedi et al. (2019); Grenioux et al. (2023), we prove these results in Section C.

Theorem 4.3. *Let $\epsilon \in (0, 1)$. Assume that $\pi_a(x)$ and $\pi_l(x)$ are both $\beta-$warm start with respect to $\pi(x|y)$ (i.e., for any Borel set $\mathcal{E}$, $\pi_a(\mathcal{E}) \leq \beta\pi(\mathcal{E})$). Assume that Assumption 4.1 and Assumption 4.2 hold. For Assumption 4.2, suppose that the Lipschitz conditions hold with $C_a(R_a) \leq \log 2\sqrt{m}/32$ and $C_l(R_l) \leq \log 2\sqrt{m}/32$, where*

$$R_a = \max\left(\sqrt{\frac{d}{m}} r\left(\frac{\epsilon}{17\beta}\right), \sqrt{\frac{d}{m}} r\left(\frac{\epsilon}{272\beta}\right) + \|x^* - x_a^*\|\right) \qquad (14)$$

$$R_l = \max\left(\sqrt{\frac{d}{m}} r\left(\frac{\epsilon}{17\beta}\right), \sqrt{\frac{d}{m}} r\left(\frac{\epsilon}{272\beta}\right) \frac{1}{\sigma_{\min}(\widetilde{\mathbf{F}}^{-1}\mathbf{F})} + \|x^* - x_l^*\|\right), \qquad (15)$$

where $r(\cdot)$ is a constant defined in Section C, $\sigma_{\min}(\cdot)$ is the smallest singular value of the matrix. Then the mixing time for `Approx-IMH` has

$$\tau_{mix}^a(\epsilon) \leq 128 \log\left(\frac{2\beta}{\epsilon}\right) \max\left(1, \frac{128^2 C_a^2(R_a)}{(\log 2)^2 m}\right). \qquad (16)$$

The mixing time for `Latent-IMH`, $\tau_{mix}^l(\epsilon)$, satisfies a similar bound with $C_a(R_a)$ replaced by $C_l(R_l)$.

Mixing time in the diagonal case. We now use Theorem 4.3 to compare `Approx-IMH` and `Latent-IMH` in a simple case, where we assume prior and noise satisfying Assumption 3.1 and $\mathbf{F}$, $\widetilde{\mathbf{F}}$, $\mathbf{O}$ are all diagonal as in Proposition 3.3.

Theorem 4.4. *The definitions and assumptions in Theorem 4.3 and Assumption 3.1 are true. Then for large d the mixing times for `Approx-IMH` and `Latent-IMH` scale as*

$$\tau_{mix}^a(\epsilon) \sim \frac{d}{m^2} \max_{i \in [d_y]} \left(1 - \alpha_i^2\right)^2 \frac{s_i^4}{\sigma^4},$$
$$\tau_{mix}^l(\epsilon) \sim \frac{d}{m^2} \max_{i \in [d]} \left(1 - \frac{1}{\alpha_i^2}\right)^2. \qquad (17)$$

By Theorem 4.4, when the perturbation α_i is fixed, the mixing time for `Approx-IMH` scales as $\mathcal{O}(d^2 d_y^2 \|\mathbf{A}\|^4/\sigma^4)$, while the mixing time for `Latent-IMH` scales as $\mathcal{O}(d^2)$. This implies that `Approx-IMH` can exhibit significantly longer mixing times when the noise level is low (i.e., $\|\mathbf{A}\|/\sigma$ is large) or when the proportion of observations is high (i.e., d_y/d is large).

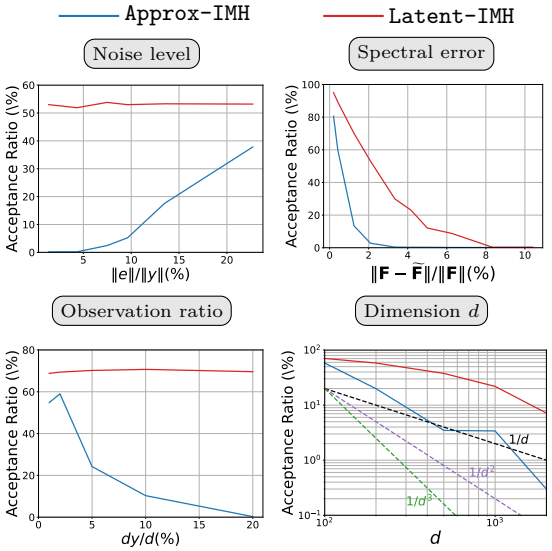

Figure 3: Acceptance ratio of `Approx-IMH` and `Latent-IMH` for different problem scenarios.

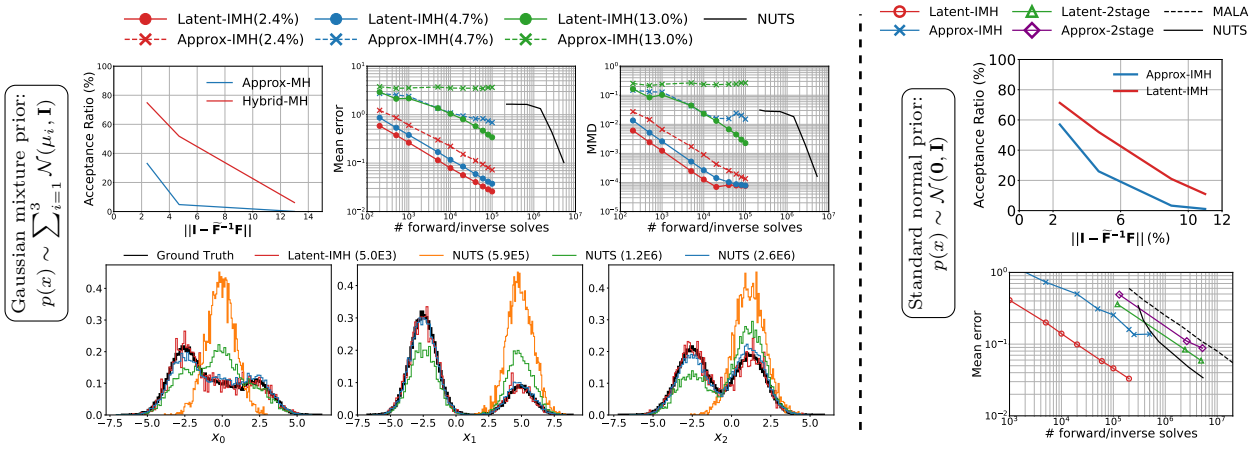

Figure 4: Sample efficiency comparison for Gaussian priors: results averaged over 5 independent runs. **Left (Gaussian mixture prior):** Numbers in parentheses in the top legend indicate spectral error $\|\mathbf{I} - \widetilde{\mathbf{F}}^{-1}\mathbf{F}\|_2$; numbers above histogram plots denote the total number of forward and inverse solves of $\mathbf{F}$ required by each sampler. **Right (standard normal prior):** The approximate operator in the mean error plot has spectral error of 4.7%.

5 NUMERICAL EXPERIMENTS

Our numerical experiments aim to investigate the following questions: How do `Latent-IMH` and `Approx-IMH` compare in terms of acceptance rate and sampling efficiency? How do they perform relative to local MCMC methods such as NUTS (Hoffman et al., 2014), MALA (Roberts and Rosenthal, 1998), and the two-stage multi-fidelity MCMC algorithm (Peherstorfer et al., 2018)? Finally, how sensitive are these methods to variations in problem structure and parameter settings?

Sensitivity tests (acceptance ratio). We first compare acceptance ratios of `Approx-IMH` and `Latent-IMH` using the same setting as in Section 3 (see Table 2). Results in Figure 3 show that as spectral error or dimensionality increases, acceptance ratios decline for both methods, but much more sharply for `Approx-IMH`. Moreover, `Approx-IMH` is highly sensitive to noise level and observation ratio, while `Latent-IMH` remains robust.

Experimental setup for efficiency tests. We evaluate sample quality using three metrics: ❶ relative sample mean error, ❷ squared bias of the second moment (Hoffman et al., 2019), ❸ maximum mean discrepancy (MMD) (see Section D). Since forward and inverse solves of the exact operator $\mathbf{F}$ are computationally expensive, we measure cost by the number of such solves: each `Latent-IMH` step requires an inverse solve of $\mathbf{F}^{-1}$, while each `Approx-IMH` step requires a forward solve of $\mathbf{F}$ to compute acceptance ratios.

Baselines include NUTS and MALA. For NUTS, we use 500-1000 warm-up steps with target acceptance 45%; for MALA, step size is tuned to yield ∼50% acceptance. We assume Gaussian noise with relative level $\|e\|/\|y\| \in [5\%, 15\%]$, and report averages over 5 independent runs for each test (see details in Section D).

Gaussian and Gaussian mixture priors We consider $d = 500$, $d_y = 50$, with synthetic $\mathbf{F}$ and $\widetilde{\mathbf{F}}$ as in Section 3, and test three priors: ❶ a standard normal $\mathcal{N}(\mathbf{0}, \mathbf{I})$, ❷ an ill-conditioned Gaussian $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ with condition number ∼1000, ❸ a three-component Gaussian mixture $\sum_{i=1}^{3} \mathcal{N}(\mu_i, \mathbf{I})$.

Results for the standard normal and Gaussian mixture are in Figure 4, while the ill-conditioned case appears in Figure 8 (Section D). For the test of standard normal prior, we include NUTS and MALA as baselines, as well as the two-stage delayed acceptance MCMC method (Peherstorfer et al., 2018), which also leverages approximate operator to improve sampling efficiency. In this method, proposals are first generated via local updates using the approximate model, and acceptance is decided using the exact model in the second stage. We use MALA for the first stage, and acceptance is then determined using the exact model. We employ MALA in the first stage, denoting the variant using `Approx-IMH posterior` as Approx-2stage and the variant using `Latent-IMH posterior` as Latent-2stage. We include more comparison with delayed-acceptance MCMC in Section D.2.

From Figure 4, `Latent-IMH` consistently outperforms other methods in sampling efficiency, with the advantage being particularly pronounced for the Gaussian

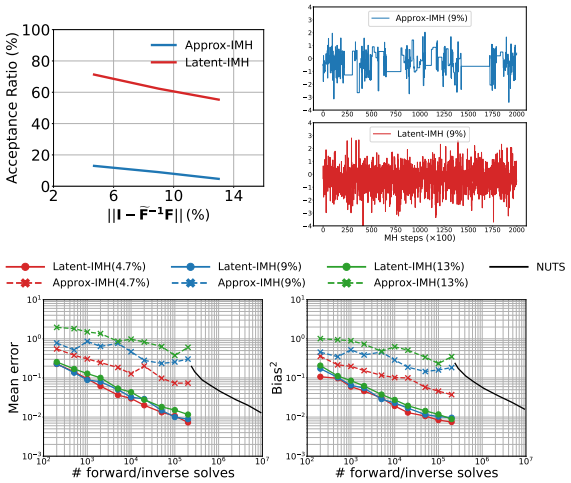

Figure 5: Results for Laplace prior test with normalizing flow. All results averaged over 5 independent runs. Numbers in parentheses indicate spectral error $\|\mathbf{I} - \widetilde{\mathbf{F}}^{-1}\mathbf{F}\|_2$.

mixture prior. In the histogram plots (bottom row of Figure 4), local MCMC methods such as NUTS exhibit poor mixing in multimodal settings due to potential barriers between modes.

Laplace prior with normalizing flow. We consider a setting with $d = 50$ and $d_y = 10$, where x follows an independent Laplace prior $p(x) = 2^{-d}\exp(-|x|1)$. The operators $\mathbf{F}$ and $\widetilde{\mathbf{F}}$ are constructed as in the Gaussian prior experiments. In `Latent-IMH`, we train a normalizing flow to approximate $\widetilde{p}(u)$ using samples $x_i, \widetilde{\mathbf{F}}x_i i = 1^N$, and then apply NUTS to sample u (Eq. (5)), with both log probabilities and gradients provided by the flow. Hence, the approximation arises from both $\widetilde{\mathbf{F}}$ and the normalizing flow, while the acceptance ratio is computed via Eq. (8a). As shown in Figure 5, `Latent-IMH` consistently outperforms both `Approx-IMH` and NUTS. Although `Latent-IMH` and NUTS achieve similar convergence rates, both substantially outperform `Approx-IMH`, which frequently rejects proposals, leading to stagnation in its trace plots compared with the more dynamic behavior of `Latent-IMH`.

Graph Laplacian test with PCG approximation. We consider problem Eq. (1) with $d_u = 8,000$, $d_x = 800$ $(= 0.1d_u)$, and $d_y = 160$ $(= 0.2d_x)$, and standard normal for the prior of x. The Laplacian $\mathbf{L}$ constructed rom a six-nearest neighbor graph on a 20^3 lattice in the unit cube; $\mathbf{B}$ and $\mathbf{O}$ are random Gaussian matrices. Since $\mathbf{F}$ is not square, we construct $\mathbf{V}_x$ and $\mathbf{Z}$ as described in Section 2 to transform (1) into (3). To approximate $\mathbf{L}^{-1}x$, we compute a randomized Cholesky factorization $\mathbf{L} \approx \mathbf{G}\mathbf{G}^\top$ (Chen et al., 2021), then solve using preconditioned conjugate gradient (PCG) with $\mathbf{G}\mathbf{G}^\top$ as preconditioner. PCG tolerance controls approximation accuracy: smaller tolerance requires more iterations and higher cost (left plot, Figure 6). Results in Figure 6 show `Latent-IMH` consistently achieves higher acceptance rates than `Approx-IMH` for the PCG tolerances (second plot). The rightmost plots show that `Latent-IMH` outperforms both `Approx-IMH` and NUTS in terms of the relative mean error and the MMD error. In particular, `Latent-IMH` at a tolerance of 0.1 (solid red) achieves even better performance than `Approx-IMH` at a much stricter tolerance of 5×10^{-4} (purple dashed), despite the latter having an even higher acceptance rate of 34.2%.

Scattering problem. We consider Bayesian inference for a 2D scattering problem introduced in Section 1. We study problem Eq. (1), where $\mathbf{L}$ is the Helmholtz operator and $\mathbf{B}$ is the lifting operator. Operators are constructed via a finite difference scheme on a fine grid ($n_u \times n_u$ nodes). For the approximation operator, we first construct coarse-grid operators $\widetilde{\mathbf{L}}$ and $\widetilde{\mathbf{B}}$ (on $\widetilde{n}_u \times \widetilde{n}_u$ nodes) and then map them to the fine grid using a prolongation operator $\mathbf{P}$. The resulting approximation to $\mathbf{L}^{-1}\mathbf{B}$ is $\mathbf{P}\widetilde{\mathbf{L}}^{-1}\widetilde{\mathbf{B}}$. We consider multiple events with multiple observations and place a total variation prior on x, i.e., $p(x) \propto \exp(-\lambda \mathrm{TV}(x))$ (details in Section D). In our experiments, we consider 4 events with $n_x = 24$, $n_u = 128$, and $\widetilde{n}_u = 60$, resulting in dimensions $d_u = 165{,}536$, $\widetilde{d}_u = 14{,}400$ $(\approx 0.2d_u)$, $d_x = 576$, and $d_y = 128$ $(\approx 0.2d_x)$. Results in Figure 1 show that `Latent-IMH` efficiently samples the posterior using the cheaper approximate operator.

6 CONCLUSIONS

The advantage of the proposed methodology is that it is simple to implement and can be effective in a wide range of scenarios, including multimodal distributions, where multimodality is related to the prior $p(x)$. The reparameterization trick in Section 2 required an SVD of the observation operator that, in general, is inexpensive to apply. In summary, we shift computational costs to constructing an approximation $\widetilde{p}(u)$. While this step can be costly, it can be performed offline and reused across multiple problem instances. Our results show that the method is not universally preferable-its benefit depends on the accuracy and cost of $\widetilde{\mathbf{F}}$, the problem dimension, and the application context, where alternatives such as NUTS, or other standard methods may be more suitable.

Limitations. Latent-IMH is most relevant for large problems where memory or computational constraints prevent forming a dense $\mathbf{F}$. Constructing $\mathbf{F}$ requires d_x applications of $\mathbf{L}^{-1}$ and $d_x d_u$ storage, and applying the dense $\mathbf{F}$ to a vector has similar cost. Note

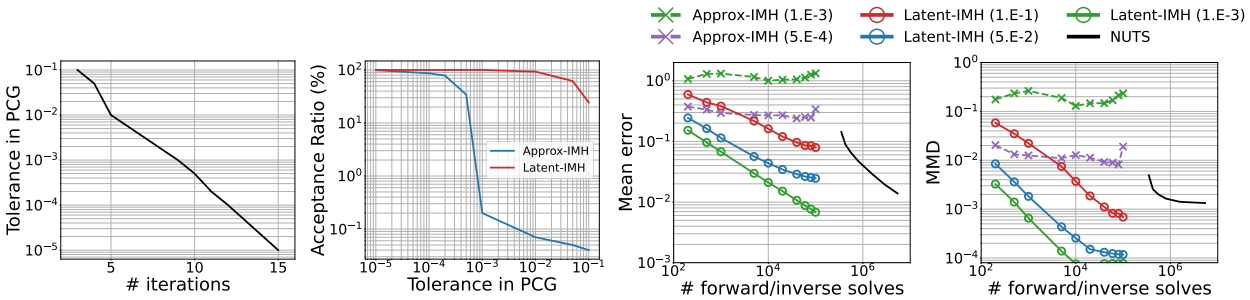

Figure 6: Results for Graph Laplacian test with PCG approximation. Left plot shows PCG iterations vs. tolerance parameter. All results averaged over 5 independent runs. Numbers in parentheses of the legend indicate PCG tolerance (smaller values yield more iterations and higher accuracy).

that in many applications $\mathbf{O}, \mathbf{L}, \mathbf{B}$ are sparse, allowing matrix-free application of $\mathbf{F}$ at $O(d_u)$ cost. These tradeoffs determine the practicality of Latent-IMH and are problem-dependent. A clear limitation of Latent-IMH is the assumption of linearity in $\mathbf{A}$; generalization to nonlinear $\mathbf{L}$ or $\mathbf{B}$ is possible if $\mathbf{L}^{-1}$ yields unique solutions, but handling nonlinear $\mathbf{O}$ is more challenging since the reparameterization trick is no longer applicable.

Another limitation is the requirement that $d_y \leq d_x$. For linear operators, our method can be generalized, but it is unclear whether it will be computationally beneficial. The approximation $\widetilde{p}(u)$ can have several components. It comprises the error due to $\widetilde{\mathbf{F}}$ and possibly a machine learning algorithm to reconstruct it for the sample. From a theoretical perspective, we would like to connect an error between $p(u)$ and $\widetilde{p}(u)$ to the overall performance of the method. Overcoming all these limitations is ongoing work in our group.

Acknowledgements

This material is based upon work supported by NSF award OAC 22042261; and Cooperative Agreement 2421782 and the Simons Foundation award MPS-AI-00010515 (NSF-Simons AI Institute for Cosmic Origins—CosmicAI, `https://www.cosmicai.org/`); by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Applied Mathematics program, Mathematical Multifaceted Integrated Capability Centers (MMICCS) program, under award number DE-SC0023171; by the U.S. Department of Energy, National Nuclear Security Administration Award Number DE-NA0003969; and by the U.S. National Institute on Aging under award number R21AG074276-01. Any opinions, findings, and conclusions or recommendations expressed herein are those of the authors and do not necessarily reflect the views of the DOE, NIH, and NSF. Computing time on the Texas Advanced Computing Centers Stampede system was provided by an allocation from TACC and the NSF.

References

Simon R Arridge and John C Schotland. Optical tomography: forward and inverse problems. *Inverse Problems*, 25(12):123010, 2009.

Owe Axelsson. *Iterative Solution Methods*. Cambridge University Press, 1994.

Lorenz Biegler, George Biros, Omar Ghattas, Matthias Heinkenschloss, David Keyes, Bani Mallick, Luis Tenorio, Bart van Bloemen Waanders, Karen Willcox, and Youssef Marzouk. *Large-scale inverse problems and quantification of uncertainty*, volume 712. Wiley Online Library, 2010.

Carey Bunks, Fatimetou M. Saleck, S. Zaleski, and Guy Chavent. Multiscale seismic waveform inversion. *Geophysics*, 50:1457–1473, 1995.

Tony F Chan and Chiu-Kwong Wong. Total variation blind deconvolution. *IEEE transactions on Image Processing*, 7(3):370–375, 1998.

Chao Chen, Tianyu Liang, and George Biros. Rchol: Randomized cholesky factorization for solving sdd linear systems. *SIAM Journal on Scientific Computing*, 43(6):C411–C438, 2021. doi: 10.1137/20M1380624. URL `https://doi.org/10.1137/20M1380624`.

Michael B Cohen, Rasmus Kyng, Gary L Miller, Jakub W Pachocki, Richard Peng, Anup B Rao, and Shen Chen Xu. Solving sdd linear systems in nearly m log1/2 n time. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 343–352, 2014.

David Colton and Rainer Kress. *Inverse Acoustic and Electromagnetic Scattering Theory, 2nd Edition*. Applied Mathematical Sciences. Springer, 1998.

Ben Cousins and Santosh Vempala. *A Cubic Algorithm for Computing Gaussian Volume*, pages 1215–1228. doi: 10.1137/1.9781611973402.

90. URL https://epubs.siam.org/doi/abs/10.1137/1.9781611973402.90.

Xiaowen Dong, Dorina Thanou, Laura Toni, Michael Bronstein, and Pascal Frossard. Graph signal processing for machine learning: A review and new perspectives. *IEEE Signal processing magazine*, 37 (6):117–127, 2020.

Raaz Dwivedi, Yuansi Chen, Martin J Wainwright, and Bin Yu. Log-concave sampling: Metropolis-hastings algorithms are fast. *Journal of Machine Learning Research*, 20(183):1–42, 2019.

Omar Ghattas and Karen Willcox. Learning physics-based models from data: perspectives from inverse problems and model reduction. *Acta Numerica*, 30: 445–554, 2021.

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

Louis Grenioux, Alain Oliviero Durmus, Éric Moulines, and Marylou Gabrié. On sampling with approximate transport maps. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Van Emden Henson and Ulrike Meier Yang. Boomeramg: a parallel algebraic multigrid solver and preconditioner. *Appl. Numer. Math.*, 41(1):155–177, 2002. ISSN 0168-9274.

Matthew D Hoffman, Andrew Gelman, et al. The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1):1593–1623, 2014.

Matthew D. Hoffman, Pavel Sountsov, Joshua V. Dillon, Ian Langmore, Dustin Tran, and Srinivas Vasudevan. Neutra-lizing bad geometry in hamiltonian monte carlo using neural transport. *arXiv: Computation*, 2019. URL https://api.semanticscholar.org/CorpusID:88517731.

J. P. Kaipio, V. Kolehmainen, E. Somersalo, and M. Vauhkonen. Statistical inversion and Monte Carlo sampling methods in electrical impedance tomography. *Inverse Problems*, 16(5):1487–1522, 2000.

Jari Kaipio and Erkki Somersalo. *Statistical and computational inverse problems*, volume 160. Springer Science & Business Media, 2006.

Juan P. Madrigal-Cianci, Fabio Nobile, and Raúl Tempone. Analysis of a class of multilevel markov chain monte carlo algorithms based on independent metropolisâĂŞhastings. *SIAM/ASA Journal on Uncertainty Quantification*, 11(1):91–138, 2023. doi: 10.1137/21M1420927. URL https://doi.org/10.1137/21M1420927.

James Martin, Lucas C Wilcox, Carsten Burstedde, and Omar Ghattas. A stochastic Newton MCMC method for large-scale statistical inverse problems with application to seismic inversion. *SIAM Journal on Scientific Computing*, 34(3):A1460–A1487, 2012.

Gonzalo Mateos, Santiago Segarra, Antonio G Marques, and Alejandro Ribeiro. Connecting the dots: Identifying network structure via graph signal processing. *IEEE Signal Processing Magazine*, 36(3): 16–43, 2019.

Antonio Ortega, Pascal Frossard, Jelena Kovacevic, José M. F. Moura, and Pierre Vandergheynst. Graph signal processing: Overview, challenges, and applications. *Proceedings of the IEEE*, 106(5):808–828, 2018. doi: 10.1109/JPROC.2018.2820126.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

Benjamin Peherstorfer, Karen Willcox, and Max Gunzburger. Survey of multifidelity methods in uncertainty propagation, inference, and optimization. *SIAM Review*, 60(3):550–591, 2018. doi: 10.1137/16M1082469. URL https://doi.org/10.1137/16M1082469.

Gareth O Roberts and Jeffrey S Rosenthal. Optimal scaling of discrete approximations to langevin diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1):255–268, 1998.

Gareth O. Roberts and Richard L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. 1996.

François-Henry Rouet, Xiaoye S. Li, Pieter Ghysels, and Artem Napov. A distributed-memory package for dense hierarchically semi-separable matrix computations using randomization. *ACM Transactions in Mathematical Software*, 42(4):27:1–27:35, June 2016. ISSN 0098-3500. doi: 10.1145/2930660. URL http://doi.acm.org/10.1145/2930660.

Malcolm Sambridge and Klaus Mosegaard. Monte carlo methods in geophysical inverse problems. *Reviews of Geophysics*, 40(3):3–1, 2002.

T Saratoon, T Tarvainen, BT Cox, and SR Arridge. A gradient-based method for quantitative photoacoustic tomography using the radiative transfer equation. *Inverse Problems*, 29(7):075006, 2013.

Yang Song, Liyue Shen, Lei Xing, and Stefano Ermon. Solving inverse problems in medical imaging with score-based generative models. *arXiv preprint arXiv:2111.08005*, 2021.

Alessio Spantini, Antti Solonen, Tiangang Cui, James Martin, Luis Tenorio, and Youssef Marzouk. Optimal low-rank approximations of bayesian linear inverse problems. *SIAM Journal on Scientific Computing*, 37(6):A2451–A2487, 2015.

Albert Tarantola. *Inverse problem theory and methods for model parameter estimation*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2005. ISBN 0-89871-572-5.

Henk A Van der Vorst. *Iterative Krylov methods for large linear systems*. Number 13. Cambridge University Press, 2003.

C.R. Vogel. *Computational methods for inverse problems*, volume 23. SIAM, 2002.

7 CHECKLIST

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

APPENDIX

APPENDIX OUTLINE

This appendix is organized as follows:

- In Section A, we derive the transformation from the rectangular matrix $\mathbf{F}$ to an invertible square matrix, corresponding to the transition from (1) to (3).
- In Section B, we begin by establishing KL-divergence bounds for `Approx-IMH posterior` and `Latent-IMH posterior` (Theorem B.2) under a more general assumption (Assumption B.1) than the diagonal case considered in Section 3. We then provide detailed proofs of Proposition 3.2, Proposition 3.3, and Theorem B.2.
- In Section C, we present proofs for the mixing time bounds of `Approx-IMH` and `Latent-IMH`, as stated in Theorem 4.3 and Theorem 4.4.
- In Section D, we include experimental details of the experiments in Section 3 and Section 5.

A REPARAMETERIZATION TRICK FOR RECTANGULAR F

Let $\mathbf{O} \in \mathbb{R}^{d_y \times d_x}$ have reduced SVD $\mathbf{O} = \mathbf{U}\mathbf{S}\mathbf{V}_y^\top$. We claim that the construction of $\mathbf{V}_x$ and $\mathbf{Z}$ as specified in Proposition A.1 guarantees that the relationship between x and y given in Eq. (1) is equivalent to the transformed system in Eq. (3).

Proposition A.1 (Construction of $\mathbf{V}_x$ and $\mathbf{Z}$ in Eq. (3))**.** *Let $\mathbf{V}_+ \in \mathbb{R}^{d_u \times (d_x - d_y)}$ be a matrix that has orthonormal columns orthogonal to $\mathbf{V}_y$, i.e. $\mathbf{V}_+^\top \mathbf{V}_y = \mathbf{0}$ and $\mathbf{V}_+^\top \mathbf{V}_+ = \mathbf{I}$. If we construct $\mathbf{V}_x$ and $\mathbf{Z}$ as described below, then the relationship between x and y in* (1) *is equivalent to that in* (3)*:*

$$\mathbf{V}_x = [\mathbf{V}_y, \mathbf{V}_+], \qquad \mathbf{Z} = \mathbf{O}\mathbf{V}_x. \tag{18}$$

Proof. It suffices to show that $\mathbf{Z}\mathbf{V}_x^\top = \mathbf{O}$. By the definitions of $\mathbf{V}_x$ and $\mathbf{Z}$ in (18), we have

$$\begin{aligned}\mathbf{Z}\mathbf{V}_x^\top &= \mathbf{O}\mathbf{V}_x\mathbf{V}_x^\top \\ &= \mathbf{U}\mathbf{S}\mathbf{V}_y^\top [\mathbf{V}_y, \mathbf{V}_1] \begin{bmatrix} \mathbf{V}_y^\top \\ \mathbf{V}_1^\top \end{bmatrix} = \mathbf{U}\mathbf{S}[\mathbf{I}, \mathbf{0}] \begin{bmatrix} \mathbf{V}_y^\top \\ \mathbf{V}_1^\top \end{bmatrix} = \mathbf{U}\mathbf{S}\mathbf{V}_y^\top = \mathbf{O}.\end{aligned}$$

□

To improve the conditioning of $\mathbf{V}_x^\top \mathbf{L}^{-1}\mathbf{B}$ and $\mathbf{V}_x^\top \widetilde{\mathbf{L}}^{-1}\mathbf{B}$, we construct $\mathbf{V}_+$ from the first $d_x - d_y$ left singular vectors of $(\mathbf{I} - \mathbf{V}_y\mathbf{V}_y^\top)\widetilde{\mathbf{L}}^{-1}\mathbf{B}$, corresponding to its $d_x - d_y$ largest singular values. This ensures that $\mathbf{V}_+$ is orthogonal to $\mathbf{V}_y$ while capturing the dominant energy of $\widetilde{\mathbf{L}}^{-1}\mathbf{B}$.

B PROOFS OF SECTION 3

Recall that we denote $\mathbb{D}_a$ and $\mathbb{D}_l$ as expected KL-divergence between `Approx-IMH posterior` and `Latent-IMH posterior` against `Exact posterior`:

$$\mathbb{D}_a := 2\,\mathbb{E}_y[\mathrm{KL}(\pi_a(x|y)||\pi(x|y))], \quad \mathbb{D}_l := 2\,\mathbb{E}_y[\mathrm{KL}(\pi_l(x|y)||\pi(x|y))]. \tag{19}$$

In this section, we first present the general results for $\mathbb{D}_a$ and $\mathbb{D}_l$ in Theorem B.2, located in Section B.1. We then provide the proofs for Proposition 3.2 in Section B.2, Proposition 3.3 in Section B.3, and Theorem B.2 in Section B.4.

B.1 EXPECTED KL-DIVERGENCE BOUNDS FOR THE GENERAL SETTING

We define matrix $\mathbf{K} \in \mathbb{R}^{d\times d}$, reduced and full SVD of $\mathbf{A}$ and $\widetilde{\mathbf{A}}$ as follows:

$$\mathbf{K} := \mathbf{F}^{-1}\widetilde{\mathbf{F}}, \tag{20}$$

$$\mathbf{A} = \mathbf{U}\mathbf{S}\mathbf{V}_{\|}^{\top} = \mathbf{U}[\mathbf{S}\dot{:}\mathbf{0}]\begin{bmatrix}\mathbf{V}_{\|}^{\top}\\ \mathbf{V}_{\perp}^{\top}\end{bmatrix}, \quad \widetilde{\mathbf{A}} = \widetilde{\mathbf{U}}\widetilde{\mathbf{S}}\widetilde{\mathbf{V}}_{\|}^{\top} = \widetilde{\mathbf{U}}[\widetilde{\mathbf{S}}\dot{:}\mathbf{0}]\begin{bmatrix}\widetilde{\mathbf{V}}_{\|}^{\top}\\ \widetilde{\mathbf{V}}_{\perp}^{\top}\end{bmatrix}, \tag{21}$$

where $\mathbf{S}$ and $\widetilde{\mathbf{S}}$ are diagonal matrices with $(\mathbf{S})_{ii} = s_i$ and $\widetilde{\mathbf{S}}_{ii} = \widetilde{s}_i$, for all $i \in [d_y]$.

Next, we introduce the constants κ_+, κ_-, ϵ and τ in Assumption B.1 to quantify the proximity between $\mathbf{F}$ and $\widetilde{\mathbf{F}}$, as well as $\mathbf{A}$ and $\widetilde{\mathbf{A}}$.

Assumption B.1 ($(\kappa_+, \kappa_-, \epsilon, \tau)$-condition)**.** *We assume there exit constants $\kappa_+, \kappa_-, \tau > 0$ and $\epsilon \in (0,1)$ satisfying the following conditions:*

- *κ_+ and κ_- are the largest and smallest singular values of $\mathbf{K}$ respectively.*
- *$\forall i \in [d]$, $\mathbf{v}_i^{\top}\widetilde{\mathbf{v}}_i \in [1-\epsilon, 1+\epsilon]$ and $\mathbf{u}_i^{\top}\widetilde{\mathbf{u}}_i \in [1-\epsilon, 1+\epsilon]$.*
- *$\|\mathbf{V}_{\|}\mathbf{S}_\sigma\mathbf{V}_{\|}^{\top} - \widetilde{\mathbf{V}}_{\|}\widetilde{\mathbf{S}}_\sigma\widetilde{\mathbf{V}}_{\|}^{\top}\|_F^2 \le \tau$ and $\|\mathbf{U}\mathbf{S}_\sigma\mathbf{U}^{\top} - \widetilde{\mathbf{U}}\widetilde{\mathbf{S}}_\sigma\widetilde{\mathbf{U}}^{\top}\|_F^2 \le \tau$, where*

$$\mathbf{S}_\sigma := \mathrm{Diag}\left[\frac{1}{1+\sigma^2/s_i^2}\right]_{i\in[d_y]}, \qquad \widetilde{\mathbf{S}}_\sigma := \mathrm{Diag}\left[\frac{1}{1+\sigma^2/\widetilde{s}_i^2}\right]_{i\in[d_y]}, \tag{22}$$

where $\{s_i\}_{i\in[d_y]}$ and $\{\widetilde{s}_i\}_{i\in[d_y]}$ are singular values of $\mathbf{A}$ and $\widetilde{\mathbf{A}}$, respectively.

Specifically, κ_+ and κ_- capture the similarity between $\mathbf{F}$ and $\widetilde{\mathbf{F}}$, while ϵ and τ measure the closeness between $\mathbf{A}$ and $\widetilde{\mathbf{A}}$. Note that when $s_i \gg \sigma$, $(\mathbf{S}_\sigma)_{ii}$ and $(\widetilde{\mathbf{S}}_\sigma)_{ii}$ are close to 1, whereas when $\sigma \ll s_i$, they are close to 0. Therefore, τ serves as a measure of how close the singular vectors of $\mathbf{A}$ and $\widetilde{\mathbf{A}}$ that correspond to relatively large singular values are. Clearly, when $\mathbf{F} = \widetilde{\mathbf{F}}$, we have $\kappa_+ = \kappa_- = 1$ and $\epsilon = \tau = 0$.

Using Assumption B.1 along with the Gaussian assumptions for the prior and noise from Assumption 3.1, we can establish upper bounds for $\mathbb{D}_a$ and $\mathbb{D}_l$ as stated in Theorem B.2.

Theorem B.2. *Given Assumption 3.1 and Assumption B.1 we obtain the following upper bounds for $\mathbb{D}_a$ and $\mathbb{D}_l$ respectively:*

$$\begin{aligned}\mathbb{D}_a \le\ & 2\epsilon d + \sum_{i\in[d_y]}\left((1+2\epsilon)\rho_i + \log\gamma_i\right)\\ & + \left(1 + \frac{\|\mathbf{A}\|^2}{\sigma^2}\right)\left\{\tau + \kappa_a d_y + \sigma^2\sum_{i\in[d_y]}\left(\left(\frac{1}{\zeta_i} - \frac{1}{\widetilde{\zeta}_i}\right)^2 + \frac{4\epsilon}{\zeta_i\widetilde{\zeta}_i}\right)\right\},\end{aligned} \tag{23}$$

$$\begin{aligned}\mathbb{D}_l \le\ & 2\epsilon_l d + \sum_{i\in[d_y]}\left((1+2\epsilon_l)\rho_i + \log\frac{\gamma_i}{\kappa_-^2}\right)\\ & + \left(2 + \frac{\|\mathbf{A}\|^2}{\sigma^2}\right)\tau + \kappa_l d_y + \sigma^2\sum_{i\in[d_y]}\left(\frac{\kappa_+^2}{\widetilde{\zeta}_i^2} + \frac{1}{\zeta_i^2} - \frac{2(\kappa_- + \epsilon(\kappa_- - 1))}{\zeta_i\widetilde{\zeta}_i}\right),\end{aligned} \tag{24}$$

where ϵ_l, κ_a and κ_l are constants defined by

$$\epsilon_l = \kappa_+(1+\epsilon) - 1, \quad \kappa_a = \max\left\{\left|\frac{1}{\kappa_-} - 1\right|, \left|\frac{1}{\kappa_+} - 1\right|\right\}^2, \quad \kappa_l = \frac{\kappa_+^2}{\kappa_-^2} - \frac{2\kappa_-}{\kappa_+} + 1, \tag{25}$$

and for all $i \in [d_y]$, γ_i, ρ_i, ζ_i and $\widetilde{\zeta}_i$ are constants defined by

$$\gamma_i = (\widetilde{s}_i^2 + \sigma^2)/(s_i^2 + \sigma^2), \qquad \rho_i = (s_i^2 - \widetilde{s}_i^2)/(\widetilde{s}_i^2 + \sigma^2), \tag{26}$$

$$\zeta_i = s_i + \sigma^2/s_i, \qquad \widetilde{\zeta}_i = \widetilde{s}_i + \sigma^2/\widetilde{s}_i. \tag{27}$$

Remark. We highlight several observations regarding the bounds established in Theorem B.2. First, note that both bounds satisfy $\mathbb{D}_a = 0$ and $\mathbb{D}_l = 0$ when $\widetilde{\mathbf{F}} = \mathbf{F}$. In many practical applications, the signal-to-noise ratio satisfies $\frac{\|\mathbf{A}\|}{\sigma} \gg 1$. Assuming $\tau = \mathcal{O}(1)$ and both d and d_y are large, the bounds can be approximated as

$$\mathbb{D}_a \sim \epsilon d + \frac{\|\mathbf{A}\|^2}{\sigma^2} d_y, \qquad \mathbb{D}_l \sim \epsilon d + d_y + \frac{\|\mathbf{A}\|^2}{\sigma^2}. \tag{28}$$

This shows that $\mathbb{D}_a$ is highly sensitive to noise. Furthermore, both $\mathbb{D}_a$ and $\mathbb{D}_l$ grow with increasing spectral approximation error and problem dimensions. This is consistent to our analysis in the simpler diagonal case (see Proposition 3.3).

B.2 Proof of Proposition 3.2

Lemma B.3 (KL-divergence between two multivariate Gaussians)**.**

$$2\mathrm{KL}(\mathcal{N}(\mu_1, \mathbf{\Sigma}_1) || \mathcal{N}(\mu_2, \mathbf{\Sigma}_2)) = \log \frac{|\mathbf{\Sigma}_2|}{|\mathbf{\Sigma}_1|} + \mathrm{Tr}(\mathbf{\Sigma}_2^{-1}\mathbf{\Sigma}_1) - d + (\mu_2 - \mu_1)\mathbf{\Sigma}_2^{-1}(\mu_2 - \mu_1). \tag{29}$$

With the above lemma, we can prove the closed-form expression for $\mathbb{D}_a$ and $\mathbb{D}_l$ in Proposition 3.2:

Proof of Proposition 3.2. When $x \sim \mathcal{N}(0, \mathbf{I})$, we have $y \sim \mathcal{N}(\mathbf{0}, \mathbf{A}\mathbf{A}^\top + \sigma^2\mathbf{I})$, which is equivalent to the following linear transformation expression for y:

$$y = \mathbf{M}z, \qquad \text{where} \quad \mathbf{M} = (\mathbf{A}\mathbf{A}^\top + \sigma^2\mathbf{I})^{1/2}, \qquad z \sim \mathcal{N}(\mathbf{0}, \mathbf{I}).$$

By Lemma B.3, we have

$$\mathbb{D}_a = \left(\log \frac{|\mathbf{\Sigma}|}{|\mathbf{\Sigma}_a|} + \mathrm{Tr}(\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_a) - d \right) + \mathbb{E}_y[(\mu_a - \mu)^\top \mathbf{\Sigma}^{-1}(\mu_a - \mu)]. \tag{30}$$

By Section 3, we can express the quadratic formula in Eq. (30) as follows:

$$\begin{aligned}
\mathbb{E}_y[(\mu_a - \mu)^\top \mathbf{\Sigma}^{-1}(\mu_a - \mu)] &= \mathbb{E}_z[z^\top \mathbf{M}^\top (\widetilde{\mathbf{A}}^\dagger - \mathbf{A}^\dagger)^\top \mathbf{\Sigma}^{-1} (\widetilde{\mathbf{A}}^\dagger - \mathbf{A}^\dagger)\mathbf{M}z] && (31)\\
&= \mathbb{E}_z[\langle \mathbf{M}^\top (\widetilde{\mathbf{A}}^\dagger - \mathbf{A}^\dagger)^\top \mathbf{\Sigma}^{-1} (\widetilde{\mathbf{A}}^\dagger - \mathbf{A}^\dagger)\mathbf{M}, zz^\top \rangle] && (32)\\
&= \mathrm{Tr}\left(\mathbf{M}^\top (\widetilde{\mathbf{A}}^\dagger - \mathbf{A})^\top \mathbf{\Sigma}^{-1} (\widetilde{\mathbf{A}}^\dagger - \mathbf{A}^\dagger)\mathbf{M} \right) && (33)\\
&= \frac{1}{\sigma^2} \|\mathbf{A}(\widetilde{\mathbf{A}}^\dagger - \mathbf{A}^\dagger)\mathbf{A}\|_F^2 + \sigma^2 \|\widetilde{\mathbf{A}}^\dagger - \mathbf{A}^\dagger\|_F^2 && (34)\\
&\quad + \|(\widetilde{\mathbf{A}}^\dagger - \mathbf{A}^\dagger)\mathbf{A}\|_F^2 + \|\mathbf{A}(\widetilde{\mathbf{A}}^\dagger - \mathbf{A}^\dagger)\|_F^2. && (35)
\end{aligned}$$

□

B.3 Proof of Proposition 3.3

Proof. By the diagonal assumptions of $\mathbf{F}$ and $\widetilde{\mathbf{F}}$, $\mathbf{K} = \mathbf{F}^{-1}\widetilde{\mathbf{F}}$ is a diagonal matrix with $\mathbf{K}_{ii} = \alpha_i$ for $\forall i \in [d]$. $\mathbf{A}, \widetilde{\mathbf{A}} \in \mathbb{R}^{d_y \times d}$ are matrices with zero elements except for:

$$\mathbf{A}_{ii} = s_i, \qquad \widetilde{\mathbf{A}}_{ii} = \alpha_i s_i, \quad \forall i \in [d_y]. \tag{36}$$

By definitions in Section 3, $\mathbf{A}^\dagger, \widetilde{\mathbf{A}}^\dagger, \mathbf{A}_l^\dagger \in \mathbb{R}^{d \times d_y}$ with zero elements except for:

$$\mathbf{A}^\dagger_{ii} = \frac{s_i}{s_i^2 + \sigma^2}, \qquad \widetilde{\mathbf{A}}^\dagger_{ii} = \frac{\alpha_i s_i}{\alpha_i^2 s_i^2 + \sigma^2}, \qquad (\mathbf{A}_l^\dagger)_{ii} = \frac{\alpha^2 s_i}{\alpha_i^2 s_i^2 + \sigma^2}, \quad \forall i \in [d_y]. \tag{37}$$

$\mathbf{\Sigma}, \mathbf{\Sigma}_a, \mathbf{\Sigma}_l \in \mathbb{R}^{d \times d}$ are diagonal matrices with diagonals as

$$\begin{cases} \mathbf{\Sigma}_{ii} = \frac{\sigma^2}{s_i^2 + \sigma^2}, \quad (\mathbf{\Sigma}_a)_{ii} = \frac{\sigma^2}{\alpha_i^2 s_i^2 + \sigma^2}, \quad (\mathbf{\Sigma}_l)_{ii} = \frac{\alpha^2 \sigma^2}{\alpha_i^2 s_i^2 + \sigma^2}, & i \le d_y, \\ \mathbf{\Sigma}_{ii} = 1, \quad (\mathbf{\Sigma}_a)_{ii} = 1, \quad (\mathbf{\Sigma}_l)_{ii} = \alpha_i^2, & i > d_y. \end{cases} \tag{38}$$

Thus $\boldsymbol{\Delta}_a, \boldsymbol{\Delta}_l \in \mathbb{R}^{d\times d_y}$ are matrices with zero elements except for:

$$(\boldsymbol{\Delta}_a)_{ii} = \widetilde{\mathbf{A}}^{\dagger}_{ii} - \mathbf{A}^{\dagger}_{ii} = \frac{s_i(\alpha_i s_i^2 - \sigma^2)(1-\alpha_i)}{(\alpha_i^2 s_i^2 + \sigma^2)(s_i^2+\sigma^2)}, \forall i \in [d_y] \tag{39}$$

$$(\boldsymbol{\Delta}_l)_{ii} = (\mathbf{A}_l^{\dagger})_{ii} - \mathbf{A}^{\dagger}_{ii} = \frac{\sigma^2 s_i(\alpha_i^2 - 1)}{(\alpha_i^2 s_i^2 + \sigma^2)(s_i^2+\sigma^2)}, \forall i \in [d_y]. \tag{40}$$

Next we derive the expressions of $\mathbb{D}_a$ and $\mathbb{D}_l$ derived in Proposition 3.2, respectively.

1. `Approx-IMH posterior`: by Eq. (38) we have

$$\log\frac{|\boldsymbol{\Sigma}|}{|\boldsymbol{\Sigma}_a|} = \sum_{i\in[d_y]} \frac{\alpha_i^2 s_i^2 + \alpha^2}{s_i^2+\sigma^2} = \sum_{i\in[d_y]} \rho_i, \quad \mathrm{Tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_a) = \sum_{i\in[d_y]} \frac{1}{\rho_i} + d - d_y. \tag{41}$$

   By Eq. (39), we have $\mathbf{A}\boldsymbol{\Delta}_a \in \mathbb{R}^{d_y\times d_y}$ is a diagonal matrix with diagonals:

$$(\mathbf{A}\boldsymbol{\Delta}_a)_{ii} = \frac{s_i^2(\alpha_i s_i^2 - \sigma^2)(1-\alpha_i)}{(\alpha_i^2 s_i^2 + \sigma^2)(s_i^2+\sigma^2)}. \tag{42}$$

   Denote $f_i = (\alpha_i^2 s_i^2 + \sigma^2)^2(s_i^2+\sigma^2)^2$, then

$$\frac{1}{\sigma^2}\|\mathbf{A}\boldsymbol{\Delta}_a\|_F^2 = \sum_{i\in[d_y]} \frac{s_i^6(\alpha_i s_i^2 - \sigma^2)^2(1-\alpha_i)^2/\sigma^2}{f_i}, \tag{43}$$

$$\sigma^2\|\boldsymbol{\Delta}_a\|_F^2 = \sum_{i\in[d_y]} \frac{\sigma^2 s_i^2(\alpha_i s_i^2 - \sigma^2)^2(1-\alpha_i)^2}{f_i}, \tag{44}$$

$$\|\boldsymbol{\Delta}_a\mathbf{A}\|_F^2 = \|\mathbf{A}\boldsymbol{\Delta}_a\|_F^2 = \sum_{i\in[d_y]} \frac{s_i^4(\alpha_i s_i^2 - \sigma^2)^2(1-\alpha_i)^2}{f_i}. \tag{45}$$

   Thus

$$\frac{1}{\sigma^2}\|\mathbf{A}\boldsymbol{\Delta}_a\|_F^2 + \sigma^2\|\boldsymbol{\Delta}_a\|_F^2 + \|\boldsymbol{\Delta}_a\mathbf{A}\|_F^2 + \|\mathbf{A}\boldsymbol{\Delta}_a\|_F^2 = \sum_{i\in[d_y]} \zeta_i^2(\alpha_i s_i^2 - \sigma^2)^2(1-\alpha_i)^2\frac{s_i^2}{\sigma^2}. \tag{46}$$

   Substitute Eqs. (41) and (46) into Eq. (10), we have the expressions for $\mathbb{D}_a$ in Eq. (11).

2. `Latent-IMH posterior`: by Eq. (38), we have

$$\log\frac{|\boldsymbol{\Sigma}|}{|\boldsymbol{\Sigma}_l|} = \sum_{i\in[d_y]} \log\frac{\rho_i}{\alpha_i^2} + \sum_{i=d_y+1}^{d} \log\frac{1}{\alpha_i^2}, \quad \mathrm{Tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_a) = \sum_{i\in[d_y]} \frac{\alpha_i^2}{\rho_i} + \sum_{i=d_y+1}^{d} \alpha_i^2. \tag{47}$$

   By Eq. (40), we have $\mathbf{A}\boldsymbol{\Delta}_l \in \mathbb{R}^{d_y\times d_y}$ is a diagonal matrix with diagonals:

$$(\mathbf{A}\boldsymbol{\Delta}_l)_{ii} = \frac{\alpha_i^2 s_i^2(\alpha_i^2 - 1)}{(\alpha_i^2 s_i^2 + \sigma^2)(s_i^2+\sigma^2)}. \tag{48}$$

   Denote $f_i = (\alpha_i^2 s_i^2 + \sigma^2)^2(s_i^2+\sigma^2)^2$, then

$$\frac{1}{\sigma^2}\|\mathbf{A}\boldsymbol{\Delta}_l\|_F^2 = \sum_{i\in[d_y]} \frac{\sigma^2 s_i^6(\alpha_i^2-1)^2)}{f_i}, \tag{49}$$

$$\sigma^2\|\boldsymbol{\Delta}_l\|_F^2 = \sum_{i\in[d_y]} \frac{\sigma^6 s_i^2(\alpha_i^2-1)^2}{f_i}, \tag{50}$$

$$\|\boldsymbol{\Delta}_l\mathbf{A}\|_F^2 = \|\mathbf{A}\boldsymbol{\Delta}_l\|_F^2 = \sum_{i\in[d_y]} \frac{\sigma^4 s_i^4(\alpha_i^2-1)^2}{f_i}. \tag{51}$$

Thus

$$\frac{1}{\sigma^2}\|\mathbf{A}\boldsymbol{\Delta}_l\|_F^2 + \sigma^2\|\boldsymbol{\Delta}_l\|_F^2 + \|\boldsymbol{\Delta}_l\mathbf{A}\|_F^2 + \|\mathbf{A}\boldsymbol{\Delta}_l\|_F^2 = \sum_{i\in[d_y]} \zeta_i^2(\alpha_i^2-1)^2 s_i^2\sigma^2. \tag{52}$$

Combining Eqs. (47) and (52) with Proposition 3.2, we have the expressions for $\mathbb{D}_a$ in Eq. (12).

□

B.4 Proof of Theorem B.2

Lemma B.4. *We have the following properties for* $\mathbf{A}^\dagger$, $\widetilde{\mathbf{A}}^\dagger$ *and* $\mathbf{A}_l^\dagger$

- $(\widetilde{\mathbf{A}}^\dagger - \mathbf{A}^\dagger)\mathbf{A} = \widetilde{\mathbf{A}}^\dagger\widetilde{\mathbf{A}}\mathbf{K}^{-1} - \mathbf{A}^\dagger\mathbf{A}$.
- $(\mathbf{A}_l^\dagger - \mathbf{A}^\dagger)\mathbf{A} = \mathbf{K}\widetilde{\mathbf{A}}^\dagger\widetilde{\mathbf{A}}\mathbf{K}^{-1} - \mathbf{A}^\dagger\mathbf{A}$.
- $\mathbf{A}(\mathbf{A}_l^\dagger - \mathbf{A}^\dagger) = \widetilde{\mathbf{A}}\widetilde{\mathbf{A}}^\dagger - \mathbf{A}\mathbf{A}^\dagger$.

Proof.

- Note that

$$\mathbf{A} = \mathbf{O}\mathbf{F} = \mathbf{O}\widetilde{\mathbf{F}}\widetilde{\mathbf{F}}^{-1}\mathbf{F} = \mathbf{O}\widetilde{\mathbf{F}}(\mathbf{F}^{-1}\widetilde{\mathbf{F}})^{-1} = \widetilde{\mathbf{A}}\mathbf{K}^{-1}. \tag{53}$$

- By Section 3, $\mathbf{A}_l^\dagger = \mathbf{K}\widetilde{\mathbf{A}}^\dagger$. The property can be obtained by combing this relation with the previous property.
- Using Eq. (53), we have

$$\mathbf{A}\mathbf{A}_l^\dagger = \mathbf{A}\mathbf{K}\widetilde{\mathbf{A}}^\dagger = \widetilde{\mathbf{A}}\mathbf{K}^{-1}\mathbf{K}\widetilde{\mathbf{A}}^\dagger = \widetilde{\mathbf{A}}\widetilde{\mathbf{A}}^\dagger. \tag{54}$$

□

Lemma B.5. *If Assumption B.1 holds, we have*

- $\|\widetilde{\mathbf{A}}^\dagger\widetilde{\mathbf{A}} - \mathbf{A}^\dagger\mathbf{A}\| \le \tau$.
- $\|\widetilde{\mathbf{A}}\widetilde{\mathbf{A}}^\dagger - \mathbf{A}\mathbf{A}^\dagger\| \le \tau$.

Proof. By the SVD notation of $\mathbf{A}$, we have

$$\mathbf{A}^\dagger = \mathbf{A}^\top(\mathbf{A}\mathbf{A}^\top + \sigma^2\mathbf{I})^{-1} = \mathbf{V}_\parallel \mathrm{Diag}\left[\frac{s_i}{s_i^2+\sigma^2}\right]_{i\in[d_y]}\mathbf{U}^\top.$$

and then by the definition of $\mathbf{S}_\sigma$ and $\widetilde{\mathbf{S}}_\sigma$, we have

$$\mathbf{A}^\dagger\mathbf{A} = \mathbf{V}_\parallel\mathbf{S}_\sigma\mathbf{V}_\parallel^\top, \qquad \mathbf{A}\mathbf{A}^\dagger = \mathbf{U}\mathbf{S}_\sigma\mathbf{U}.$$

Similar argument holds for the expressions of $\widetilde{\mathbf{A}}^\dagger\widetilde{\mathbf{A}}$ and $\widetilde{\mathbf{A}}\widetilde{\mathbf{A}}^\dagger$. By Assumption B.1, we have $\|\widetilde{\mathbf{A}}^\dagger\widetilde{\mathbf{A}} - \mathbf{A}^\dagger\mathbf{A}\| \le \tau$ and $\|\widetilde{\mathbf{A}}\widetilde{\mathbf{A}}^\dagger - \mathbf{A}\mathbf{A}^\dagger\| \le \tau$. □

Lemma B.6. *We have the following bounds for the covariance related terms in the KL-divergence for* `Approx-IMH posterior` *and* `Latent-IMH posterior`*:*

$$\log\frac{|\boldsymbol{\Sigma}|}{|\boldsymbol{\Sigma}_a|} + \mathrm{Tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_a) - d \le 2\epsilon d + \sum_{i\in[d_y]}(1+2\epsilon)\rho_i + \log\gamma_i, \tag{55}$$

$$\log\frac{|\boldsymbol{\Sigma}|}{|\boldsymbol{\Sigma}_l|} + \mathrm{Tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_l) - d \le 2\epsilon_l d + \sum_{i\in[d_y]}(1+2\epsilon_l)\rho_i + \log\frac{\gamma_i}{\kappa_-^2}, \tag{56}$$

where

$$\epsilon_l = \kappa_+(1+\epsilon) - 1, \qquad \gamma_i = (\widetilde{s}_i^2+\sigma^2)/(s_i^2+\sigma^2), \qquad \rho_i = (s_i^2 - \widetilde{s}_i^2)/(\widetilde{s}_i^2+\sigma^2), \quad \forall i\in[d_y]. \tag{57}$$

Proof. For `Approx-IMH posterior`, by the SVD notation of $\mathbf{A}$ and $\widetilde{\mathbf{A}}$, we have

$$\boldsymbol{\Sigma}^{-1} = (\mathbf{I} - \mathbf{A}^\dagger\mathbf{A})^{-1} = \mathbf{V}\mathrm{Diag}\Big[\underbrace{\cdots, \frac{s_i^2}{\sigma^2} + 1, \cdots}_{i \in [d_y]}, \underbrace{1, 1, \cdots}_{d - d_y}\Big]\mathbf{V}^\top =: \mathbf{V}\mathbf{S}_1\mathbf{V}^\top,$$

$$\boldsymbol{\Sigma}_a = \widetilde{\mathbf{V}}\mathrm{Diag}\Big[\underbrace{\cdots, \frac{\sigma^2}{\widetilde{s}_i^{\,2} + \sigma^2}, \cdots}_{i \in [d_y]}, \underbrace{1, 1, \cdots}_{d - d_y}\Big]\widetilde{\mathbf{V}}^\top =: \widetilde{\mathbf{V}}\mathbf{S}_2\widetilde{\mathbf{V}}^\top.$$

Denote $\mathbf{V}^\top\widetilde{\mathbf{V}} = \mathbf{I} + \boldsymbol{\Delta}$, by the Assumption B.1, we have diagonals $\boldsymbol{\Delta}_{ii} \le \epsilon$, for all $i \in [d]$. Thus

$$\begin{aligned}
\mathrm{Tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_a) - d &= \mathrm{Tr}(\mathbf{V}\mathbf{S}_1\mathbf{V}^\top\widetilde{\mathbf{V}}\mathbf{S}_2\widetilde{\mathbf{V}}^\top) - d \\
&= \mathrm{Tr}(\mathbf{S}_1(\mathbf{I} + \boldsymbol{\Delta})\mathbf{S}_2(\mathbf{I} + \boldsymbol{\Delta}^\top)) - d \\
&\overset{(a)}{=} \mathrm{Tr}(\mathbf{S}_1\mathbf{S}_2) + 2\,\mathrm{Tr}(\mathbf{S}_1\mathbf{S}_2\boldsymbol{\Delta}) - d \\
&\le (1 + 2\epsilon)\,\mathrm{Tr}(\mathbf{S}_1\mathbf{S}_2) - d \\
&= (1 + 2\epsilon)\sum_{i \in [d_y]} \frac{s_i^2 + \sigma^2}{\widetilde{s}_i^{\,2} + \sigma^2} + (1 + 2\epsilon)(d - d_y) - d \\
&= (1 + 2\epsilon)\sum_{i \in [d_y]} \frac{s_i^2 - \widetilde{s}_i^{\,2}}{\widetilde{s}_i^{\,2} + \sigma^2} + 2\epsilon d,
\end{aligned} \tag{58}$$

where step (a) neglected second order term on $\boldsymbol{\Delta}$.

For `Latent-IMH posterior`, we can use a similar argument. Since $\boldsymbol{\Sigma}_l = \mathbf{K}\boldsymbol{\Sigma}_a\mathbf{K}^\top$, we have

$$\mathrm{Tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_l) = \mathrm{Tr}(\mathbf{V}\mathbf{S}_1\mathbf{V}^\top\mathbf{K}\widetilde{\mathbf{V}}\mathbf{S}_2\widetilde{\mathbf{V}}^\top\mathbf{K}^\top) = \mathrm{Tr}(\mathbf{S}_1(\mathbf{V}^\top\mathbf{K}\widetilde{\mathbf{V}})\mathbf{S}_2(\mathbf{V}\mathbf{K}\widetilde{\mathbf{V}})^\top). \tag{59}$$

Denote $\mathbf{V}^\top\mathbf{K}\widetilde{\mathbf{V}} = \mathbf{I} + \boldsymbol{\Delta}'$, then the i-th diagonal entry of $\boldsymbol{\Delta}'$ has the following bound:

$$\begin{aligned}
\boldsymbol{\Delta}'_{ii} = \mathbf{v}_i^\top\mathbf{K}\widetilde{\mathbf{v}}_i - 1 &= \frac{1}{2}\,\mathrm{Tr}(\mathbf{K}(\mathbf{v}_i\widetilde{\mathbf{v}}_i^\top + \widetilde{\mathbf{v}}_i\mathbf{v}_i^\top)) - 1 \\
&\le \kappa_+(1 + \epsilon) - 1,
\end{aligned} \tag{60}$$

where the last inequality follows by Assumption B.1. Now by Eq. (59) we have

$$\begin{aligned}
\mathrm{Tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_l) - d &= \mathrm{Tr}(\mathbf{S}_1(\mathbf{I} + \boldsymbol{\Delta}')\mathbf{S}_2(\mathbf{I} + \boldsymbol{\Delta}'^\top)) - d \\
&= \mathrm{Tr}(\mathbf{S}_1\mathbf{S}_2) + 2\,\mathrm{Tr}(\mathbf{S}_1\mathbf{S}_2\boldsymbol{\Delta}') - d \\
&\le [2\kappa_+(1 + \epsilon) - 1]\,\mathrm{Tr}(\mathbf{S}_1\mathbf{S}_2) - d \\
&= [2\kappa_+(1 + \epsilon) - 1]\Big[\sum_{i \in [d_y]} \frac{s_i^2 + \sigma^2}{\widetilde{s}_i^{\,2} + \sigma^2} + d - d_y\Big] - d \\
&= [2\kappa_+(1 + \epsilon) - 1]\sum_{i \in [d_y]} \frac{s_i^2 - \widetilde{s}_i^{\,2}}{\widetilde{s}_i^{\,2} + \sigma^2} + 2[\kappa_+(1 + \epsilon) - 1]d.
\end{aligned} \tag{61}$$

□

Proof of Theorem B.2. For the ease of notation, we denote the quadratic terms in the KL-divergence expressions for `Approx-IMH posterior` and `Latent-IMH posterior` in Proposition 3.2 by

$$f = \underbrace{\frac{1}{\sigma^2}\|\mathbf{A}(\widetilde{\mathbf{A}}^\dagger - \mathbf{A}^\dagger)\mathbf{A}\|_F^2}_{f_1} + \underbrace{\sigma^2\|(\widetilde{\mathbf{A}}^\dagger - \mathbf{A}^\dagger)\|_F^2}_{f_2} + \underbrace{\|(\widetilde{\mathbf{A}}^\dagger - \mathbf{A}^\dagger)\mathbf{A}\|_F^2}_{f_3} + \underbrace{\|\mathbf{A}(\widetilde{\mathbf{A}}^\dagger - \mathbf{A}^\dagger)\|_F^2}_{f_4}. \tag{62}$$

$$g = \underbrace{\frac{1}{\sigma^2}\|\mathbf{A}(\mathbf{A}_l^\dagger - \mathbf{A}^\dagger)\mathbf{A}\|_F^2}_{g_1} + \underbrace{\sigma^2\|(\mathbf{A}_l^\dagger - \mathbf{A}^\dagger)\|_F^2}_{g_2} + \underbrace{\|(\mathbf{A}_l^\dagger - \mathbf{A}^\dagger)\mathbf{A}\|_F^2}_{g_3} + \underbrace{\|\mathbf{A}(\mathbf{A}_l^\dagger - \mathbf{A}^\dagger)\|_F^2}_{g_4} \tag{63}$$

1. `Approx-IMH posterior`: By Lemma B.4, we have

$$f_1 = \frac{1}{\sigma^2}\|\mathbf{A}(\widetilde{\mathbf{A}}^\dagger - \mathbf{A}^\dagger)\mathbf{A}\|_F^2 = \frac{1}{\sigma^2}\|\mathbf{A}(\widetilde{\mathbf{A}}^\dagger\widetilde{\mathbf{A}}\mathbf{K}^{-1} - \mathbf{A}^\dagger\mathbf{A})\|_F^2 \le \frac{\|\mathbf{A}\|^2}{\sigma^2}\|\widetilde{\mathbf{A}}^\dagger\widetilde{\mathbf{A}}\mathbf{K}^{-1} - \mathbf{A}^\dagger\mathbf{A}\|_F^2, \tag{64}$$

$$f_3 = \|(\widetilde{\mathbf{A}}^\dagger - \mathbf{A})\mathbf{A}\|_F^2 = \|\widetilde{\mathbf{A}}^\dagger\widetilde{\mathbf{A}}\mathbf{K}^{-1} - \mathbf{A}^\dagger\mathbf{A}\|_F^2. \tag{65}$$

By Assumption B.1,

$$\begin{aligned}
\|\widetilde{\mathbf{A}}^\dagger\widetilde{\mathbf{A}}\mathbf{K}^{-1} - \mathbf{A}^\dagger\mathbf{A}\|_F^2 &= \|\widetilde{\mathbf{V}}_\parallel\widetilde{\mathbf{S}}_\sigma\widetilde{\mathbf{V}}_\parallel^\top\mathbf{K}^{-1} - \mathbf{V}_\parallel\mathbf{S}_\sigma\mathbf{V}_\parallel^\top\|_F^2 \\
&\le \|\widetilde{\mathbf{V}}_\parallel\widetilde{\mathbf{S}}_\sigma\widetilde{\mathbf{V}}_\parallel^\top\mathbf{K}^{-1} - \widetilde{\mathbf{V}}_\parallel\widetilde{\mathbf{S}}_\sigma\widetilde{\mathbf{V}}_\parallel^\top\|_F^2 + \|\widetilde{\mathbf{V}}_\parallel\widetilde{\mathbf{S}}_\sigma\widetilde{\mathbf{V}}_\parallel^\top - \mathbf{V}_\parallel\mathbf{S}_\sigma\mathbf{V}_\parallel^\top\|_F^2 \\
&\le \|\mathbf{I} - \mathbf{K}^{-1}\|^2\|\widetilde{\mathbf{V}}_\parallel\widetilde{\mathbf{S}}_\sigma\widetilde{\mathbf{V}}_\parallel^\top\|_F^2 + \|\widetilde{\mathbf{V}}_\parallel\widetilde{\mathbf{S}}_\sigma\widetilde{\mathbf{V}}_\parallel^\top - \mathbf{V}_\parallel\mathbf{S}_\sigma\mathbf{V}_\parallel^\top\|_F^2 \\
&\le \max\left\{\left(\frac{1}{\kappa_-} - 1\right)^2, \left(\frac{1}{\kappa_+} - 1\right)^2\right\} d_y + \tau,
\end{aligned} \tag{66}$$

where the last inequality is derived by Assumption B.1 and the fact that $\widetilde{\mathbf{S}}_\sigma$ is a diagonal matrix with elements less than 1 (Eq. (22)).

Substitute Eq. (66) into Eqs. (64) and (65), we have

$$f_1 + f_3 \le \left(1 + \frac{\|\mathbf{A}\|^2}{\sigma^2}\right)\left[\max\left\{\left(\frac{1}{\kappa_-} - 1\right)^2, \left(\frac{1}{\kappa_+} - 1\right)^2\right\} d_y + \tau\right]. \tag{67}$$

Now we consider f_2 and f_4:

$$\begin{aligned}
f_2 + f_4 &= \sigma^2\|\widetilde{\mathbf{A}}^\dagger - \mathbf{A}^\dagger\|_F^2 + \|\mathbf{A}(\widetilde{\mathbf{A}}^\dagger - \mathbf{A}^\dagger)\|_F^2 \\
&\le (\sigma^2 + \|\mathbf{A}\|^2)\|\widetilde{\mathbf{A}}^\dagger - \mathbf{A}^\dagger\|_F^2.
\end{aligned} \tag{68}$$

Define diagonal matrices $\mathbf{S}_1 = \mathrm{Diag}\left[s_i/(s_i^2 + \sigma^2)\right]_{i\in[d_y]}$ and $\mathbf{S}_2 = \mathrm{Diag}\left[\widetilde{s}_i/(\widetilde{s}_i^2 + \sigma^2)\right]_{i\in[d_y]}$. By Lemma B.5, we have

$$\begin{aligned}
\|\widetilde{\mathbf{A}}^\dagger - \mathbf{A}^\dagger\|_F^2 &\le \|\mathbf{V}_\parallel\mathbf{S}_1\mathbf{U}^\top - \widetilde{\mathbf{V}}_\parallel\mathbf{S}_2\widetilde{\mathbf{U}}^\top\|_F^2 \\
&= \mathrm{Tr}(\mathbf{V}_\parallel\mathbf{S}_1^2\mathbf{V}_\parallel^\top) + \mathrm{Tr}(\widetilde{\mathbf{V}}_\parallel\mathbf{S}_2^2\widetilde{\mathbf{V}}_\parallel^\top) - 2\,\mathrm{Tr}(\mathbf{V}_\parallel\mathbf{S}_1\mathbf{U}^\top\widetilde{\mathbf{U}}\mathbf{S}_2\widetilde{\mathbf{V}}_\parallel^\top) \\
&= \sum_{i\in[d_y]}\left(\frac{s_i^2}{(s_i^2 + \sigma^2)^2} + \frac{\widetilde{s}_i^2}{(\widetilde{s}_i^2 + \sigma^2)^2}\right) - 2\,\mathrm{Tr}(\mathbf{S}_1\mathbf{U}^\top\widetilde{\mathbf{U}}\mathbf{S}_2\widetilde{\mathbf{V}}_\parallel^\top\mathbf{V}_\parallel).
\end{aligned} \tag{69}$$

Denote $\mathbf{U}^\top\widetilde{\mathbf{U}} = \mathbf{I} + \mathbf{\Delta}_\mathbf{U}$ and $\mathbf{V}_\parallel^\top\widetilde{\mathbf{V}}_\parallel = \mathbf{I} + \mathbf{\Delta}_\mathbf{V}$, then by Assumption B.1, we have

$$\left(\mathbf{\Delta}_\mathbf{U}\right)_{ii} = \mathbf{u}_i^\top\widetilde{\mathbf{u}}_i - 1 \ge -\epsilon, \qquad \left(\mathbf{\Delta}_\mathbf{V}\right)_{ii} = \mathbf{v}_i^\top\widetilde{\mathbf{v}}_i - 1 \ge -\epsilon. \tag{70}$$

Thus

$$\begin{aligned}
\mathrm{Tr}(\mathbf{S}_1\mathbf{U}^\top\widetilde{\mathbf{U}}\mathbf{S}_2\widetilde{\mathbf{V}}_\parallel^\top\mathbf{V}_\parallel) &= \mathrm{Tr}(\mathbf{S}_1(\mathbf{I} + \mathbf{\Delta}_\mathbf{U})\mathbf{S}_2(\mathbf{I} + \mathbf{\Delta}_\mathbf{V})^\top) \\
&= \mathrm{Tr}(\mathbf{S}_1\mathbf{S}_2) + \mathrm{Tr}(\mathbf{\Delta}_\mathbf{U}\mathbf{S}_2\mathbf{S}_1) + \mathrm{Tr}(\mathbf{\Delta}_\mathbf{V}^\top\mathbf{S}_1\mathbf{S}_2) \\
&\ge (1 - 2\epsilon)\,\mathrm{Tr}(\mathbf{S}_1\mathbf{S}_2) = (1 - 2\epsilon)\sum_{i\in[d_y]}\frac{s_i\cdot\widetilde{s}_i}{(s_i^2 + \sigma^2)(\widetilde{s}_i^2 + \sigma^2)},
\end{aligned} \tag{71}$$

where we neglect the second order term in the derivation of second equality. Substitute Eq. (71) into Eq. (69), we have

$$\|\widetilde{\mathbf{A}}^\dagger - \mathbf{A}^\dagger\|_F^2 \le \sum_{i\in[d_y]}\left(\frac{s_i}{(s_i^2 + \sigma^2)} - \frac{\widetilde{s}_i}{(\widetilde{s}_i^2 + \sigma^2)}\right)^2 + \frac{4\epsilon s_i\widetilde{s}_i}{(s_i^2 + \sigma^2)(\widetilde{s}_i^2 + \sigma^2)}. \tag{72}$$

Substitute Eq. (72) into Eq. (68) and combine with Eq. (67), we have

$$f = \sum_{i\in[4]} f_i \le \left(1 + \frac{\|\mathbf{A}\|^2}{\sigma^2}\right)\left\{\tau + \kappa_a d_y + \sigma^2 \sum_{i\in[d_y]}\left(\left(\frac{1}{\zeta_i} - \frac{1}{\widetilde{\zeta}_i}\right)^2 + \frac{4\epsilon}{\zeta_i\widetilde{\zeta}_i}\right)\right\}, \tag{73}$$

where

$$\begin{aligned}\kappa_a &= \max\left\{\left(\frac{1}{\kappa_-} - 1\right)^2, \left(\frac{1}{\kappa_+} - 1\right)^2\right\},\\ \zeta_i &= s_i + \sigma^2/s_i, \qquad \widetilde{\zeta}_i = \widetilde{s}_i + \sigma^2/\widetilde{s}_i, \qquad \forall i \in [d_y].\end{aligned} \tag{74}$$

By combining Eq. (73) and Eq. (55), and substituting them into the closed-form expression for $\mathbb{D}_a$ given in Proposition 3.2, we obtain the bound for $\mathbb{D}_a$ as presented in Eq. (23).

2. `Latent-IMH posterior`: first by Lemma B.5 and Lemma B.4 we have

$$\begin{aligned}g_1 + g_4 &= \frac{1}{\sigma^2}\|\mathbf{A}(\mathbf{A}_l^\dagger - \mathbf{A}^\dagger)\mathbf{A}\|_F^2 + \|\mathbf{A}(\mathbf{A}_l^\dagger - \mathbf{A}^\dagger)\|_F^2\\ &\le \left(1 + \frac{\|\mathbf{A}\|^2}{\sigma^2}\right)\|\widetilde{\mathbf{A}}\widetilde{\mathbf{A}}^\dagger - \mathbf{A}\mathbf{A}^\dagger\|_F^2 \le \left(1 + \frac{\|\mathbf{A}\|^2}{\sigma^2}\right)\tau.\end{aligned} \tag{75}$$

Define diagonal matrices $\mathbf{S}_1 = \operatorname{Diag}\left[s_i/(s_i^2+\sigma^2)\right]_{i\in[d_y]}$ and $\mathbf{S}_2 = \operatorname{Diag}\left[\widetilde{s}_i/(\widetilde{s}_i^2+\sigma^2)\right]_{i\in[d_y]}$. By Lemma B.5, we have

$$\begin{aligned}g_2 &= \sigma^2\|\mathbf{A}_l^\dagger - \mathbf{A}^\dagger\|_F^2 = \sigma^2\|\mathbf{K}\widetilde{\mathbf{A}}^\dagger - \mathbf{A}^\dagger\|_F^2 = \sigma^2\big(\|\mathbf{K}\widetilde{\mathbf{A}}^\dagger\|_F^2 + \|\mathbf{A}^\dagger\|_F^2 - 2\operatorname{Tr}(\mathbf{K}\widetilde{\mathbf{A}}^\dagger\mathbf{A}^{+\top})\big)\\ &\le \sigma^2\big[\kappa_+^2\operatorname{Tr}(\mathbf{S}_2^2) + \operatorname{Tr}(\mathbf{S}_1^2) - 2\operatorname{Tr}(\mathbf{K}\widetilde{\mathbf{V}}_\parallel\mathbf{S}_2\widetilde{\mathbf{U}}^\top\mathbf{U}\mathbf{S}_1\mathbf{V}_\parallel^\top)\big].\end{aligned} \tag{76}$$

Let $\widetilde{\mathbf{U}}^\top\mathbf{U} = \mathbf{I} + \mathbf{\Delta}_\mathbf{U}$ and $\mathbf{V}_\parallel^\top\mathbf{K}\widetilde{\mathbf{V}}_\parallel = \mathbf{I} + \mathbf{\Delta}_\mathbf{V}$, then by Assumption B.1,

$$(\mathbf{\Delta}_\mathbf{U})_{ii} = \mathbf{u}_i^\top\widetilde{\mathbf{u}}_i - 1 \ge -\epsilon, \qquad (\mathbf{\Delta}_\mathbf{V})_{ii} = \mathbf{v}_i^\top\mathbf{K}\widetilde{\mathbf{v}}_i - 1 \ge \kappa_-(1-\epsilon) - 1. \tag{77}$$

Thus

$$\begin{aligned}\operatorname{Tr}(\mathbf{K}\widetilde{\mathbf{V}}_\parallel\mathbf{S}_2\widetilde{\mathbf{U}}^\top\mathbf{U}\mathbf{S}_1\mathbf{V}_\parallel^\top) &= \operatorname{Tr}(\mathbf{S}_2(\mathbf{I}+\mathbf{\Delta}_\mathbf{U})\mathbf{S}_1(\mathbf{I}+\mathbf{\Delta}_\mathbf{V}^\top))\\ &= \operatorname{Tr}(\mathbf{S}_2\mathbf{S}_1) + \operatorname{Tr}(\mathbf{S}_2\mathbf{S}_1\mathbf{\Delta}_\mathbf{V}^\top) + \operatorname{Tr}(\mathbf{S}_1\mathbf{S}_2\mathbf{\Delta}_\mathbf{U})\\ &\ge (\kappa_- + \epsilon(\kappa_- - 1))\operatorname{Tr}(\mathbf{S}_2\mathbf{S}_1)\\ &= (\kappa_- + \epsilon(\kappa_- - 1))\sum_{i\in[d_y]}\frac{1}{\zeta_i\widetilde{\zeta}_i}.\end{aligned} \tag{78}$$

Substitute Eq. (78) into Eq. (76), we have

$$g_2 \le \sigma^2\left(\sum_{i\in[d_y]}\frac{\kappa_+^2}{\widetilde{\zeta}_i^2} + \frac{1}{\zeta_i^2} - \frac{2(\kappa_- + \epsilon(\kappa_- - 1))}{\zeta_i\widetilde{\zeta}_i}\right). \tag{79}$$

Now we consider g_3:

$$\begin{aligned}g_3 &= \|(\mathbf{A}_l^\dagger - \mathbf{A}^\dagger)\mathbf{A}\|_F^2 = \|\mathbf{K}\widetilde{\mathbf{A}}^\dagger\widetilde{\mathbf{A}}\mathbf{K}^{-1} - \mathbf{A}^\dagger\mathbf{A}\|_F^2\\ &\le \|\mathbf{K}\widetilde{\mathbf{V}}_\parallel\widetilde{\mathbf{S}}_\sigma\widetilde{\mathbf{V}}_\parallel^\top\mathbf{K}^{-1} - \mathbf{V}_\parallel\mathbf{S}_\sigma\mathbf{V}_\parallel^\top\|_F^2\\ &\le \|\mathbf{V}_\parallel\mathbf{S}_\sigma\mathbf{V}_\parallel^\top - \widetilde{\mathbf{V}}_\parallel\widetilde{\mathbf{S}}_\sigma\widetilde{\mathbf{V}}_\parallel^\top\|_F^2 + \|\mathbf{K}\widetilde{\mathbf{V}}_\parallel\widetilde{\mathbf{S}}_\sigma\widetilde{\mathbf{V}}_\parallel^\top\mathbf{K}^{-1} - \widetilde{\mathbf{V}}_\parallel\widetilde{\mathbf{S}}_\sigma\widetilde{\mathbf{V}}_\parallel^\top\|_F^2,\end{aligned} \tag{80}$$

where

$$\begin{aligned}\|\mathbf{K}\widetilde{\mathbf{V}}_\parallel\widetilde{\mathbf{V}}_\parallel^\top\mathbf{K}^{-1} - \widetilde{\mathbf{V}}_\parallel\widetilde{\mathbf{V}}_\parallel^\top\|_F^2 &= \|\mathbf{K}\widetilde{\mathbf{V}}_\parallel\widetilde{\mathbf{V}}_\parallel^\top\mathbf{K}^{-1}]\|_F^2 + \|\widetilde{\mathbf{V}}_\parallel\widetilde{\mathbf{V}}_\parallel^\top\|_F^2 - 2\operatorname{Tr}(\mathbf{K}\widetilde{\mathbf{V}}_\parallel\widetilde{\mathbf{V}}_\parallel^\top\mathbf{K}^{-1}\widetilde{\mathbf{V}}_\parallel\widetilde{\mathbf{V}}_\parallel^\top)\\ &\le \left(\frac{\kappa_+^2}{\kappa_-^2} + 1 - \frac{2\kappa_-}{\kappa_+}\right)\|\widetilde{\mathbf{V}}_\parallel\widetilde{\mathbf{V}}_\parallel^\top\|_F^2 = \left(\frac{\kappa_+^2}{\kappa_-^2} + 1 - \frac{2\kappa_-}{\kappa_+}\right)d_y.\end{aligned} \tag{81}$$

Combining Eq. (81), Assumption B.1 and Eq. (80), we cane get

$$g_3 \leq \tau + \left(\frac{\kappa_+^2}{\kappa_-^2} + 1 - \frac{2\kappa_-}{\kappa_+}\right) d_y. \tag{82}$$

Now combine Eqs. (75), (79) and (82), we have

$$g = \sum_{i\in[4]} g_i \leq \left(2 + \frac{\|\mathbf{A}\|^2}{\sigma^2}\right)\tau + \kappa_l d_y + \sigma^2 \sum_{i\in[d_y]} \left(\frac{\kappa_+^2}{\widetilde{\zeta}_i^2} + \frac{1}{\zeta_i^2} - \frac{2(\kappa_- + \epsilon(\kappa_- - 1))}{\zeta_i \widetilde{\zeta}_i}\right), \tag{83}$$

where

$$\kappa_l = \frac{\kappa_+^2}{\kappa_-^2} - \frac{2\kappa_-}{\kappa_+} + 1. \tag{84}$$

By combining Eq. (83) and Eq. (56), and substituting them into the closed-form expression for $\mathbb{D}_l$ given in Proposition 3.2, we obtain the bound for $\mathbb{D}_l$ as presented in Eq. (24).

□

C PROOFS OF SECTION 4

We begin this section by presenting the supporting lemma in Section C.1, followed by the proof of Theorem 4.3 in Section C.2. Finally, we provide the proof of Theorem 4.4 in Section C.3. For the ease of notation, given $s \in (0, 1/2)$, we define constant $r(s)$ by

$$r(s) = 2 + 2\max\left\{\frac{-\log^{1/4}(s)}{d^{1/4}}, \frac{-\log^{1/2}(s)}{d^{1/2}}\right\}. \tag{85}$$

C.1 Supporting lemma

Our proof is based on the theory developed in Dwivedi et al. (2019); Grenioux et al. (2023). We show the general mixing time bound of the strongly log-concave distribution for IMH in Proposition C.3. We first denote π as the ground truth posterior, and $\widehat{\pi}$ denote the proposal distribution in the MH step. Define the log-weight function by

$$\widehat{w}(x) = \log \frac{\pi(x)}{\widehat{\pi}(x)}.$$

Assumption C.1.

- $x^* = \arg\max \log \pi(x)$, $\widehat{x}^* = \arg\max \log \widehat{\pi}(x)$.
- *π satisfies isoperimetric inequality with isoperimetric constant $\psi(\pi)$ i.e., for any partition $\{\mathsf{S}_1, \mathsf{S}_2, \mathsf{S}_3\}$ of $\mathbb{R}^d$, we have*

$$\pi(\mathsf{S}_3) \geq \psi(\pi)\mathrm{dist}(\mathsf{S}_1, \mathsf{S}_2)\pi(\mathsf{S}_1, \mathsf{S}_2), \tag{86}$$

  where $\mathrm{dist}(\mathsf{S}_1, \mathsf{S}_2) = \inf\{\|x - y\| : x \in \mathsf{S}_1, y \in \mathsf{S}_2\}$.
- *For all $R > 0$, there exists $C_a(R) \geq 0$ such that for all $(x, x') \in \mathsf{Ball}(x^*, R)$, $\|\widehat{w}(x) - \widehat{w}(x')\| \leq C_R\|x - x'\|$.*

Assumption C.2 (α, s)**.** *There exists $\Delta \in (0, 1)$, $\delta_\Delta > 0$ and $R > 0$ satisfying*

$$\int_{B_\Delta} \pi(dx)\widehat{\pi}(dy) \geq 1 - \delta_\Delta, \tag{87}$$

where

$$B_\Delta = \{(x, y) \in \mathsf{Ball}(x^*, R) \times \mathsf{Ball}(x^*, R) : \|x - y\| \leq R_\Delta\}, \quad \textit{where} \quad R_\Delta = \frac{-\log(\Delta)}{C_R}, \tag{88}$$

and

$$\frac{1}{1-\alpha}\delta_\Delta \leq \min\left(\frac{s}{4}, \frac{(2\alpha - 1)\psi(\pi)s}{64C_R}\right). \tag{89}$$

Proposition C.3 (Corollary E.5 in Grenioux et al. (2023)). *Let $\alpha(0,1/2)$, $\epsilon \in (0,1)$ and g a β-warm initial distribution with respect to π. Assume that Assumption C.1 and Assumption C.2 hold, then we have the following upper bound on the mixing time*

$$\tau_{mix}(g,\epsilon) \leq \frac{8}{(1-\alpha^2)} \log\left(\frac{2\beta}{\epsilon}\right) \max\left(1, \frac{64^2 C_R^2}{\psi(\pi)^2(2\alpha-1)^2}\right). \tag{90}$$

C.2 Proof of Theorem 4.3

Proof. Denote $s = \epsilon/2\beta$, let $c_1 > 0$ be a constant such that $c_1 s \in (0,1/2)$. By Lemma 1 in Dwivedi et al. (2019), we have $\pi(\mathsf{Ball}(x^*, R_1)) \geq 1 - c_1 s$, where

$$R_1 = \sqrt{\frac{d}{m}} r(c_1 s). \tag{91}$$

1. `Approx-IMH`: since $-\log p(x)$ is m-strongly convex, $-\log \pi_a(x)$ (Eq. (4)) is also m-strongly convex. Then let $c_2 > 0$ be a constant s.t. $c_2 s \in (0,1/2)$, we have $\pi_a(\mathsf{Ball}(x_a^*, R_2)) \geq 1 - c_2 s$, where

$$R_2 = \sqrt{\frac{d}{m}} r(c_2 s). \tag{92}$$

Define $R_{2,a} = R_2 + \|x^* - x_a^*\|$, $R_\Delta = R_1 + R_{2,a}$ and $R = \max(R_1, R_{2,a})$. Then for any $(x,y) \in \mathbf{Ball}(x^*, R_1) \times \mathbf{Ball}(x_a^*, R_{2,a})$, we have

$$\begin{aligned} &\|x - y\| \leq \|x^* - x_a^*\| + R_1 + R_2 = R_\Delta, \\ &\|x - x^*\| \leq R_1 \leq R, \qquad \|y - x^*\| \leq R_{2,a} + \|x^* - x_a^*\| \leq R. \end{aligned}$$

Thus for the set B_Δ defined in Assumption C.2, we have

$$\mathbf{Ball}(x^*, R_1) \times \mathbf{Ball}(x_a^*, R_{2,a}) \subset B_\Delta. \tag{93}$$

Thus the condition of Eq. (87) in Assumption C.2 is satisfied since

$$\begin{aligned} \int_{B_\Delta} \pi(dx)\pi_a(dy) &\geq \int_{\mathbf{Ball}(x^*,R)\times\mathbf{Ball}(x_a^*,R_{2,a})} \pi(dx)\pi_a(dy) \\ &\geq (1 - c_1 s)(1 - c_2 s) \geq 1 - (c1 + c2)s =: 1 - \delta_\Delta. \end{aligned} \tag{94}$$

Now we need to determine c_1 and c_2 such that Eq. (89) is satisfied:

$$\frac{\delta_\Delta}{1-\alpha} \leq \min\left(\frac{s}{4}, \frac{(2\alpha-1)s}{64 C_R}\psi(\pi)\right) \Longleftrightarrow \begin{cases} \delta_\Delta \leq \frac{1-\alpha}{4} s \\ \delta_\Delta \leq \frac{(1-\alpha)(2\alpha-1)}{64 C_R}\psi(\pi) s. \end{cases} \tag{95}$$

Let $\alpha = 3/4$, by Eq. (95) and Eq. (94), it is enough if the following conditions are satisfied:

$$c_1 + c_2 \leq \frac{1}{16}, \qquad C_R \leq \frac{\psi(\pi)}{512(c_1 + c_2)}. \tag{96}$$

By Theorem 4.4 in Cousins and Vempala, a m-strongly log concave distribution π is isoperimetric with the isoperimetric constant $\psi(\pi) = \log(2\sqrt{m})$. Let $c_1 = 1/17$, $c_2 = 1/16 - 1/17$. Thus if

$$C_a(R_a) \leq \frac{\log 2\sqrt{m}}{32}, \tag{97}$$

where

$$R_a = \max\left(\sqrt{\frac{d}{m}} r\left(\frac{\epsilon}{17\beta}\right), \sqrt{\frac{d}{m}} r\left(\frac{\epsilon}{272\beta}\right) + \|x^* - x_a^*\|\right), \tag{98}$$

conditions of (96) are satisfied and thus Assumption C.1 and Assumption C.2 are satisfied. Then by Proposition C.3, we have

$$\tau_{mix}^a(\epsilon) \leq 128 \log\left(\frac{2\beta}{\epsilon}\right) \max\left(1, \frac{128^2 C_a(R_a)^2}{(\log 2)^2 m}\right). \tag{99}$$

2. `Latent-IMH`: the result can be proved by an argument similar to the proof for `Approx-IMH`. By Eq. (6), we have $\pi_l(x) \propto q(y - \mathbf{A}x)p(\widetilde{\mathbf{F}}^{-1}\mathbf{F}x)$. It is easy to see that $\pi_l(x)$ is $m\sigma^2_{\min}(\widetilde{\mathbf{F}}^{-1}\mathbf{F})$. Thus we have $\pi_l(\mathsf{Ball}(x_l^*, R_2)) \geq 1 - c_2 s$, where

$$R_2 = \sqrt{\frac{d}{m}} \frac{1}{\sigma_{\min}(\widetilde{\mathbf{F}}^{-1}\mathbf{F})} r(c_2 s). \tag{100}$$

Then if $C_l(R_l) \leq \frac{\log 2\sqrt{m}}{32}$, where

$$R_l = \max\left(\sqrt{\frac{d}{m}} r\left(\frac{\epsilon}{17\beta}\right), \sqrt{\frac{d}{m}} \frac{1}{\sigma_{\min}(\widetilde{\mathbf{F}}^{-1}\mathbf{F})} r\left(\frac{\epsilon}{272\beta}\right) + \|x^* - x_l^*\|\right), \tag{101}$$

conditions of Eq. (96) are satisfied and the bound for mixing time of `Latent-IMH` is proved.

□

C.3 Proof of Theorem 4.4

Proof. Without loss of generality, we assume that the posterior mode x^* is located at the origin. The key step is to establish the local Lipschitz continuity of the log-weight function on the ball $\mathsf{Ball}(x^*, R_a)$.

1. `Approx-IMH`: the log-weight function has the following form:

$$\begin{aligned} w_a(x) = \log \frac{\pi(x)}{\pi_a(x)} &= -\frac{1}{2\sigma^2}\left[\|y - \mathbf{A}x\|^2 - \|y - \widetilde{\mathbf{A}}x\|^2\right] \\ &= -\frac{1}{2\sigma^2}\left[x^\top(\mathbf{A} - \widetilde{\mathbf{A}})^\top\big(2y - (\mathbf{A} + \widetilde{\mathbf{A}})x\big)\right] \\ &= -\frac{1}{2\sigma^2}\left[x^\top \widetilde{\mathbf{C}} x + 2x^\top \Delta\mathbf{A}^\top y\right], \end{aligned} \tag{102}$$

where we define

$$\widetilde{\mathbf{C}} := \mathbf{A}^\top\mathbf{A} - \widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}} = \mathrm{Diag}\big(\underbrace{\cdots, (1-\alpha_i^2)s_i^2, \cdots}_{\text{first } d_y \text{ coordinates}}, 0, \cdots\big), \quad \Delta\mathbf{A} := \mathbf{A} - \widetilde{\mathbf{A}}. \tag{103}$$

Thus for any $x_1, x_2 \in \mathsf{Ball}(x^*, R_a)$, we have

$$\begin{aligned} |w_a(x_1) - w_a(x_2)| &= \frac{1}{2\sigma^2}\left|x_1^\top \widetilde{\mathbf{C}} x_1 + 2x_1^\top \Delta\mathbf{A}^\top y - x_2^\top \widetilde{\mathbf{C}} x_2 - 2x_2^\top \Delta\mathbf{A}^\top y\right| \\ &\leq \frac{1}{2\sigma^2}\left|x_1^\top \widetilde{\mathbf{C}} x_1 - x_2^\top \widetilde{\mathbf{C}} x_2\right| + \frac{1}{\sigma^2}\left|(x_1 - x_2)^\top \Delta\mathbf{A}^\top y\right|. \end{aligned} \tag{104}$$

For the first term of the last inequality of Eq. (104), we have

$$\begin{aligned} \left|x_1^\top \widetilde{\mathbf{C}} x_1 - x_2^\top \widetilde{\mathbf{C}} x_2\right| &= |(x_1 - x_2)^\top \widetilde{\mathbf{C}}(x_1 + x_2)| \\ &\leq \|\widetilde{\mathbf{C}}\|\|x_1 + x_2\|\|x_1 - x_2\| \leq 2R_a \left(\max_{i\in[d_y]} |1 - \alpha_i^2| s_i^2\right) \|x_1 - x_2\|. \end{aligned} \tag{105}$$

For the second term of last inequality of Eq. (104), we have

$$\left|(x_1 - x_2)^\top \Delta\mathbf{A}^\top y\right| \leq \left(\max_{i\in[d_y]} |1 - \alpha_i| s_i\right) \|y\|\|x_1 - x_2\|. \tag{106}$$

Substitute Eqs. (105) and (106) into Eq. (104), we have

$$|w_a(x_1) - w_a(x_2)| \leq \frac{1}{2\sigma^2}\left\{2R_a \left(\max_{i\in[d_y]} |1 - \alpha_i^2| s_i^2\right) + \left(\max_{i\in[d_y]} |1 - \alpha_i| s_i\right) \|y\|\right\} \|x_1 - x_2\|. \tag{107}$$

By Theorem 4.3, $R_a \sim 2\sqrt{d/m}$ when d is large. Thus, the Lipschitz constant scale as the following when d is large,

$$C_a(R_a) \sim \frac{d}{m} \max_{i\in[d_y]} |1-\alpha_i^2| \frac{s_i^2}{\sigma^2} \tag{108}$$

2. `Latent-IMH`: the log-weight function has the following form

$$\begin{aligned} w_l(x) &= \log \frac{\pi(x)}{\widehat{\boldsymbol{\pi}}(x)} = \log \frac{p(x)}{p(\widetilde{\mathbf{F}}^{-1}\mathbf{F}x)} \\ &= -\frac{1}{2}\big[\|x\|^2 - \|\widetilde{\mathbf{F}}^{-1}\mathbf{F}x\|^2\big]. \end{aligned} \tag{109}$$

Then

$$\begin{aligned} |w_l(x_1) - w_l(x_2)| &= \frac{1}{2}\left|\|x_1\|^2 - \|x_2\|^2 - \|\widetilde{\mathbf{F}}^{-1}\mathbf{F}x_1\|^2 + \|\widetilde{\mathbf{F}}^{-1}\mathbf{F}x_2\|^2\right| \\ &= \frac{1}{2}|x_1^\top \widehat{\mathbf{C}} x_1 - x_2^\top \widehat{\mathbf{C}} x_2|, \end{aligned} \tag{110}$$

where we define

$$\widehat{\mathbf{C}} := \mathbf{I} - (\widetilde{\mathbf{F}}^{-1}\mathbf{F})^\top (\widetilde{\mathbf{F}}^{-1}\mathbf{F}) = \mathrm{Diag}\left(1 - \frac{1}{\alpha^2}\right)_{i\in[d]}. \tag{111}$$

Thus

$$\begin{aligned} |w_l(x_1) - w_l(x_2)| &= \frac{1}{2}\left|(x_1-x_2)^\top \widehat{\mathbf{C}}(x_1+x_2)\right| \\ &\le \frac{1}{2}\|\widehat{\mathbf{C}}\|\|x_1+x_2\|\|x_1-x_2\| \le R_l \left(\max_{i\in[d]}\left|1-\frac{1}{\alpha_i^2}\right|\right)\|x_1-x_2\|. \end{aligned} \tag{112}$$

By Theorem 4.3, $R_l \sim \sqrt{d/m}$ when d is large. Thus, the Lipschitz constant scale as the following when d is large,

$$C_l(R_l) \sim \frac{d}{m} \max_{i\in[d]}\left|1-\frac{1}{\alpha_i^2}\right|. \tag{113}$$

Substitute results for $C_a(R_a)$ and $C_l(C_l)$ into Theorem 4.3, we can get the results for mixing time.

□

D ADDITIONAL EXPERIMENTAL DETAILS

All experiments were conducted on the Vista 600 Grace Hopper (GH) nodes at the Texas Advanced Computing Center (TACC).

D.1 Sensitivity test in Section 3

The design of the sensitivity tests for comparing the KL-divergence between `Approx-IMH posterior`, `Latent-IMH posterior` with `Exact posterior` is presented in Table 2. In Figure 7 shows both the absolute KL-divergence and relative KL-divergence results for the sensitivity tests.

D.2 Additional details for Section 5

Comparison with delayed-acceptance MCMC. We compare the sampling efficiency of different MCMC methods under two approximate operators (with varying levels of accuracy) in Table 3. Both `Latent-2stage` and `Approx-2stage` improve efficiency relative to standard MALA. For example, to achieve comparable accuracy, `Latent-2stage` requires only about 25% of the exact forward/inverse solves used by MALA. However, despite these improvements, both two-stage methods still underperform compared to Latent-IMH.

Table 2: Experimental setup for comparing `Approx-IMH posterior` and `Latent-IMH posterior` with `Exact posterior`.

Effects	$\log_{10}$ SNR	$\frac{\|\mathbf{F}-\widetilde{\mathbf{F}}\|}{\|\mathbf{F}\|}$	d_y/d	d
Noise level	0.5 —4.0	6%	0.2	500
Spectral error	2.5	2% —21%	0.2	500
Observation ratio	2.5	6%	0.05 —0.5	500
Dimension	2.5	6%	0.2	100 —2000

Table 3: Relative mean error comparisons between different MCMC sampling methods. Each entry shows: (# exact forward/adjoint solves, relative mean error). The approximate operator used in Latent-2stage, Approx-2stage, and Latent-IMH has approximation error $\|\mathbf{I}-\widetilde{\mathbf{F}}^{-1}\mathbf{F}\| = \mathbf{2.4}\%$ (top) and $\|\mathbf{I}-\widetilde{\mathbf{F}}^{-1}\mathbf{F}\| = \mathbf{9.0}\%$ (bottom).

Latent-2stage	**Approx-2stage**	**MALA**	**NUTS**	**Latent-IMH**
(1.2×10^5, 0.37)	(1.2×10^5, 0.38)	(5.0×10^5, 0.36)	(1.8×10^5, 0.20)	(1.0×10^3, 0.32)
(2.4×10^6, 0.08)	(2.4×10^6, 0.088)	(1.0×10^7, 0.08)	(7.5×10^5, 0.066)	(2.0×10^4, 0.07)
(4.8×10^6, 0.057)	(4.8×10^6, 0.062)	(2.0×10^7, 0.055)	(1.3×10^6, 0.044)	(2.0×10^5, 0.023)

Latent-2stage	**Approx-2stage**	**Latent-IMH**
(1.2×10^5, 0.36)	(1.3×10^5, 0.49)	(5.0×10^3, 0.59)
(2.4×10^6, 0.084)	(2.6×10^6, 0.11)	(1.0×10^5, 0.12)
(4.9×10^6, 0.059)	(5.2×10^6, 0.088)	(5.0×10^5, 0.056)

Evaluation Metric – Maximum Mean Discrepancy (MMD). MMD is a kernel-based distance between two distributions P and Q, defined as

$$\mathrm{MMD}^2(P,Q) = \mathbb{E}_{x,x'\sim P}[k(x,x')] + \mathbb{E}_{y,y'\sim Q}[k(y,y')] - 2\,\mathbb{E}_{x\sim P,y\sim Q}[k(x,y)],$$

where $k(\cdot,\cdot)$ is a positive-definite kernel. In our experiments, we use the *RBF kernel*

$$k(x,y) = \exp\Big(-\gamma\|x-y\|^2\Big),$$

and set the kernel parameter γ using the *median of pairwise distances* between all ground truth samples (the median heuristic). Intuitively, a smaller MMD indicates that the generated samples are closer to the ground-truth distribution in the reproducing kernel Hilbert space defined by k.

Ill-conditioned Gaussian prior. The sample efficiency comparison for the ill-conditioned Gaussian prior test is presented in Figure 8.

Scattering problem. For the TV prior $p(x)\exp(-\lambda\mathrm{TV}(x))$, we use a smoothed TV distance using finite difference:

$$TV_\epsilon(x) = \sum_{i,j}\sqrt{(x_{i+1,j}-x_{i,j})^2 + (x_{i,j+1}-x_{i,j})^2 + \epsilon^2}. \tag{114}$$

Here $x_{i,j}$ is the pixel at row i and column j. We use $\epsilon > 0$ ensures differentiability for gradient-based samplers like NUTS.

To assess the convergence of the samples from `Approx-IMH` and `Latent-IMH`, we compare the sample mean and variance of the generated samples with those computed from ground-truth posterior samples.

$$\mu_{\text{true}} = \frac{1}{N_{\text{true}}}\sum_{i=1}^{N_{\text{true}}} x_i^{\text{true}}, \qquad \sigma^2_{\text{true}} = \frac{1}{N_{\text{true}}-1}\sum_{i=1}^{N_{\text{true}}}\left(x_i^{\text{true}} - \mu_{\text{true}}\right)^2. \tag{115}$$

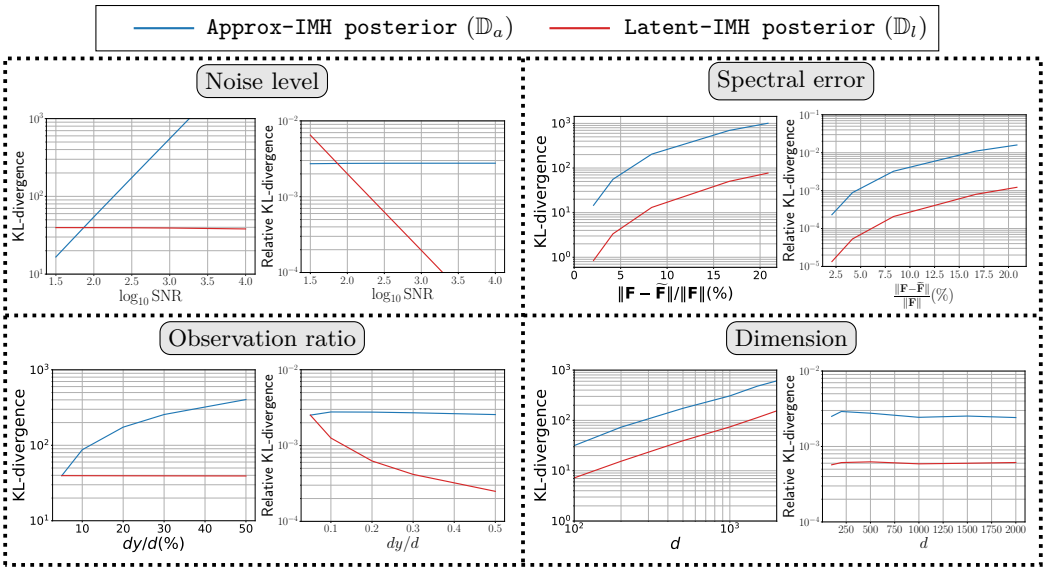

Figure 7: Sensitivity test results for the expected KL-divergence of `Approx-IMH posterior` and `Latent-IMH posterior` relative to `Exact posterior`.

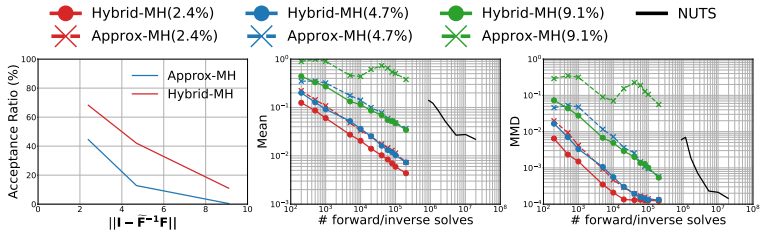

Figure 8: Sample efficiency comparison for ill-conditioned Gaussian prior $p(x) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, where $\kappa(\mathbf{\Sigma}) \approx 1000$. The results in the plots reflect the average of 5 independent runs for each test. The numbers in parentheses in the top legend for `Approx-IMH` and `Latent-IMH` indicate the spectral error $\|\mathbf{I} - \widetilde{\mathbf{F}}^{-1}\mathbf{F}\|_2$. The numbers in parentheses above the histogram plots denote the total number of forward and inverse solves of $\mathbf{F}$ required by each sampler.

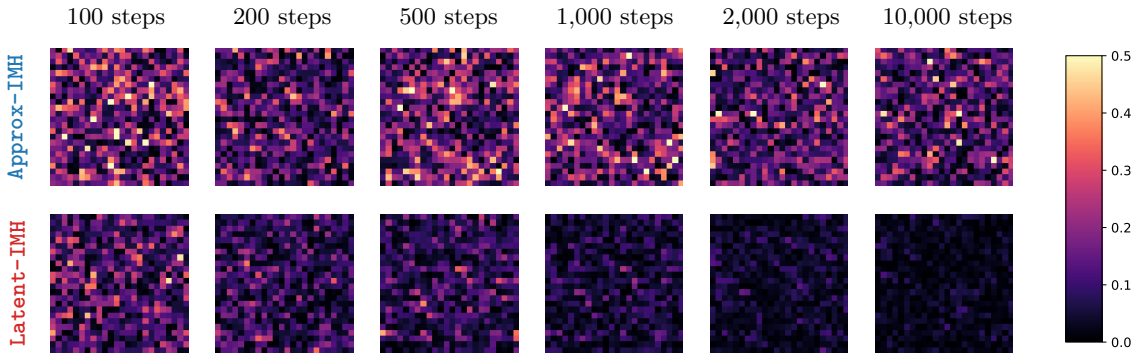

Figure 9: Convergence of the mean error map for the scattering problem test.

and for the MCMC chain up to iteration t:

$$\hat{\mu}_t = \frac{1}{t}\sum_{i=1}^{t} x_i^{\text{mcmc}}, \qquad \hat{\sigma}_t^2 = \frac{1}{t-1}\sum_{i=1}^{t}\left(x_i^{\text{mcmc}} - \hat{\mu}_t\right)^2. \tag{116}$$

The ground-truth samples are generated using NUTS for the true posterior, collecting 2×10^5 samples from 10

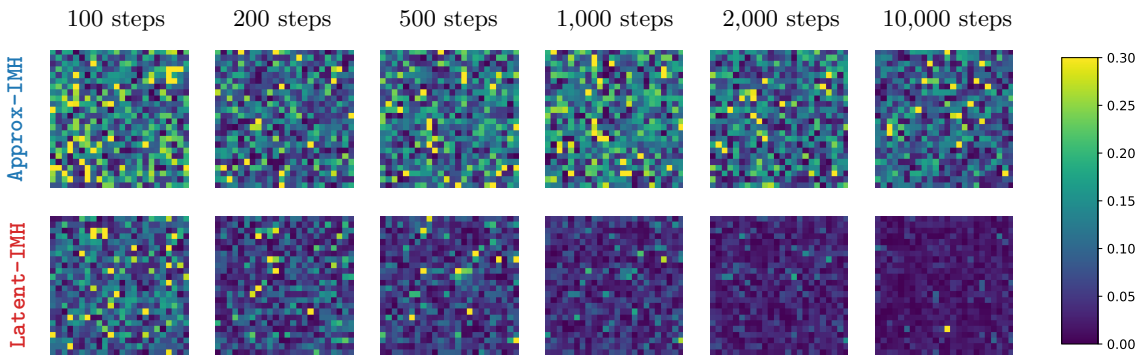

Figure 10: Convergence of the variance error map for the scattering problem test.

independent runs. Both μ and σ^2 are images, so we visualize convergence via the mean error map $|\hat{\mu}t - \mu\text{true}|$ and variance error map $|\hat{\sigma}t^2 - \sigma\text{true}^2|$ shown in Figure 9 and Figure 10. The plots indicate that `Latent-IMH` achieves faster convergence than `Approx-IMH`.