Corrected Integrated Laplace Approximation for Bayesian Inference in Latent Gaussian Models

Jinlin Lai
Manning College of Information and Computer Sciences
University of Massachusetts Amherst
jinlinlai@cs.umass.edu

Charles C. Margossian
Department of Statistics
University of British Columbia
charles.margossian@ubc.ca

Daniel Sheldon
Manning College of Information and Computer Sciences
University of Massachusetts Amherst
sheldon@cs.umass.edu

Abstract

Latent Gaussian models (LGMs) are a popular class of Bayesian hierarchical models that include Gaussian processes, as well as certain spatial models and mixed-effect models. Efficient Bayesian inference of LGMs often requires marginalizing out the latent variables. For LGMs with a non-Gaussian likelihood, exact marginalization is not possible and a popular approach is to do approximate marginalization with an integrated Laplace approximation (ILA). Using ILA produces an approximate posterior which, in some settings, can differ significantly from the correct posterior, which impacts downstream applications. We propose an importance sampling scheme to correct the error introduced by ILA. By increasing the number of samples in importance sampling, the posterior with ILA converges to the correct posterior. This idea is realized with various techniques, including pseudo-marginalization, quasi-Monte Carlo and randomized quasi-Monte Carlo. We implement our methods in an automatic differentiation framework to support gradient-based algorithms when doing inference on the hyperparameters. For the latter, we specifically consider the use of Hamiltonian Monte Carlo. We demonstrate the benefits of reduced error in various applied models.

1 Introduction

Latent Gaussian models (LGMs) are a popular class of Bayesian models, formulated as a joint density $\pi(\theta, z, y) = \pi(\theta)\pi(z|\theta)\pi(y|\theta, z)$, where θ is often termed the "hyperparameter", z is the latent Gaussian variable and y is the observation. In detail, the hierarchical prior $\pi(z|\theta)$ is Gaussian. Examples of LGMs include Gaussian process models [46] and generalized linear models [32]. Doing Bayesian inference on such models amounts to computing the joint posterior distribution $\pi(\theta, z|y)$. This is a well-known challenge because z is often high-dimensional and the posterior has the intricate geometry characteristic of hierarchical models [5, 35].

A common approach to tackle this problem is to marginalize out z and then perform Bayesian inference on $\pi(\theta|y)$, a target with a lower dimension and often a simpler geometry. Posterior samples

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

for z are then obtained in a post-inference step. Marginalization can be done exactly when $\pi(y|\theta, z)$ is Gaussian, else it must be done *approximately* with a Laplace approximation [41, 46, 38]. This idea underlies the popular *integrated nested Laplace approximation* [INLA, 38]. INLA works effectively when $\pi(\theta|y)$ is well characterized around the mode and low-dimensional ("2–5, not more than 20" [39]). To overcome these limitations and scale to cases where θ is high-dimensional, it has been proposed to run Markov chain Monte Carlo (MCMC) over $\pi(\theta|y)$ [25, 30, 27] and in particular gradient-based samplers, such as Hamiltonian Monte Carlo [HMC, 17, 31] which are known to scale favorably with dimension. For this, Margossian et al. [27] developed *adjoint-differentiated Laplace approximation* (ADLA), an automatic differentiation algorithm which efficiently computes the gradient of the approximate marginal log posterior, $\nabla_\theta \log \hat{\pi}(\theta|y)$, obtained with an integrated Laplace approximation.

Unfortunately, all the above described methods are subject to the approximation error introduced by the Laplace approximation. An alternative is to simply run MCMC over the full posterior $\pi(\theta, z|y)$ but this approach, while asymptotically exact, is often computationally expensive, if not infeasible. In this work, we develop methods that both benefit from marginalization with Laplace approximation and give an asymptotically correct posterior.

Our main idea is importance sampling. We show that the integrated Laplace approximation can be reformulated as an importance sampling estimator with one sample and the error of this estimator can be reduced with multiple samples. First, we propose ADLA using pseudo-marginalization [3], an approach which targets an extended distribution over the model parameters and auxiliary random variables used for importance sampling, and we show that this approach is asymptotically correct. Next, we propose to fix the randomness using quasi-Monte Carlo (QMC). While QMC-based approaches remain asymptotically correct, we find that in practice they can have a large error, even when the number of samples used for importance sampling is large. Finally, we unify pseudo-marginalization and QMC and propose ADLA with randomized QMC, an approach which utilizes the Laplace approximation and is asymptotically correct. We implement the proposed methods in a reverse-mode automatic differentiation framework and test them across several applications. We find that our proposed methods provide reduced error with the computational benefits of marginalization. The code is available at `https://github.com/lll6924/cila.git`.

2 Background

2.1 Latent Gaussian models

LGMs observe the following hierarchical structure,

$$\theta \sim \pi(\theta),\ z \sim \text{normal}(0, K(\theta)),\ y \sim \pi(y|\theta, z), \tag{1}$$

where θ is the hyperparameter, z comprises the latent Gaussian effects, and y is the observation. Here, the prior Gaussian on z is centered at 0 but off-set from 0 can be encoded in the likelihood $\pi(y|\theta, z)$. Without loss of generality, we split θ into two, $\theta = [\xi, \eta]$, where ξ parametrizes the prior covariance K, meaning $K(\theta) = K(\xi)$, and η parametrizes the likelihood, meaning $\pi(y|\theta, z) = \pi(y|\eta, z)$.

2.2 Adjoint-differentiated Laplace Approximation

A general way of approximately marginalizing z from LGMs is the Laplace approximation. In detail, the Laplace approximation approximates the conditional posterior $\pi(z|\theta, y)$ by a Gaussian, which matches the mode and curvature of the conditional posterior,

$$\hat{z} = \arg\max_z \log \pi(y, z|\theta), \quad \Sigma^{-1} = -\nabla_z \nabla_z \log \pi(y, \hat{z}|\theta), \tag{2}$$

where $\hat{z}$ can be obtained from any optimization algorithm (Newton's method is often used in practice). Then $\pi(z|\theta, y)$ is approximated by $\hat{\pi}(z|\theta, y) = \text{normal}(\hat{z}, \Sigma)$. The approximation is exact when the likelihood $\pi(y|\theta, z)$ is normal and so, by conjugacy, $\pi(z|\theta, y)$ is also normal. In practice, this approximation is well justified when $\pi(y|\theta, z)$ is log-concave [40], although it has also been explored in other contexts [e.g., 42].

For any fixed z, the marginalized model can be approximated by

$$\hat{\pi}_z(\theta, y) \coloneqq \frac{\pi(\theta)\pi(z|\theta)\pi(y|\theta, z)}{\hat{\pi}(z|\theta, y)} \approx \frac{\pi(\theta)\pi(z|\theta)\pi(y|\theta, z)}{\pi(z|\theta, y)} = \pi(\theta, y). \tag{3}$$

Table 1: Comparison of different methods of marginalization on posterior correctness, dimension change and continuity of the objective.

Algorithm	Correct posterior	Dimension	Continuity
ADLA [28]	✗	Lower	✓
PM-ADLA (Sec. 4)	✓	Possibly Higher	✓
QMC-ADLA (Sec. 5)	✓(Asymptotically)	Lower	✓
RQMC-ADLA (Sec. 5.1)	✓	Same	✗

This approximation is valid for any z, but $z = \hat{z}$ from the optimizer is usually used. Note that $\hat{z} = \hat{z}(\theta)$ is a function of θ. We omit this dependence here for simplicity, but it is important when computing the gradients.

With approximation (3), we can run inference on a lower-dimensional model $\hat{\pi}_{\hat{z}}(\theta, y)$ and get an approximate marginal posterior $\hat{\pi}_{\hat{z}}(\theta|y)$. In this work, we focus on Hamiltonian Monte Carlo (HMC) [17, 31], an MCMC sampler that scales well in high-dimensions but requires computing $\nabla_\theta \log \hat{\pi}_{\hat{z}}(\theta, y)$. The calculation of this gradient must be done carefully: in particular, we need to account for the fact that $\hat{z}$ depends implicitly on θ and that the Laplace approximation is obtained using second-order derivatives of $\log \pi(z|y, \theta)$ to compute Σ. Calculating the gradient can be done efficiently using the implicit function theorem and adjoint methods of automatic differentiation, as in the adjoint-differentiated Laplace approximation (ADLA) [27, 28]. Appendix B provides details on ADLA. Here, we implement ADLA in JAX [8], a high-performance automatic differentiation library.

HMC generates asymptotically unbiased samples from the approximate marginal posterior $\hat{\pi}_{\hat{z}}(\theta|y)$. When the approximation in Eq. (3) is not exact, the posterior approximation from HMC differs from the true posterior $\pi(\theta|y)$, even asymptotically. We call this issue 'posterior error' and propose to correct the error by closing the gap in Eq. (3). In this paper, we construct approximate marginal models with importance sampling that better approximate $\pi(\theta, y)$, and then derive an algorithm to compute gradients.

3 Reducing error with importance sampling

We propose an unbiased estimator of the marginal model $\pi(\theta, y)$. This estimator is similar to the approximation in Eq. (3), except that the Laplace approximation is not evaluated at the mode $\hat{z}$ but at a random point z drawn from the Laplace approximation.

Proposition 1. *Consider the Laplace approximation $\hat{\pi}(z|\theta, y)$ in Eq. (2) and let*

$$\hat{\pi}_z(\theta, y) = \frac{\pi(\theta)\pi(z|\theta)\pi(y|\theta, z)}{\hat{\pi}(z|\theta, y)}, \tag{4}$$

where $z \sim \hat{\pi}(z|\theta, y)$. Then $\hat{\pi}_z(\theta, y)$ is an unbiased estimator of $\pi(\theta, y)$.

Proof. Observe that,

$$\hat{\pi}_z(\theta, y) = \pi(\theta, y)\frac{\pi(z|\theta, y)}{\hat{\pi}(z|\theta, y)}.$$

Here, we may recognize that $\hat{\pi}_z(\theta, y)$ is an importance sampling estimator of $\pi(\theta, y)$ and so must be unbiased. In detail,

$$\mathbb{E}[\hat{\pi}_z(\theta, y)] = \int \pi(\theta, y)\frac{\pi(z|\theta, y)}{\hat{\pi}(z|\theta, y)}\hat{\pi}(z|\theta, y)\mathrm{d}z = \pi(\theta, y)\int \pi(z|\theta, y)\mathrm{d}z = \pi(\theta, y).$$

□

If we have n i.i.d. samples, $z_1, \ldots, z_n \sim \hat{\pi}(z|\theta, y)$, we can construct an averaged estimator for the true marginal

$$\hat{\pi}_{z_{1:n}}(\theta, y) = \frac{1}{n}\sum_{i=1}^{n} \hat{\pi}_{z_i}(\theta, y).$$

This estimator is also unbiased from linearity of expectation, and its variance is of order $\mathcal{O}(n^{-1})$ if the variance of the importance sampling weights is finite.

In order to use our unbiased estimator of the marginal posterior in HMC, we must overcome two challenges. First, standard HMC only works for a fixed target density, rather than a target density that varies with an auxiliary random variable z_i. In the next two sections, we consider different ways to handle this random target density. Second, we must generalize the automatic differentiation algorithm used by ADLA to compute $\nabla_\theta \log \hat{\pi}_{z_{1:n}}(\theta, y)$. Indeed, in the classical setting of the integrated Laplace approximation, it suffices to propagate derivatives through $\hat{z}$. Here, we must propagate derivatives through z_i. We detail this procedure in Appendix B.

4 Pseudo-marginal ADLA

To build a fixed approximate marginal model, our first idea is to target an extended distribution that includes the auxiliary random variables z_i introduced to construct the approximate marginal model in Eq. (4). This approach is known as pseudo-marginalization in the literature [3]. In detail, we express each sample z_i as a transformation over some random auxiliary variables ϵ_i,

$$z_i = \mathcal{T}_{\theta,y}(\epsilon_i) \coloneqq \hat{z} + \sqrt{\Sigma}\epsilon_i, \text{ where } \epsilon_i \sim \pi(\epsilon_i) = \text{normal}(0, I),$$

for $i = 1, 2, ..., n$. Here $\sqrt{\Sigma}$ is a matrix square root such that $\sqrt{\Sigma}\sqrt{\Sigma}^T = \Sigma$, so each z_i still follows $\hat{\pi}(z|\theta, y) = \text{normal}(\hat{z}, \Sigma)$. Then, our approximate marginal model is the importance sampling estimator multiplied by the density of the auxiliary variables:

$$\hat{\pi}^{\text{PM}}(\theta, \epsilon_{1:n}, y) \coloneqq \prod_{i=1}^{n} \pi(\epsilon_i) \left(\frac{1}{n} \sum_{i=1}^{n} \frac{\pi(\theta)\pi(\mathcal{T}_{\theta,y}(\epsilon_i)|\theta)\pi(y|\theta, \mathcal{T}_{\theta,y}(\epsilon_i))}{\hat{\pi}(\mathcal{T}_{\theta,y}(\epsilon_i)|\theta, y)} \right), \tag{5}$$

where we have replaced z_i with $\mathcal{T}_{\theta,y}(\epsilon_i)$ when computing the density. This approximate marginal model has several nice properties. First, it shares the same marginal over (θ, y) as the original model.

Proposition 2. *$\hat{\pi}^{\text{PM}}$, as defined in eq.* (5)*, admits $\pi(\theta, y)$ as its marginal.*

The proof is in Appendix A.1. By sampling from the extended posterior $\hat{\pi}^{\text{PM}}(\theta, \epsilon_{1:n}|y)$, we also obtain samples from the marginal posterior $\pi(\theta|y)$. Here, the Laplace approximation does the marginalization in a "pseudo" way: the latent Gaussian is marginalized out but we augment the state space with additional noise variables ϵ. We call this approach PM-ADLA.

We now argue that this model offers the benefits of the Laplace approximation and of importance sampling. To see this, write

$$\hat{\pi}^{\text{PM}}(\theta, \epsilon_{1:n}, y) = \pi(\theta, y) \prod_{i=1}^{n} \pi(\epsilon_i) \left(\frac{1}{n} \sum_{i=1}^{n} \frac{\pi(\mathcal{T}_{\theta,y}(\epsilon_i)|\theta, y)}{\hat{\pi}(\mathcal{T}_{\theta,y}(\epsilon_i)|\theta, y)} \right).$$

When n is large, the last term $n^{-1} \sum_{i=1}^{n} \pi(\mathcal{T}_{\theta,y}(\epsilon_i)|\theta, y)/\hat{\pi}(\mathcal{T}_{\theta,y}(\epsilon_i)|\theta, y)$ approaches 1 and the model is the true marginal $\pi(\theta, y)$ multiplied by n independent normal distributions. Then, the augmented model does not possess the intricate geometry of the original posterior $\pi(\theta, z|y)$, since ϵ is independent of θ.

Compared with ADLA, PM-ADLA corrects the posterior error, but does not reduce the problem dimension. Instead, if the dimension of z is d_z, the model dimension even increases by $(n-1)d_z$ for $n > 1$. In addition, at each step of HMC, PM-ADLA requires n evaluations of the model density and its gradient. (However, we still only need to compute the Laplace approximation once.)

5 ADLA with quasi-Monte Carlo

Instead of introducing auxiliary random variables and in doing so increasing the model dimension, we can instead fix those variables. A naive approach is to sample the auxiliary variables $\epsilon_1, \ldots, \epsilon_n$ independently once in the beginning and reuse them during inference to ensure a fixed approximate model. The approximate marginal model is

$$\hat{\pi}^{\text{IS}}_{\epsilon_{1:n}}(\theta, y) \coloneqq \frac{1}{n} \sum_{i=1}^{n} \frac{\pi(\theta)\pi(\mathcal{T}_{\theta,y}(\epsilon_i)|\theta)\pi(y|\theta, \mathcal{T}_{\theta,y}(\epsilon_i))}{\hat{\pi}(\mathcal{T}_{\theta,y}(\epsilon_i)|\theta, y)}.$$

When the estimation variance of importance sampling is $\mathcal{O}(n^{-1})$, this converges to the correct posterior density as $n \to \infty$.

A more efficient way is to sample correlated auxiliary variables to further reduce the estimation variance of importance sampling. Intuitively, we want the randomness to spread more uniformly than independent sampling. To achieve this, we can use a low-discrepancy sequence (LDS) in the space of the unit cube as the auxiliary variables. The approach, called quasi-Monte Carlo (QMC) [12], further reduces the estimation variance of importance sampling to $\mathcal{O}(n^{-2})$ under proper conditions. We demonstrate that QMC works better than raw importance sampling for correcting the error in Appendix D.1. In this work, we use the Sobol sequence, but other sequences, such as the Halton sequence, could also be applied. To use auxiliary variables in the unit cube, we change our reparameterization to,

$$z_i = \mathcal{T}_{\theta,y}(u_i) = \hat{z} + \sqrt{\Sigma}\Phi^{-1}(u_i), \text{ where } u_i \sim \text{uniform}([0,1]^{d_z}),$$

for $i = 1, \ldots, n$. Here Φ^{-1} is the inverse-CDF function of the unit normal distribution, so $z_i \sim \hat{\pi}(z|\theta, y)$, ensuring our estimator is a valid importance sampling estimator. With this parameterization, we fix an LDS $u_1, \ldots u_n$ and denote the QMC-based approximate marginal density by

$$\hat{\pi}^{\text{QMC}}_{u_{1:n}}(\theta, y) \coloneqq \frac{1}{n}\sum_{i=1}^{n} \frac{\pi(\theta)\pi(\mathcal{T}_{\theta,y}(u_i)|\theta)\pi(y|\theta, \mathcal{T}_{\theta,y}(u_i))}{\hat{\pi}(\mathcal{T}_{\theta,y}(u_i)|\theta, y)}. \tag{6}$$

We call this method QMC-ADLA. The approximate marginal distribution with the original ADLA can be viewed as QMC-ADLA where each $u_i = 1/2$, making $z_i = \hat{z}$.

$\hat{\pi}^{\text{QMC}}_{u_{1:n}}(\theta, y)$ defines an approximate posterior $\hat{\pi}^{\text{QMC}}_{u_{1:n}}(\theta|y)$. The next proposition shows that the total variation distance between $\hat{\pi}^{\text{QMC}}_{u_{1:n}}(\theta, y)$ and $\pi(\theta|y)$ vanishes as n increases.

Proposition 3. *Let* $f(u) = \frac{\pi(\theta)\pi(\mathcal{T}_{\theta,y}(u)|\theta)\pi(y|\theta,\mathcal{T}_{\theta,y}(u))}{\hat{\pi}(\mathcal{T}_{\theta,y}(u)|\theta,y)}$. *If* $f(u)$ *has bounded variation [34], and there exists a function* $g(\theta)$ *such that* $\hat{\pi}^{\text{QMC}}_{u_{1:n}}(\theta, y) < g(\theta)$ *and* $\int g(\theta)d\theta < \infty$, *then* $\int \left|\hat{\pi}^{\text{QMC}}_{u_{1:n}}(\theta|y) - \pi(\theta|y)\right| d\theta = 0$ *as* $n \to \infty$,

The proof is in Appendix A.2. Note that the condition of bounded variation is required for the convergence rate of QMC but is difficult to verify and may not hold in practice. Our framework allows $\hat{\pi}(z|\theta, y)$ to be replaced with any approximation of $\pi(z|\theta, y)$, including those with better tail behaviors. This flexibility may allow the conditions of QMC to be met more easily.

Naturally, for finite n, some posterior error still persists, and in our experiments, we find cases where the error is large even when $n = 64$. Also, similar to PM-ADLA, QMC-ADLA requires n density evaluations at each step of HMC.

5.1 ADLA with randomized QMC

It is possible to combine QMC and PM to construct an approximate marginal model that converges to the true marginal and does not increase the model dimension. The technique is randomized QMC (RQMC) [33], which introduces a single shift variable $U \sim \text{uniform}([0,1]^{d_z})$ as the only auxiliary variable. Next, let

$$u_i = (v_i + U)\%1$$

for $i = 1, \ldots, n$, where "%" is the modulo operator and $v_1, \ldots, v_n$ is a *fixed* LDS in the space of $[0,1]^{d_z}$. To simplify our notation, we can define the mapping from U to each z_i by

$$z_i = \mathcal{S}_i(U) \coloneqq \mathcal{T}_{\theta,y}((v_i + U)\%1).$$

Then, the RQMC-based approximate marginal model is

$$\hat{\pi}^{\text{RQMC}}(\theta, U, y) \coloneqq \pi(U)\left(\frac{1}{n}\sum_{i=1}^{n} \frac{\pi(\theta)\pi(\mathcal{S}_i(U)|\theta)\pi(y|\theta, \mathcal{S}_i(U))}{\hat{\pi}(\mathcal{S}_i(U)|\theta, y)}\right). \tag{7}$$

We call this estimator RQMC-ADLA, which is also correct for marginalization according to the following proposition.

Proposition 4. $\hat{\pi}^{\text{RQMC}}$, *as defined in eq.* (7), *admits* $\pi(\theta, y)$ *as its marginal.*

The proof is in Appendix A.3. Similar to PM-ADLA, as $n \to \infty$, the model converges to the true marginal multiplied by an independent uniform distribution over U. Since U has the same dimension as z, RQMC-ADLA does not increase the model dimension, and can be viewed as a tunable reparameterization with the hyperparameter n.

However, unlike previous approaches, the function $\mathcal{S}_i(U)$ is not continuous due to modulo operators, so the density $\hat{\pi}^{\mathrm{RQMC}}(\theta, U, y)$ is not continuous in U and HMC is not directly applicable. Fortunately, given a fixed U, $\hat{\pi}^{\mathrm{RQMC}}(\theta, U, y)$ is continuous in θ, so it is possible to use a Metropolis-within-Gibbs (MwG) sampler [19]. In each round, we update θ conditional on U with HMC, and then update U conditional on θ with Metropolis-Hastings. Details of the approach can be found in Appendix B.5.

We summarize the high-level trade-offs of all variants of ADLA in Table 1.

Now we analyze the time complexity. With HMC, in every leapfrog step, we need to compute the model density once. If we use ADLA, an additional $\mathcal{O}(d_z^3)$ complexity is introduced for the Laplace approximation. Our methods are more expensive than ADLA or HMC. To construct our approximate marginal posterior, the importance sampling estimator requires n copies of the latent variables $z_1, \ldots, z_n$, which implies computing the original model density n times. Furthermore, the reparameterization from the auxiliary variables $\epsilon_1, \ldots, \epsilon_n$ to $z_1, \ldots, z_n$ takes $\mathcal{O}(nd_z^2)$ time. These additional computations occur in every leapfrog step. In RQMC-ADLA, we also have a Gibbs sampling step in addition to the HMC step. In the Gibbs sampling step, the d_z-dimensional variable U is updated one coordinate at a time. To update each coordinate, we need $\mathcal{O}(nd_z^2)$ time for reparameterization to get $z_1, \ldots, z_n$ plus n density computations. Therefore, RQMC-ADLA requires $\mathcal{O}(nd_z^3)$ time complexity and nd_z density computations for Gibbs sampling. This implies that RQMC-ADLA scales more poorly with n or d_z compared with other methods. Although our methods are more expensive per step, this could be offset by a higher effective sample size per step, which eventually leads to more efficient sampling.

6 Recovering marginalized variables

In ADLA, the posterior distribution of z can be constructed using $\hat{\pi}(z|\theta, y)$. For example, within HMC, for each sample of θ, we generate $z \sim \hat{\pi}(z|\theta, y)$ such that $(\theta, z) \sim \hat{\pi}(\theta, z|y) \approx \pi(\theta, z|y)$. We can use the same procedure in our methods. However, sampling from the Laplace approximation introduces an error. With importance sampling, we can get posterior samples of z that are closer to the true posterior $\pi(z|\theta, y)$.

In any of our methods, for each θ, we have n unnormalized importance weights

$$w_i \coloneqq \frac{\pi(\theta)\pi(z_i|\theta)\pi(y|\theta, z_i)}{\hat{\pi}(z_i|\theta, y)},$$

for $i = 1, \ldots, n$, and also n samples $z_1, \ldots, z_n$. Then, we let

$$z \sim \hat{\pi}(z|z_{1:n}, \theta, y) := \frac{\sum_{i=1}^n w_i \delta_{z_i}(z)}{\sum_{i=1}^n w_i}.$$

The distribution of z then matches the true distribution $\pi(z|\theta, y)$ by the following proposition.

Proposition 5. *If for each i, $z_i \sim \hat{\pi}(\cdot|\theta, y)$, let $\hat{\pi}(\theta, z_{1:n}, y) = \left(\frac{1}{n}\sum_{i=1}^n w_i\right)\hat{\pi}(z_{1:n}|\theta, y)$, then*

$$\hat{\pi}(\theta, z, y) \coloneqq \int \hat{\pi}(z|z_{1:n}, \theta, y)\hat{\pi}(\theta, z_{1:n}, y)dz_{1:n} = \pi(\theta, z, y).$$

The proof is in Appendix A.4. In the proposition, $\hat{\pi}(\theta, z_{1:n}, y)$ is the importance sampling estimator multiplied by the density of $z_1, \ldots z_n$, equivalent to the approximate marginal model in Eq. 5 with a change of variable. Coupled with the recovery distribution $\hat{\pi}(z|z_{1:n}, \theta, y)$, we get a joint model $\hat{\pi}(\theta, z_{1:n}, z, y)$ whose marginal is the model $\pi(\theta, z, y)$. For PM-ADLA and RQMC-ADLA, since each z_i follows $\hat{\pi}(\cdot|\theta, y)$, we have $\hat{\pi}(z|\theta, y) = \pi(z|\theta, y)$. At each step of sampling, we can recover z directly from already computed intermediate results. For QMC-ADLA, we can follow the same procedure, but the recovery becomes approximate. This derivation is related to importance weighted variational inference [15, 16].

7 Related Work

The integrated Laplace approximation plays an important role in the influential software packages INLA [38], TMB [25], and GPstuff [44]. TMB's inverse subset algorithm [25] and ADLA [27, 28] propose implementations of the integrated Laplace approximation in probabilistic programming languages which use automatic differentiation, respectively TMB and Stan [13]. Within these frameworks, users have a great deal of flexibility when specifying their model, however the accuracy of the Laplace approximation is not guaranteed to hold for an arbitrary model. To truly take advantage of the flexibility afforded by probabilistic programming languages, users need broadly applicable guarantees such as the ones we obtain for our methods.

Several papers have proposed replacing the integrated Laplace approximation with a more accurate approximation, for example using a copula-based correction [18] or a higher-order Taylor expansion of $\log \pi(z|\theta, y)$ [40, 14]. These methods can readily be combined with our proposed estimators, since our estimators (eqs. 5, 6, and 7) do not in fact rely on $\pi(z|\theta, y)$ being approximated by a standard Laplace approximation. (On the other hand, designing differentiation algorithms for new approximations would require additional engineering effort.)

There is a rich literature on the use of importance sampling for approximate Bayesian inference. Berild et al. [4] combine importance sampling and the integrated Laplace approximation. Specifically, importance sampling is used as a backend inference algorithm (instead of MCMC) to approximate the marginal posterior $\pi(\theta|y)$; by contrast, we construct an importance sampling estimator of the marginal joint $\pi(\theta, y)$. The two methods can be combined and more generally, any (gradient-based) inference method can be used with our proposed estimators of $\pi(\theta, y)$. Importance sampling has also been used to improve posterior approximation of variational inference [47, 15]. Domke and Sheldon [16] generalize the idea and discuss QMC-based variational objectives. The recovery step of our methods is closely related to these works. There is another line of work that fixes the randomness of importance sampling-based variational objectives [10, 20]. Finally, Alenlöv et al. [2] explore using importance sampling within pseudo-marginalization for HMC. Different from all above works, the importance sampling in our work is applied to only part of the model, and uses the Laplace approximation as the proposal distribution.

Our work can also be related to recent studies on learning the best reparametererization of Bayesian models [21, 36, 22, 24]. RQMC-ADLA can be viewed as a training-free reparameterization with convergence guarantees.

8 Experiments

Our methods are implemented in BlackJAX [11], using the default No-U-turn sampler (NUTS) [23] of HMC in NumPyro [6, 37] as the inference algorithm. We demonstrate the performance of our methods with synthetic and real examples. All experiments are conducted on an Intel Xeon Platinum 8352Y CPU. In this section, we use 'Base' to represent HMC without marginalization, and call our methods 'PM' (Section 4), 'QMC' (Section 5) and 'RQMC' (Section 5.1). When computing the estimation error, we aggregate the results of all 'Base' runs as the ground-truth.

8.1 Gaussian process

A widely used family of LGMs is Gaussian process models. We synthesized a dataset of features $X \in \mathbb{R}^{N \times d}$ and responses $y \in \mathbb{N}^N$, where $N = 100$ and $d = 1$, and modeled these data with a log-Gaussian Cox process:

$$\rho \sim \Gamma^{-1}(3, 2),\ \alpha \sim \text{Exponential}(1),\ f \sim \text{MVN}(0, K(X, \rho, \alpha)),\ y \sim \text{Poisson}(e^f),$$

where $K_{ij}(X, \rho, \alpha) = \alpha^2 \exp\left(-\frac{1}{2\rho^2}(x_i - x_j)^2\right)$. Details can be found in the Appendix. The Laplace approximation can be applied to marginalize out the latent variables f.

This model is difficult to sample with HMC directly. We find that, even with a non-centered parameterization [35], there are still on average 205 (out of 100,000) divergent transitions, which indicate occasional numerical instability due to the geometry of the posterior [5]. Getting full HMC to work also requires increasing the target acceptance probability from a default value of

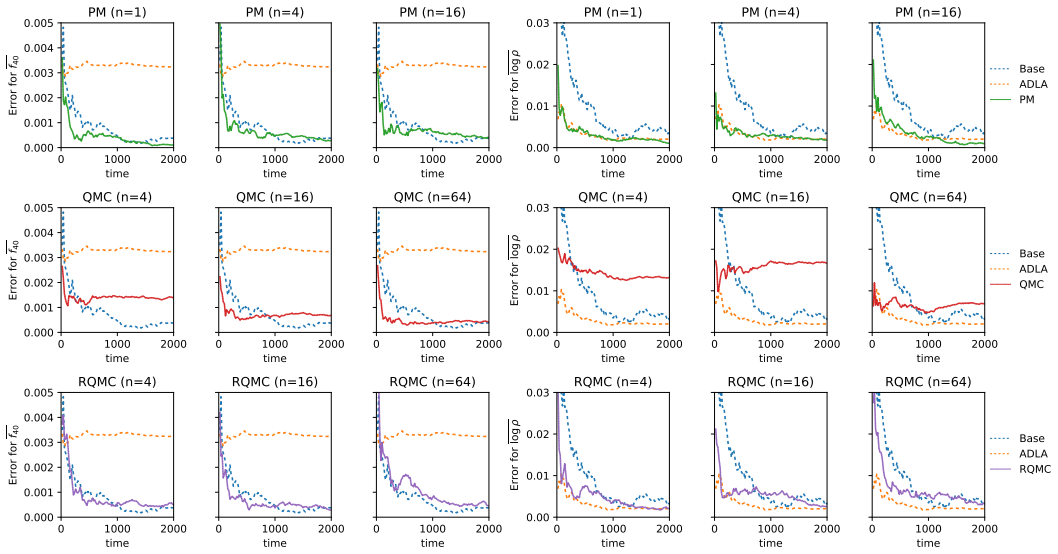


Figure 1: Error of estimating the means of parameters as a function of time in seconds for the synthesized Gaussian process with Poisson likelihood. Results are averaged from 5 independent runs. Ground-truth is estimated from NUTS on the unmarginalized model.

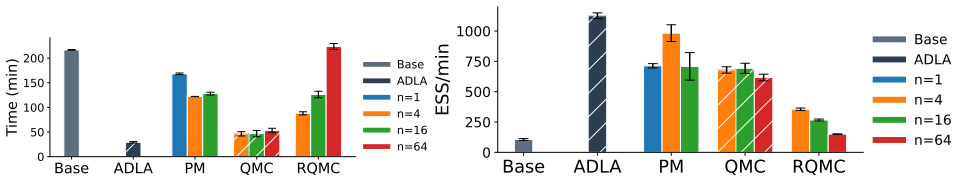


Figure 2: Running time to collect 100,000 samples and average ESS/min of $(\log \rho, \log \alpha)$ for the synthesized Gaussian process with Poisson likelihood. Hatched bars represent methods with error in the posterior. Results are collected from 5 independent runs.

0.8 to 0.99. In contrast, ADLA and our methods all report zero divergences with the default target acceptance probability without any reparameterization, meaning the *approximate* marginal posterior has a well-behaved geometry. We then use the samples to estimate the mean of the model parameters.

In Figure 1, we plot the estimation error as a function of time. When estimating $\mathbb{E}[\log \rho]$, ADLA has a lower error than HMC throughout the period. However, ADLA's estimation error for $\mathbb{E}[f_{40}]$ is evident from the beginning. In contrast, our proposed methods enjoy the speed-up of ADLA and reduced error as sampling proceeds. In the second row, we find that QMC gives better estimates for $\mathbb{E}[f_{40}]$ but worse estimates for $\mathbb{E}[\log \rho]$, and performs better with larger n. In the first and third rows, we see that both PM and RQMC have less error than ADLA for estimating $\mathbb{E}[f_{40}]$, and faster inference than HMC for estimating $\mathbb{E}[\log \rho]$. Despite using multiple samples, we find that QMC's bias may be higher than ADLA. When PM or RQMC is applicable, we find it is possible to correct the error of ADLA with comparable convergence speed.

In Figure 2, we further compare the running time and the effective sample size (ESS) per minute of different methods. An important caveat with this metric is that the ESS characterizes the variance of the Monte Carlo estimator *but not its bias*. A high ESS/min can be interpreted as HMC efficiently sampling from the approximate posterior. Only if this approximation is accurate, as verified in Figure 1, can we further interpret a high ESS/min as efficient sampling from the correct posterior.

ADLA and QMC run fastest but can suffer from a high bias for certain parameters. We find that, although PM is multiple times slower than ADLA, the fast mixing of the MCMC chains leads to a high ESS/min. Furthermore, PM produces a low error across parameters. RQMC is also slow because each Gibbs step requires multiple additional evaluations of the density function, which leads to the lowest ESS/min among the marginalization methods.

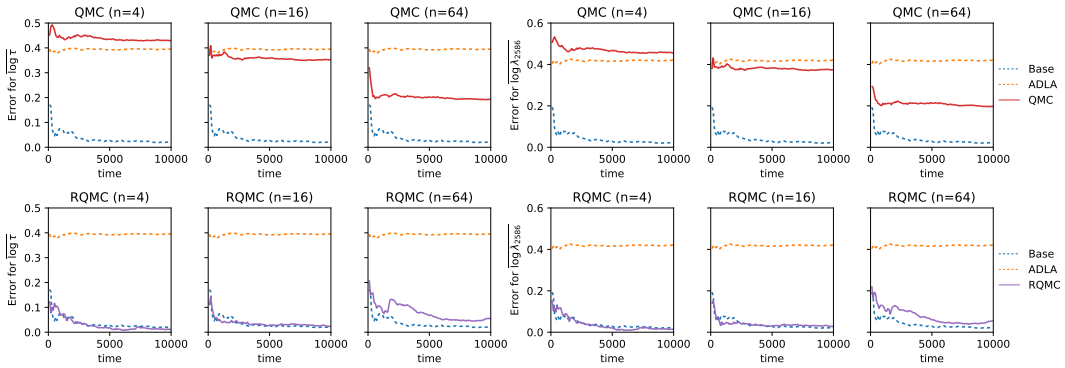


Figure 3: Error of estimating the means of parameters as a function of time in seconds for the sparse kernel interaction model. Results are averaged from 5 independent runs. Ground-truth is estimated from NUTS on the unmarginalized model.

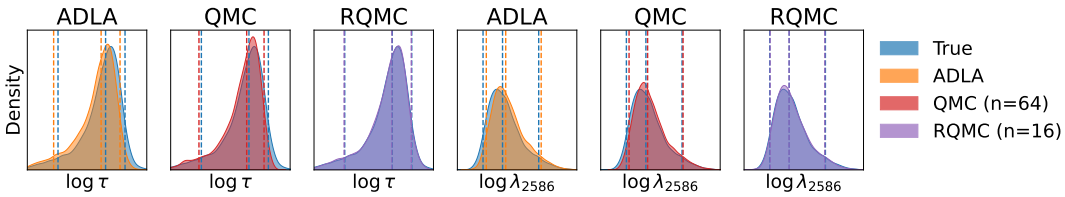


Figure 4: Comparing the sampled posterior against the ground-truth from NUTS. We demonstrate with two difficult variables: $\log \tau$ and $\log \lambda_{2586}$. Each posterior is estimated from 10,000 samples. The dashed lines are 5%, 50%, 95% quantiles.

8.2 Sparse kernel interaction model

Next, we demonstrate our methods on a model applied to real data. Sparse kernel interaction models (SKIMs) [1] are an extension of generalized linear models that include interaction terms. We use the same data and model as Margossian et al. [27]. Details can be found in the appendix. We monitor two difficult variables in the model, $\log \tau$ and $\log \lambda_{2586}$. In this model, we find that PM is too slow to produce results in a reasonable time. The estimation error of other methods can be found in Figure 3. With a non-centered parameterization and NUTS, the sampling is still challenging: there are 153 divergent transitions (out of 300,000). To get full HMC work, we still need to increase the target acceptance probability to 0.99. ADLA solves this problem, but there is also an evident error in estimating the mean of the variables. With QMC, the bias can be reduced, and as we increase n, the error becomes smaller. The error with RQMC is smaller than with ADLA and comparable to HMC. We demonstrate this further in Figure 4. Even with $n = 64$ samples, QMC still leads to error in the posterior as seen in the right tail of $\log \tau$ and left tail of $\log \lambda_{2586}$. RQMC effectively reduces this error with $n = 16$ samples.

8.3 Mixed-effects models

Another class of LGMs are mixed-effects models, which are generalized linear models with both fixed and random effects. We use the `Epil2` dataset from glmmTMB [9] to demonstrate the bias reduction ability of our methods. `Epil2` studies the seizure counts for patients in a clinical trial [7], which has the likelihood

$$y_i \sim \text{NegativeBinomial}(\exp(x_i^T \beta + z_i^T u_{g_i}), \phi)$$

where for each data i, y_i is the seizure count, x_i is the fixed-effect predictor in $\mathbb{R}^6$, z_i is the random-effect predictor in $\mathbb{R}^2$, g_i is the subject index. Parameters in the model include the fixed effect β, the random effects $u_1, u_2, ..., u_G$ for each subject, and the dispersion parameter ϕ. We marginalize out the random effects for inference. More details can be found in the appendix.

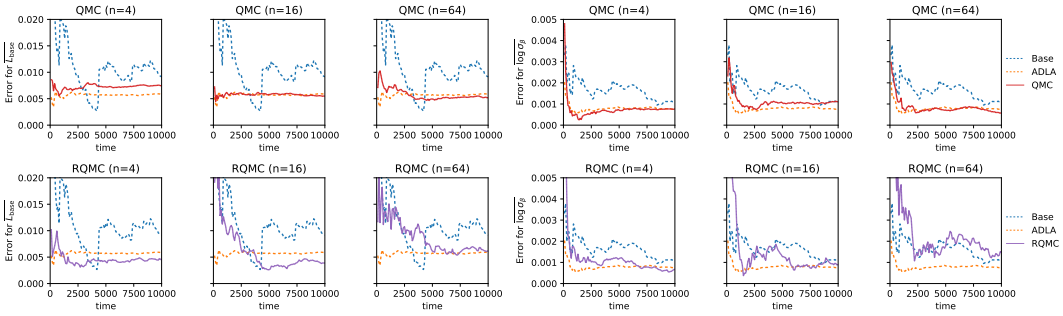


Figure 5: Error of estimating $\mathbb{E}[L_{\text{base}}]$ and $\mathbb{E}[\log \sigma_\beta]$ as a function of time in seconds for the `Epil2` model (L_{base} is transformed from L in the unconstrained space). Results are averaged from 5 independent runs. Ground-truth is estimated from NUTS with non-centered parameterization.

For this model, we find NUTS with non-centered parameterization of random effects very efficient for inference, so we use it to get estimated ground-truths. In Figure 5 we compare vanilla NUTS (without non-centered parameterization), ADLA, QMC, and RQMC by plotting the estimation error for means of two parameters as a function of time. In general, ADLA provides a fast estimation but has a non-vanishing estimation error. When estimating $\mathbb{E}[L_{\text{base}}]$, QMC with $n = 64$ and RQMC with $n = 4$ converge similarly as ADLA and have lower error at the end. When estimating $\mathbb{E}[\log \sigma_\beta]$, QMC is comparable with ADLA, but RQMC may give a higher error. This indicates that our methods may not always help inference and the hyperparameters should be carefully selected.

9 Discussion

We propose methods for reducing the error of the integrated Laplace approximation with an importance sampling estimator. The estimator is realized with various methods, including pseudo-marginalization, quasi-Monte Carlo and randomized quasi-Monte Carlo. Future work can explore different directions. First, a byproduct of importance sampling is its variance estimation, which may be used to diagnose how well the approximation is working. Second, in our experiments, HMC is used, but our methods can be combined with other (gradient-based) inference algorithms, such as variational inference (which has been combined with ADLA [29]). Third, while we focus on latent Gaussian models, our methods may extend to non-Gaussian latent variables, which would require different approximate marginalization methods and different correction mechanisms. Finally, we believe it would be useful to develop high-performance implementations of our methods for the various libraries which use an integrated Laplace approximation.

Limitations. We also note several limitations that may inspire future work. First, QMC-ADLA under-performs vanilla ADLA in several settings. We suspect it is because the QMC points are not symmetric around the mode, as opposed to the vanilla ADLA that is evaluated at the mode. Second, the benchmarking of different methods across models is limited. It will be interesting to test on more models with varying latent dimensions. Third, there is much space for exploration in the selection of methods and the choice of n for each method. We summarized the high-level trade-offs in Table 1 and presented experimental results for several choices of n. But an important part is analyzing importance sampling with correlated samples, such as those obtained from the reparameterization of QMC series. This is nontrivial and remains a difficult, though certainly important, research question. Fourth, we also conducted experiments on the first two models to obtain insights on the variance of the importance weights. In these experiments, we fitted a Pareto distribution to the importance weights for each posterior sample, and found that k-hat [45] (from the Pareto fit) was on average 0.5, albeit rarely above 0.7. This suggests that the variance of the importance weight may not be finite so the $\mathcal{O}(n^{-1})$ scaling in Proposition 1 applies conditionally. Still, our methods can be extended to use other (possibly more elaborate) proposal distributions that approximate $\pi(z|\theta, y)$ and may lead to better importance weights. Also, for PM-ADLA and RQMC-ADLA to have the correct posterior, the only requirement is the unbiasedness of importance sampling. In practice, it is possible to have difficult importance sampling but manageable MCMC sampling as shown in our experiments, but this may require further investigation.

Acknowledgment

The authors would like to thank the anonymous reviewers for their valuable feedback and insightful suggestions, which greatly improved the quality of this manuscript. CM acknowledges the support of the Natural Sciences and Engineering Research Council (NSERC) of Canada: RGPIN-2026-07219, DGECR-2026-00308.

References

[1] Raj Agrawal, Brian Trippe, Jonathan Huggins, and Tamara Broderick. The kernel interaction trick: Fast Bayesian discovery of pairwise interactions in high dimensions. In *International Conference on Machine Learning*, pages 141–150. PMLR, 2019.

[2] Johan Alenlöv, Arnoud Doucet, and Fredrik Lindsten. Pseudo-marginal Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 22(141):1–45, 2021.

[3] Christophe Andrieu and Gareth O. Roberts. The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics*, 37(2), April 2009. ISSN 0090-5364.

[4] Martin Outzen Berild, Sara Martino, Virgilio Gómez-Rubio, and Håvard Rue. Importance sampling with the integrated nested Laplace approximation. *Journal of Computational and Graphical Statistics*, 31(4):1225–1237, 2022. doi: 10.1080/10618600.2022.2067551.

[5] Michael Betancourt and Mark Girolami. Hamiltonian Monte Carlo for hierarchical models. In *Current Trends in Bayesian Methodology with Applications*, page 24. Chapman and Hall/CRC, 2015. doi: 10.1201/b18502-5.

[6] Eli Bingham, Jonathan P Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul Szerlip, Paul Horsfall, and Noah D Goodman. Pyro: Deep universal probabilistic programming. *Journal of Machine Learning Research*, 20(28):1–6, 2019.

[7] James G Booth, George Casella, Herwig Friedl, and James P Hobert. Negative binomial loglinear mixed models. *Statistical Modelling*, 3(3):179–191, 2003.

[8] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL `http://github.com/jax-ml/jax`.

[9] Mollie E. Brooks, Kasper Kristensen, Koen J. van Benthem, Arni Magnusson, Casper W. Berg, Anders Nielsen, Hans J. Skaug, Martin Maechler, and Benjamin M. Bolker. glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, 9(2):378–400, 2017. doi: 10.32614/RJ-2017-066.

[10] Javier Burroni, Justin Domke, and Daniel Sheldon. Sample average approximation for black-box variational inference. In *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, volume 244 of *Proceedings of Machine Learning Research*, pages 471–498. PMLR, 15–19 Jul 2024.

[11] Alberto Cabezas, Adrien Corenflos, Junpeng Lao, and Rémi Louf. Blackjax: Composable Bayesian inference in JAX, 2024.

[12] Russel E Caflisch. Monte Carlo and quasi-Monte Carlo methods. *Acta Numerica*, 7:1–49, 1998.

[13] Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76:1–32, 2017.

[14] Cristian Chiuchiolo, Janet van Niekerk, and Håvard Rue. An extended simplified Laplace strategy for approximate Bayesian inference of latent Gaussian models using R-INLA. *arXiv:2203.14304*, 2022.

[15] Justin Domke and Daniel R Sheldon. Importance weighting and variational inference. *Advances in Neural Information Processing Systems*, 31, 2018.

[16] Justin Domke and Daniel R Sheldon. Divide and couple: Using Monte Carlo variational objectives for posterior approximation. *Advances in Neural Information Processing Systems*, 32, 2019.

[17] Simon Duane, Anthony D Kennedy, Brian J Pendleton, and Duncan Roweth. Hybrid Monte Carlo. *Physics letters B*, 195(2):216–222, 1987.

[18] Egil Ferkingstad and Håvard Rue. Improving the INLA approach for approximate Bayesian inference for latent Gaussian models. *Electronic Journal of Statistics*, 9(2):2706–2731, 2015. doi: 10.1214/15-EJS1092.

[19] Wally R Gilks, Nicky G Best, and Keith KC Tan. Adaptive rejection Metropolis sampling within Gibbs sampling. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 44 (4):455–472, 1995.

[20] Ryan Giordano, Martin Ingram, and Tamara Broderick. Black box variational inference with a deterministic objective: Faster, more accurate, and even more black box. *Journal of Machine Learning Research*, 25(18):1–39, 2024.

[21] Maria Gorinova, Dave Moore, and Matthew Hoffman. Automatic reparameterisation of probabilistic programs. In *International Conference on Machine Learning*, pages 3648–3657. PMLR, 2020.

[22] Matthew 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 preprint arXiv:1903.03704*, 2019.

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

[24] Joohwan Ko and Justin Domke. Model-informed flows for Bayesian inference. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[25] Kasper Kristensen, Anders Nielsen, Casper W Berg, Hans Skaug, and Bradley M Bell. TMB: automatic differentiation and Laplace approximation. *Journal of Statistical Software*, 70:1–21, 2016.

[26] Jinlin Lai, Justin Domke, and Daniel R Sheldon. Hamiltonian Monte Carlo inference of marginalized linear mixed-effects models. *Advances in Neural Information Processing Systems*, 37:29435–29463, 2024.

[27] Charles Margossian, Aki Vehtari, Daniel Simpson, and Raj Agrawal. Hamiltonian Monte Carlo using an adjoint-differentiated Laplace approximation: Bayesian inference for latent Gaussian models and beyond. *Advances in Neural Information Processing Systems*, 33:9086–9097, 2020.

[28] Charles C Margossian. General adjoint-differentiated Laplace approximation. *arXiv preprint arXiv:2306.14976*, 2023.

[29] Charles C. Margossian and Lawrence K. Saul. Generalized guarantees for variational inference in the presence of even and elliptical symmetry. *arXiv:2511.01064*, 2025.

[30] Cole C Monnahan and Kasper Kristensen. No-U-turn sampling for fast Bayesian inference in ADMB and TMB: Introducing the adnuts and tmbstan R packages. *PloS one*, 13(5):e0197954, 2018.

[31] Radford M. Neal. MCMC using Hamiltonian dynamics. In *Handbook of Markov Chain Monte Carlo*. Chapman & Hall / CRC Press, 2012.

[32] John Ashworth Nelder and Robert WM Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 135(3):370–384, 1972.

[33] Art B Owen. Monte Carlo, quasi-Monte Carlo, and randomized quasi-Monte Carlo. In *Monte-Carlo and Quasi-Monte Carlo Methods 1998: Proceedings of a Conference held at the Claremont Graduate University, Claremont, California, USA, June 22–26, 1998*, pages 86–97. Springer, 2000.

[34] Art B Owen. Multidimensional variation for quasi-Monte Carlo. In *International Conference on Statistics in honour of Professor Kai-Tai Fang's 65th birthday*, pages 49–74. World Scientific, 2005.

[35] Omiros Papaspiliopoulos, Gareth O Roberts, and Martin Sköld. A general framework for the parametrization of hierarchical models. *Statistical Science*, pages 59–73, 2007.

[36] Matthew D Parno and Youssef M Marzouk. Transport map accelerated Markov chain Monte Carlo. *SIAM/ASA Journal on Uncertainty Quantification*, 6(2):645–682, 2018.

[37] Du Phan, Neeraj Pradhan, and Martin Jankowiak. Composable effects for flexible and accelerated probabilistic programming in NumPyro. *arXiv preprint arXiv:1912.11554*, 2019.

[38] Håvard Rue, Sara Martino, and Nicolas Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 71(2):319–392, 2009.

[39] Håvard Rue, Andrea Riebler, Sigrunn Sorbye, Janine Illian, Daniel Simson, and Finn Lindgren. Bayesian computing with INLA: A review. *Annual Review of Statistics and its Application*, 4: 395 – 421, 2017. doi: https://doi.org/10.1146/annurev-statistics-060116-054045.

[40] Z. Shun and P. McCullagh. Laplace approximation of high dimensional integrals. *Journal of the Royal Statistical Society: Series B*, 57(4):749–760, 1995.

[41] Luke Tierney and Joseph B Kadane. Accurate approximations for posterior moments and marginal densities. *Journal of the American statistical Association*, 81(393):82–86, 1986.

[42] Jarno Vanhatalo, Pasi Jylänki, and Aki Vehtari. Gaussian process regression with a student-t likelihood. *Advances in Neural Information Processing Systems*, 22:1910–1918, 2009.

[43] Jarno Vanhatalo, Ville Pietiläinen, and Aki Vehtari. Approximate inference for disease mapping with sparse Gaussian processes. *Statistics in medicine*, 29(15):1580–1607, 2010.

[44] Jarno Vanhatalo, Jaakko Riihimäki, Jouni Hartikainen, Pasi Jylänki, Ville Tolvanen, and Aki Vehtari. GPstuff: Bayesian modeling with Gaussian processes. *The Journal of Machine Learning Research*, 14(1):1175–1179, 2013.

[45] Aki Vehtari, Daniel Simpson, Andrew Gelman, Yuling Yao, and Jonah Gabry. Pareto smoothed importance sampling. *Journal of Machine Learning Research*, 25(72):1–58, 2024.

[46] Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

[47] Yuling Yao, Aki Vehtari, Daniel Simpson, and Andrew Gelman. Yes, but did it work?: Evaluating variational inference. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5581–5590. PMLR, 2018.

Appendices to "Corrected Integrated Laplace Approximation"

A Proof of the theories

In this section, we restate and prove each of the propositions.

A.1 Proof of Proposition 2

Proposition 2. *$\hat{\pi}^{\mathrm{PM}}$, as defined in eq.* (5)*, admits $\pi(\theta, y)$ as its marginal.*

Proof. We need to show that $\int \hat{\pi}^{\mathrm{PM}}(\theta, \epsilon_{1:n}, y) d\epsilon_{1:n} = \pi(\theta, y)$.

$$\begin{aligned}
\int \hat{\pi}^{\mathrm{PM}}(\theta, \epsilon_{1:n}, y) du_{1:n} &= \int \prod_{i=1}^{n} \pi(\epsilon_i) \left(\frac{1}{n} \sum_{i=1}^{n} \frac{\pi(\theta)\pi(\mathcal{T}_{\theta,y}(\epsilon_i)|\theta)\pi(y|\theta, \mathcal{T}_{\theta,y}(\epsilon_i))}{\hat{\pi}(\mathcal{T}_{\theta,y}(\epsilon_i)|\theta, y)} \right) d\epsilon_{1:n} \\
&= \frac{1}{n} \sum_{i=1}^{n} \int \prod_{j=1}^{n} \pi(\epsilon_j) \frac{\pi(\theta)\pi(\mathcal{T}_{\theta,y}(\epsilon_i)|\theta)\pi(y|\theta, \mathcal{T}_{\theta,y}(\epsilon_i))}{\hat{\pi}(\mathcal{T}_{\theta,y}(\epsilon_i)|\theta, y)} d\epsilon_{1:n} \\
&= \frac{1}{n} \sum_{i=1}^{n} \int \frac{\pi(\theta)\pi(\mathcal{T}_{\theta,y}(\epsilon_i)|\theta)\pi(y|\theta, \mathcal{T}_{\theta,y}(\epsilon_i))}{\hat{\pi}(\mathcal{T}_{\theta,y}(\epsilon_i)|\theta, y)} \pi(\epsilon_i) d\epsilon_i \\
&= \frac{1}{n} \sum_{i=1}^{n} \int \frac{\pi(\theta)\pi(z_i|\theta)\pi(y|\theta, z_i)}{\hat{\pi}(z_i|\theta, y)} \hat{\pi}(z_i|\theta, y) dz_i \\
&= \frac{1}{n} \sum_{i=1}^{n} \pi(\theta, y) \\
&= \pi(\theta, y). \qquad (8)
\end{aligned}$$

□

A.2 Proof of Proposition 3

Proposition 3. *Let $f(u) = \frac{\pi(\theta)\pi(\mathcal{T}_{\theta,y}(u)|\theta)\pi(y|\theta,\mathcal{T}_{\theta,y}(u))}{\hat{\pi}(\mathcal{T}_{\theta,y}(u)|\theta,y)}$. If $f(u)$ has bounded variation [34], and there exists a function $g(\theta)$ such that $\hat{\pi}^{\mathrm{QMC}}_{u_{1:n}}(\theta, y) < g(\theta)$ and $\int g(\theta) d\theta < \infty$, then $\int \left| \hat{\pi}^{\mathrm{QMC}}_{u_{1:n}}(\theta|y) - \pi(\theta|y) \right| d\theta = 0$ as $n \to \infty$.*

Proof. By the bounded variation of f, we invoke the properties of QMC, which suggests that $\hat{\pi}^{\mathrm{QMC}}_{u_{1:n}}(\theta, y) \to \pi(\theta, y)$ pointwise as $n \to \infty$. With dominated convergence theorem,

$$\lim_{n\to\infty} \hat{\pi}^{\mathrm{QMC}}_{u_{1:n}}(y) = \lim_{n\to\infty} \int \hat{\pi}^{\mathrm{QMC}}_{u_{1:n}}(\theta, y) d\theta = \int \lim_{n\to\infty} \hat{\pi}^{\mathrm{QMC}}_{u_{1:n}}(\theta, y) d\theta = \int \pi(\theta, y) d\theta = \pi(y). \quad (9)$$

Therefore,

$$\lim_{n\to\infty} \hat{\pi}^{\mathrm{QMC}}_{u_{1:n}}(\theta|y) = \lim_{n\to\infty} \frac{\hat{\pi}^{\mathrm{QMC}}_{u_{1:n}}(\theta, y)}{\hat{\pi}^{\mathrm{QMC}}_{u_{1:n}}(y)} = \frac{\lim_{n\to\infty} \hat{\pi}^{\mathrm{QMC}}_{u_{1:n}}(\theta, y)}{\lim_{n\to\infty} \hat{\pi}^{\mathrm{QMC}}_{u_{1:n}}(y)} = \frac{\pi(\theta, y)}{\pi(y)} = \pi(\theta|y). \quad (10)$$

Also, $\int \hat{\pi}^{\mathrm{QMC}}_{u_{1:n}}(\theta|y) d\theta = 1 = \int \pi(\theta|y) d\theta$, so by Scheffé's lemma, $\lim_{n\to\infty} \int \left| \hat{\pi}^{\mathrm{QMC}}_{u_{1:n}}(\theta|y) - \pi(\theta|y) \right| d\theta = 0$. □

A.3 Proof of Proposition 4

Proposition 4. *$\hat{\pi}^{\mathrm{RQMC}}$, as defined in eq.* (7)*, admits $\pi(\theta, y)$ as its marginal.*

Proof. We need to show that $\int \hat{\pi}^{\mathrm{RQMC}}(\theta, U, y)dU = \pi(\theta, y)$. For each i, we can apply the change of variable $z_i = \mathcal{S}_i(U)$.

$$\begin{aligned}
\int \hat{\pi}^{\mathrm{RQMC}}(\theta, U, y)dU &= \int \pi(U) \left(\frac{1}{n} \sum_{i=1}^{n} \frac{\pi(\theta)\pi(\mathcal{S}_i(U)|\theta)\pi(y|\theta, \mathcal{S}_i(U))}{\hat{\pi}(\mathcal{S}_i(U)|\theta, y)} \right) dU \\
&= \frac{1}{n} \sum_{i=1}^{n} \int \pi(U) \frac{\pi(\theta)\pi(\mathcal{S}_i(U)|\theta)\pi(y|\theta, \mathcal{S}_i(U))}{\hat{\pi}(\mathcal{S}_i(U)|\theta, y)} dU \\
&= \frac{1}{n} \sum_{i=1}^{n} \int \frac{\pi(\theta)\pi(z_i|\theta)\pi(y|\theta, z_i)}{\hat{\pi}(z_i|\theta, y)} \hat{\pi}(z_i|\theta, y)dz_i \\
&= \int \pi(\theta)\pi(z|\theta)\pi(y|\theta, z)dz \\
&= \pi(\theta, y). \qquad (11)
\end{aligned}$$

□

A.4 Proof of Proposition 5

Proposition 5. *If for each i, $z_i \sim \hat{\pi}(z|\theta, y)$, let $\hat{\pi}(\theta, z_{1:n}, y) = \left(\frac{1}{n}\sum_{i=1}^{n} w_i\right) \hat{\pi}(z_{1:n}|\theta, y)$, then*

$$\hat{\pi}(\theta, z, y) \coloneqq \int \hat{\pi}(z|z_{1:n}, \theta, y)\hat{\pi}(\theta, z_{1:n}, y)dz_{1:n} = \pi(\theta, z, y).$$

Proof.

$$\begin{aligned}
\hat{\pi}(\theta, z, y) &= \int \hat{\pi}(z|z_{1:n}, \theta, y)\hat{\pi}(\theta, z_{1:n}, y)dz_{1:n} \\
&= \int \frac{\sum_{i=1}^{n} w_i \delta_{z_i}(z)}{\sum_{i=1}^{n} w_i} \left(\frac{1}{n} \sum_{i=1}^{n} w_i \right) \hat{\pi}(z_{1:n}|\theta, y)dz_{1:n} \\
&= \frac{1}{n} \sum_{i=1}^{n} \int w_i \delta_{z_i}(z) \hat{\pi}(z_{1:n}|\theta, y)dz_{1:n} \\
&= \frac{1}{n} \sum_{i=1}^{n} \int w_i \hat{\pi}(z_{1:n}|\theta, y)|_{z_i=z} dz_{-i} \\
&= \frac{1}{n} \sum_{i=1}^{n} w_i \hat{\pi}(z_i|\theta, y)|_{z_i=z} \\
&= \frac{1}{n} \sum_{i=1}^{n} \pi(\theta, z, y) \\
&= \pi(\theta, z, y). \qquad (12)
\end{aligned}$$

□

B Details of ADLA and our methods

B.1 Newton's method

We use the procedure in Algorithm 1 to get $\hat{z}$ in Laplace approximation, following Margossian [28].

Note at any z,

$$\nabla_z(\log \pi(y|\theta, z) + \log \pi(z|\theta)) = \nabla_z \log \pi(y|\theta, z) - K^{-1}z.$$

So at $\hat{z}$, we have $\nabla_z \log \pi(y|\theta, \hat{z}) - K^{-1}\hat{z} = 0$, which means $\hat{z} = K\nabla_z \log \pi(y|\theta, \hat{z})$. This property will be used in deriving the gradients. In the Newton's method, we use $\|\nabla_z \log \pi(y|\theta, z) - a\|_\infty \leq 10^{-4}$ as the stopping rule of the optimization.

Algorithm 1 Newton's method

Require: Init position z_0

$K \leftarrow K(\theta)$
$z \leftarrow z_0$
while $\|\nabla_z \log \pi(y|\theta, z) - a\|_\infty \leq 10^{-4}$ **do**
  $c \leftarrow \nabla_z \log \pi(y|\theta, z)$
  $W \leftarrow -\nabla_z \nabla_z \log \pi(y|\theta, z)$
  $L \leftarrow \text{Cholesky}(I + W^{\frac{1}{2}} K W^{\frac{1}{2}})$
  $b \leftarrow Wz + c$
  $a \leftarrow b - W^{\frac{1}{2}} L^{-T} L^{-1} W^{\frac{1}{2}} Kb$
  $z \leftarrow Ka$
end while
$\hat{z} \leftarrow z$

B.2 Derivation of automatic differentiation over ξ

Our derivation of automatic differentiation follows the first case in Margossian [28] ($B = I + W^{1/2} K W^{1/2}$) that handles $\hat{z}$, and then generalizes it for z_i. In this section, we derive the gradients for ξ, the hyperparameters for z, contained in θ. Given θ, the target density of Laplace approximation is

$$f(z) = \log \pi(z|\theta) + \log \pi(y|\theta, z).$$

Define $W \coloneqq -\nabla_z \nabla_z \log \pi(y|\theta, \hat{z})$, then at the optimum $\hat{z}$, the Hessian of $f(z)$ is

$$\nabla_z \nabla_z f(\hat{z}) = -K^{-1} - W.$$

Define $A = (K^{-1} + W)^{-1}$, then the Laplace approximation is $\hat{\pi}(z|\theta, y) = \text{normal}(\hat{z}, A)$.

The ADLA log likelihood is

$$\begin{aligned}
\log \hat{\pi}_{\hat{z}}(y|\theta) &= \log \pi(\hat{z}|\theta) + \log \pi(y|\theta, \hat{z}) - \log \hat{\pi}(\hat{z}|\theta, y) \\
&= -\frac{1}{2}\hat{z}^T K \hat{z} - \frac{1}{2}\log|K| + \log \pi(y|\theta, \hat{z}) + \frac{1}{2}\log|A| \\
&= -\frac{1}{2}\hat{z}^T K \hat{z} + \log \pi(y|\theta, \hat{z}) - \frac{1}{2}\log|K||K^{-1} + W| \\
&= -\frac{1}{2}\hat{z}^T K \hat{z} + \log \pi(y|\theta, \hat{z}) - \frac{1}{2}\log|I + KW|. \qquad (13)
\end{aligned}$$

We assume W to be diagonal and denote $B = I + W^{1/2} K W^{1/2}$. Then we have

$$\begin{aligned}
\log \hat{\pi}_{\hat{z}}(y|\theta) &= -\frac{1}{2}\hat{z}^T K \hat{z} + \log \pi(y|\theta, \hat{z}) - \frac{1}{2}\log|I + W^{1/2} K W^{1/2}| \\
&= -\frac{1}{2}\hat{z}^T K \hat{z} + \log \pi(y|\theta, \hat{z}) - \frac{1}{2}\log|B|.
\end{aligned}$$

Also, A can be computed from the Woodbury formula.

$$\begin{aligned}
A &= (K^{-1} + W)^{-1} \\
&= K - KW^{1/2}(I + W^{1/2} K W^{1/2})^{-1} W^{1/2} K \\
&= K - KW^{1/2} B^{-1} W^{1/2} K.
\end{aligned}$$

By processing the Cholesky decomposition $B = LL^T$, and compute $C = L^{-1} W^{1/2} K$, we can get $A = K - C^T C$.

Next, we derive the gradients. We consider the gradient for a single element ξ_i. $\nabla_{\xi_i} \log \hat{\pi}_{\hat{z}}(y|\theta)$ has two parts. First, the explicit gradient (from Eq. (13))

$$\begin{aligned}
\left.\frac{\partial \log \hat{\pi}_{\hat{z}}(y|\theta)}{\partial \xi_i}\right|_{\text{explicit}} &= \frac{1}{2}\hat{z}^T K^{-1} \frac{\partial K}{\partial \xi_i} K^{-1} \hat{z} - \frac{1}{2}\text{trace}\left((W^{-1} + K)^{-1} \frac{\partial K}{\partial \xi_i}\right) \\
&= \frac{1}{2}\hat{z}^T K^{-1} \frac{\partial K}{\partial \xi_i} K^{-1} \hat{z} - \frac{1}{2}\text{trace}\left(R \frac{\partial K}{\partial \xi_i}\right)
\end{aligned}$$

where we define $R = (W^{-1} + K)^{-1} = W^{1/2}B^{-1}W^{1/2}$. The implicit gradient through $\hat{z}$ is

$$\left.\frac{\partial \log \hat{\pi}_{\hat{z}}(y|\theta)}{\partial \xi_i}\right|_{\text{implicit}} = \frac{\partial \log \hat{\pi}_{\hat{z}}(y|\theta)}{\partial \hat{z}} \cdot \frac{\partial \hat{z}}{\partial \xi_i}.$$

Since $\hat{z}$ is the optimum, in the first term,

$$\begin{aligned}\frac{\partial \log \hat{\pi}_{\hat{z}}(y|\theta)}{\partial \hat{z}_j} &= -\frac{1}{2}\frac{\partial \log |I + KW|}{\partial \hat{z}_j} \\ &= -\frac{1}{2}\text{trace}\left((K^{-1} + W)^{-1}\frac{\partial W}{\partial \hat{z}_j}\right).\end{aligned}$$

Therefore,

$$\begin{aligned}\frac{\partial \log \hat{\pi}_{\hat{z}}(y|\theta)}{\partial \hat{z}} &= \frac{1}{2}(\text{diag}((K^{-1} + W)^{-1})\nabla_z^3 \log \pi(y|\theta, \hat{z})|)^T \\ &= \frac{1}{2}(\text{diag}(A)\nabla_z^3 \log \pi(y|\theta, \hat{z})|)^T\end{aligned}$$

For the second implicit term, by differentiating $\hat{z} = K\nabla_z \log \pi(y|\theta, \hat{z})$, we get

$$\begin{aligned}\frac{\partial \hat{z}}{\partial \xi_i} &= \frac{\partial K}{\partial \xi_i}\nabla_z \log \pi(y|\theta, \hat{z}) + K\nabla_z^2 \log \pi(y|\theta, \hat{z})\frac{\partial \hat{z}}{\partial \xi_i} \\ &= \frac{\partial K}{\partial \xi_i}\nabla_z \log \pi(y|\theta, \hat{z}) - KW\frac{\partial \hat{z}}{\partial \xi_i}.\end{aligned}$$

Therefore, let $c = \nabla_z \log \pi(y|\theta, \hat{z})$, we have

$$\begin{aligned}\frac{\partial \hat{z}}{\partial \xi_i} &= (I + KW)^{-1}\frac{\partial K}{\partial \xi_i}c \\ &= W^{-1}(W^{-1} + K)^{-1}\frac{\partial K}{\partial \xi_i}c \\ &= W^{-1}R\frac{\partial K}{\partial \xi_i}c \\ &= (KR + W^{-1}R - KR)\frac{\partial K}{\partial \xi_i}c \\ &= (I - KR)\frac{\partial K}{\partial \xi_i}c.\end{aligned}$$

In summary,

$$\begin{aligned}\nabla_{\xi_i} \log \hat{\pi}_{\hat{z}}(y|\theta) =& \frac{1}{2}\hat{z}^T K^{-1}\frac{\partial K}{\partial \xi_i}K^{-1}\hat{z} - \frac{1}{2}\text{trace}\left(R\frac{\partial K}{\partial \xi_i}\right) \\ &+ \frac{1}{2}(\text{diag}(A)\nabla_z^3 \log \pi(y|\theta, \hat{z})|)^T(I - KR)\frac{\partial K}{\partial \xi_i}c.\end{aligned}$$

Observe that $a = K^{-1}\hat{z}$ and define $s = \frac{1}{2}\text{diag}(A)\nabla_z^3 \log \pi(y|\theta, \hat{z})$, we further have

$$\begin{aligned}\nabla_{\xi_i} \log \hat{\pi}_{\hat{z}}(y|\theta) &= \frac{1}{2}a^T\frac{\partial K}{\partial \xi_i}a - \frac{1}{2}\text{trace}\left(R\frac{\partial K}{\partial \xi_i}\right) + s(I - KR)\frac{\partial K}{\partial \xi_i}c \\ &= \frac{1}{2}\left\langle a, \frac{\partial K}{\partial \xi_i}a\right\rangle_F - \frac{1}{2}\left\langle R, \frac{\partial K}{\partial \xi_i}\right\rangle_F + \left\langle s - RKs, \frac{\partial K}{\partial \xi_i}c\right\rangle_F \\ &= \frac{1}{2}\left\langle aa^T, \frac{\partial K}{\partial \xi_i}\right\rangle_F - \frac{1}{2}\left\langle R, \frac{\partial K}{\partial \xi_i}\right\rangle_F + \left\langle (s - RKs)c^T, \frac{\partial K}{\partial \xi_i}\right\rangle_F \\ &= \left\langle \frac{1}{2}aa^T - \frac{1}{2}R + (s - RKs)c^T, \frac{\partial K}{\partial \xi_i}\right\rangle_F.\end{aligned}$$

To get the last term, we can compute the vector Jacobian product (VJP) of K, with the initial tangent of $\Omega = \frac{1}{2}aa^T - \frac{1}{2}R + (s - RKs)c^T$.

In this work, we also need to compute $\nabla_\xi z_i$, which reduces to computing $\nabla_\xi \hat{z}$ and $\nabla_\xi A$. For $\hat{z}$,

$$\begin{aligned}\left\langle v, \frac{\partial \hat{z}}{\partial \xi_i} \right\rangle_F &= \left\langle v, (I - KR)\frac{\partial K}{\partial \xi_i} c \right\rangle_F \\ &= \left\langle (I - RK)vc^T, \frac{\partial K}{\partial \xi_i} \right\rangle_F .\end{aligned}$$

For A which is the same as Σ,

$$\begin{aligned}\frac{\partial A}{\partial \xi_i} &= \frac{\partial (K^{-1} + W)^{-1}}{\partial \xi_i} \\ &= -(K^{-1} + W)^{-1}\frac{\partial K^{-1}}{\partial \xi_i}(K^{-1} + W)^{-1} \\ &= (K^{-1} + W)^{-1}K^{-1}\frac{\partial K}{\partial \xi_i}K^{-1}(K^{-1} + W)^{-1} \\ &= (I + KW)^{-1}\frac{\partial K}{\partial \xi_i}(I + WK)^{-1} \\ &= (I - KR)\frac{\partial K}{\partial \xi_i}(I - RK).\end{aligned}$$

Therefore, given an arbitrary matrix V,

$$\begin{aligned}\left\langle V, \frac{\partial A}{\partial \xi_i} \right\rangle_F &= \left\langle V, (I - KR)\frac{\partial K}{\partial \xi_i}(I - RK) \right\rangle_F \\ &= \left\langle (I - RK)V(I - KR), \frac{\partial K}{\partial \xi_i} \right\rangle_F .\end{aligned}$$

Define $D = I - RK$, then

$$\begin{aligned}\left\langle v, \frac{\partial \hat{z}}{\partial \xi_i} \right\rangle_F &= \left\langle Dvc^T, \frac{\partial K}{\partial \xi_i} \right\rangle_F , \\ \left\langle V, \frac{\partial A}{\partial \xi_i} \right\rangle_F &= \left\langle DVD^T, \frac{\partial K}{\partial \xi_i} \right\rangle_F .\end{aligned}$$

So the gradient computation for z_i also reduces to the VJP of K.

B.3 Derivation of automatic differentiation over η

Next we derive the gradients for η, the hyperparameters for y, contained in θ. Recall the ADLA log likelihood is

$$\log \hat{\pi}_{\hat{z}}(y|\theta) = -\frac{1}{2}\hat{z}^T K \hat{z} + \log \pi(y|\theta, \hat{z}) - \frac{1}{2}\log |I + KW|. \tag{14}$$

We consider the gradient for a single element η_i. $\nabla_{\eta_i} \log \hat{\pi}_{\hat{z}}(y|\theta)$ has two parts. First, the explicit gradient is

$$\begin{aligned}\left.\frac{\partial \log \hat{\pi}_{\hat{z}}(y|\theta)}{\partial \eta_i}\right|_{\text{explicit}} &= \nabla_{\eta_i} \log \pi(y|\theta, \hat{z}) - \frac{1}{2}\text{trace}\left((K^{-1} + W)^{-1}\frac{\partial W}{\partial \eta_i}\right) \\ &= \nabla_{\eta_i} \log \pi(y|\theta, \hat{z}) - \frac{1}{2}\text{trace}\left(A\frac{\partial W}{\partial \eta_i}\right)\end{aligned}$$

For the implicit gradient, we need $\frac{\partial \hat{z}}{\partial \eta_i}$, which is obtained from differentiating $\hat{z} = K\nabla_z \log \pi(y|\theta, \hat{z})$:

$$\begin{aligned}\frac{\partial \hat{z}}{\partial \eta_i} &= K\frac{\partial}{\partial \eta_i}\nabla_z \log \pi(y|\theta, \hat{z}) + K\nabla_z^2 \log \pi(y|\theta, \hat{z})\frac{\partial \hat{z}}{\partial \eta_i} \\ &= K\frac{\partial}{\partial \eta_i}\nabla_z \log \pi(y|\theta, \hat{z}) - KW\frac{\partial \hat{z}}{\partial \eta_i}.\end{aligned}$$

Therefore, recall $c = \nabla_z \log \pi(y|\theta, \hat{z})$, we have

$$\begin{aligned}\frac{\partial \hat{z}}{\partial \eta_i} &= (I + KW)^{-1} K \frac{\partial c}{\partial \eta_i} \\ &= (K^{-1} + W)^{-1} \frac{\partial c}{\partial \eta_i} \\ &= A \frac{\partial c}{\partial \eta_i}.\end{aligned}$$

Combined with the fact that $\frac{\partial \log \hat{\pi}_z(y|\theta)}{\partial \hat{z}} = \frac{1}{2}(\mathrm{diag}(A)\nabla_z^3 \log \pi(y|\theta, \hat{z})|)^T$ and the explicit gradient, we have

$$\begin{aligned}\frac{\partial \log \hat{\pi}_{\hat{z}}(y|\theta)}{\partial \eta_i} &= \nabla_{\eta_i} \log \pi(y|\theta, \hat{z}) - \frac{1}{2}\mathrm{trace}\left(A \frac{\partial W}{\partial \eta_i}\right) + \frac{1}{2}(\mathrm{diag}(A)\nabla_z^3 \log \pi(y|\theta, \hat{z})|)^T A \frac{\partial c}{\partial \eta_i} \\ &= \nabla_{\eta_i} \log \pi(y|\theta, \hat{z}) - \frac{1}{2}\left\langle A, \frac{\partial W}{\partial \eta_i} \right\rangle_F + \left\langle As, \frac{\partial c}{\partial \eta_i} \right\rangle_F.\end{aligned}$$

So the gradient computation for η also reduces to VJPs of W and c.

In this work, we need to compute $\frac{\partial \hat{z}}{\partial \eta_i}$ and $\frac{\partial A}{\partial \eta_i}$. For $\hat{z}$, consider an arbitrary vector v, then

$$\begin{aligned}\left\langle v, \frac{\partial \hat{z}}{\partial \eta_i} \right\rangle_F &= \left\langle v, A \frac{\partial c}{\partial \eta_i} \right\rangle_F \\ &= \left\langle Av, \frac{\partial c}{\partial \eta_i} \right\rangle_F.\end{aligned}$$

For $A = \Sigma$, then

$$\begin{aligned}\frac{\partial A}{\partial \eta_i} &= \frac{\partial (K^{-1} + W)^{-1}}{\partial \eta_i} \\ &= -(K^{-1} + W)^{-1} \frac{\partial W}{\partial \eta_i} (K^{-1} + W)^{-1} \\ &= -A \frac{\partial W}{\partial \eta_i} A.\end{aligned}$$

So with a tangent V,

$$\begin{aligned}\left\langle V, \frac{\partial A}{\partial \eta_i} \right\rangle_F &= \left\langle V, -A \frac{\partial W}{\partial \eta_i} A \right\rangle_F \\ &= \left\langle -AVA, \frac{\partial W}{\partial \eta_i} \right\rangle_F.\end{aligned}$$

Therefore, the VJP of $\hat{z}$ reduces to the VJP of c, and the VJP of A reduces to the VJP of W.

B.4 Implementation within JAX

As shown above, the gradient computation reduces to Frobenius inner products between a vector/matrix and a gradient. In JAX, this can be implemented via reverse-mode automatic differentiation with vector Jacobian products. We summarize the gradient computation for ADLA in Algorithm 2, and our generalization in Algorithm 3.

B.5 Details of the inference algorithms

In RQMC-ADLA, we have the following model.

$$\hat{\pi}^{\mathrm{RQMC}}(\theta, U, y) = \pi(U) \left(\frac{1}{n} \sum_{i=1}^{n} \frac{\pi(\theta)\pi(\mathcal{S}_i(U)|\theta)\pi(y|\theta, \mathcal{S}_i(U))}{\hat{\pi}(\mathcal{S}_i(U)|\theta, y)} \right).$$

The model is not continuous with respect to U, so it is not possible to use HMC for inference. Instead, we use the Metropolis-within-Gibbs (MwG) sampler. Note the dimension of U is d_z, the sampling pipeline is as follows. Here HMCkernel represents the HMC kernel to update θ. The hyperparameters of this kernel are tuned during the warm-up phase with a fixed U. The MHkernel represents the Metropolis-Hastings step for updating U. We choose $\mathrm{Uniform}(-0.1, 0.1)$ as the proposal distribution, but it can be replaced by any other distribution.

Algorithm 2 Adjoint-differentiated Laplace approximation [28]

Require: $\hat{z}$ from the optimizer

$K \leftarrow K(\theta)$
$c \leftarrow \nabla_z \log \pi(y|\theta, \hat{z})$
$W \leftarrow -\nabla_z \nabla_z \log \pi(y|\theta, \hat{z})$
$L \leftarrow \text{Cholesky}(I + W^{\frac{1}{2}} K W^{\frac{1}{2}})$
$b \leftarrow W\hat{z} + c$
$a \leftarrow b - W^{\frac{1}{2}} L^{-T} L^{-1} W^{\frac{1}{2}} K b$
} Restored from Newton's method
$R \leftarrow W^{\frac{1}{2}} L^{-T} L^{-1} W^{\frac{1}{2}}$
$A \leftarrow K - K W^{\frac{1}{2}} L^{-T} L^{-1} W^{\frac{1}{2}} K$
$s \leftarrow \text{diag}(A) \nabla_{\tilde{z}}^3 \log \pi(y|\theta, \hat{z})/2$
$\Omega \leftarrow aa^T/2 - R/2 + (s - RKs)c^T$
$\nabla_\xi \hat{\pi}_{\hat{z}}(y|\theta) \leftarrow \text{VJP}_K(\xi, \Omega)$
$\nabla_\eta \hat{\pi}_{\hat{z}}(y|\theta) \leftarrow \nabla_\eta \log \pi(y|\theta, \hat{z}) + \text{VJP}_c(\eta, As) - \frac{1}{2}\text{VJP}_W(\eta, A)$
$\nabla_\theta \hat{\pi}_{\hat{z}}(y|\theta) \leftarrow \text{CONCAT}([\nabla_\xi \hat{\pi}_{\hat{z}}(y|\theta), \nabla_\eta \hat{\pi}_{\hat{z}}(y|\theta)])$
$\nabla_\theta \hat{\pi}_{\hat{z}}(\theta, y) \leftarrow \nabla_\theta \hat{\pi}_{\hat{z}}(y|\theta) + \nabla_\theta \log p(\theta)$

Algorithm 3 Computing $\nabla_\theta \hat{z}$ and $\nabla_\theta \Sigma$

Require: Variables from Algorithm 2

$D \leftarrow I - RK$
$\text{VJP}_{\hat{z}}(\xi, v) \leftarrow \text{VJP}_K(\xi, Dvc^T)$
$\text{VJP}_\Sigma(\xi, V) \leftarrow \text{VJP}_K(\xi, DVD^T)$
$\text{VJP}_{\hat{z}}(\eta, v) \leftarrow \text{VJP}_c(\eta, Av)$
$\text{VJP}_\Sigma(\eta, V) \leftarrow \text{VJP}_W(\eta, -AVA)$

C Model details

In this section, we present more details of the models.

C.1 Synthesized Gaussian process

In this model, we generate the dataset as follows:

$$x_i \sim \begin{cases} \text{Uniform}(0, 2) & i \leq 50 \\ \text{Uniform}(2, 8) & i > 50 \end{cases}$$
$$y_i \sim \begin{cases} \text{Poisson}(\exp(\sin(2x_i) + 2)) & i \leq 50 \\ \text{Poisson}(\exp(\sin(2x_i) - 2)) & i > 50 \end{cases} \tag{15}$$

This construction makes the inference hard: the posterior inference should adapt to both easier parts ($i \leq 50$, larger y_i) and harder parts ($i > 50$, smaller y_i). In the experiments of this model, we use 10,000 warmup samples before collecting samples for estimation.

C.2 Sparse kernel interaction model

Following Margossian et al. [27], we use the dataset from Vanhatalo et al. [43] and define the model as follows.

$$\lambda_i \sim \text{StudentT}(\nu_{\text{local}}, 0, 1),\ \tau \sim \text{StudentT}(\nu_{\text{global}}, 0, s_{\text{global}}),\ c_{\text{aux}}, \chi \sim \Gamma^{-1}(s_{\text{df}}/2, s_{\text{df}}/2),$$
$$c = s_{\text{slab}}\sqrt{c_{\text{aux}}},\ \tilde{\lambda}_i^2 = c^2\lambda_i^2/(\tau^2 + c^2\lambda_i^2),\ \eta_2 = \tau^2\chi/c^2,$$
$$K_1 = X\text{diag}(\tilde{\lambda}^2)X^T,\ K_2 = (X \circ X)\text{diag}(\tilde{\lambda}^2)(X \circ X)^T$$
$$K = \frac{1}{2}\eta_2^2(K_1 + 1) \circ (K_1 + 1) - \frac{1}{2}\eta_2^2 K_2 - (\tau^2 - \eta_2^2)K_1 + c_0^2 - \frac{1}{2}\eta_2^2$$
$$z \sim \text{MVN}(0, K),\ y \sim \text{Bernoulli}(\text{logits} = z), \tag{16}$$

where $\text{local}, \text{global}, s_{\text{global}}, s_{\text{df}}, s_{\text{slab}}, c_0$ are given in the dataset, X is the data matrix and $\circ$ represents the element-wise multiplication.

Algorithm 4 RQMC-ADLA with MwG sampler

Require: Init positions θ_0, U_0
$\theta, U \leftarrow \theta_0, U_0$
for $t = 1, 2, \ldots, T$ **do**
  $\theta \leftarrow$ HMCkernel$(\hat{\pi}^{\mathrm{RQMC}}(\cdot, U, y), \theta)$
  for $d = 1, 2, \ldots, d_z$ **do**
    $U' \leftarrow U$
    $u \sim \mathrm{Uniform}(-0.1, 0.1)$
    $U'[d] \leftarrow (U[d] + u)\%1$
    $U \leftarrow$ MHkernel$(\hat{\pi}^{\mathrm{RQMC}}(\theta, \cdot, y), U, U')$
  end for
  $\theta_t, U_t \leftarrow \theta, U$
end for
return $\theta_{1:T}, U_{1:T}$

For this experiment, we use 200 covariates (indexed between 2500 and 2700) to construct X. During inference, we warm up with 10,000 samples before collecting samples.

C.3 Mixed-effects models

The `Epil2` model has the likelihood

$$y_i \sim \mathrm{NegativeBinomial}(\exp(x_i^T \beta + z_i^T u_{g_i}), \phi).$$

And the priors are

$$\sigma_\beta \sim \mathrm{normal}^+(0, 1),\ T \sim \mathrm{MVN}^+(0, \mathrm{diag}(1, 1)),\ L \sim \mathrm{LKJ}(2, 1),$$
$$\phi \sim \mathrm{Exponential}(1),\ \beta_k \sim \mathrm{normal}(0, \sigma_\beta^2),\ k = 1, ..., 6,\ u_j \sim \mathrm{MVN}(0, TLL^T T),\ j = 1, ..., G.$$

For efficient inference, we may want to marginalize each u_j from the model, as suggested in Lai et al. [26]. To be practical for Laplace approximation, we replace the distributions of $u_{1:G}$ and y_i as

$$u' \sim \mathrm{MVN}(0, Z\Sigma_{u_{1:G}} Z^T),\ y_i \sim \mathrm{NegativeBinomial}(\exp(x_i^T \beta + u_i'), \phi),$$

where Z is the design matrix for random effects and $\Sigma_{u_{1:G}}$ is a block diagonal matrix consisting of G blocks of $TLL^T T$. This model is equivalent to the previous model, but is better for Laplace approximation because the W matrix in Algorithm 2 becomes diagonal. During Laplace approximation, we also use the fact that $Z\Sigma_{u_{1:G}} Z^T$ is a block-diagonal matrix to speed up computation.

D Additional results

D.1 The raw importance sampling estimator

As a sanity check, we also tested the importance sampling-based model $\hat{\pi}^{\mathrm{IS}}_{\epsilon_{1:n}}(\theta, y)$ on the Gaussian process model. The results can be found in Figure 6. We find that it gives a higher error than QMC-ADLA, despite having the same cost. This is consistent in our other experiments, so we do not include it in our main results.

D.2 Additional sampling results

We include additional sampling results of our experiments in Tables 2, 3 and 4. In all models in our experiments, the Base method presents a non-negligible number of divergent transitions in HMC sampling. ADLA and our methods solve or mitigate this problem.

D.3 More convergence results

We selected two variables for each model to demonstrate the error convergence in the main text. In this part, we include more convergence results. For the synthesized Gaussian process model, we also estimate $\mathbb{E}[f_{80}]$ and $\mathbb{E}[\log \alpha]$ with the samples. The results are in Figure 7. We find our methods correct the error of ADLA and have faster or comparable convergence as HMC. For the sparse kernel

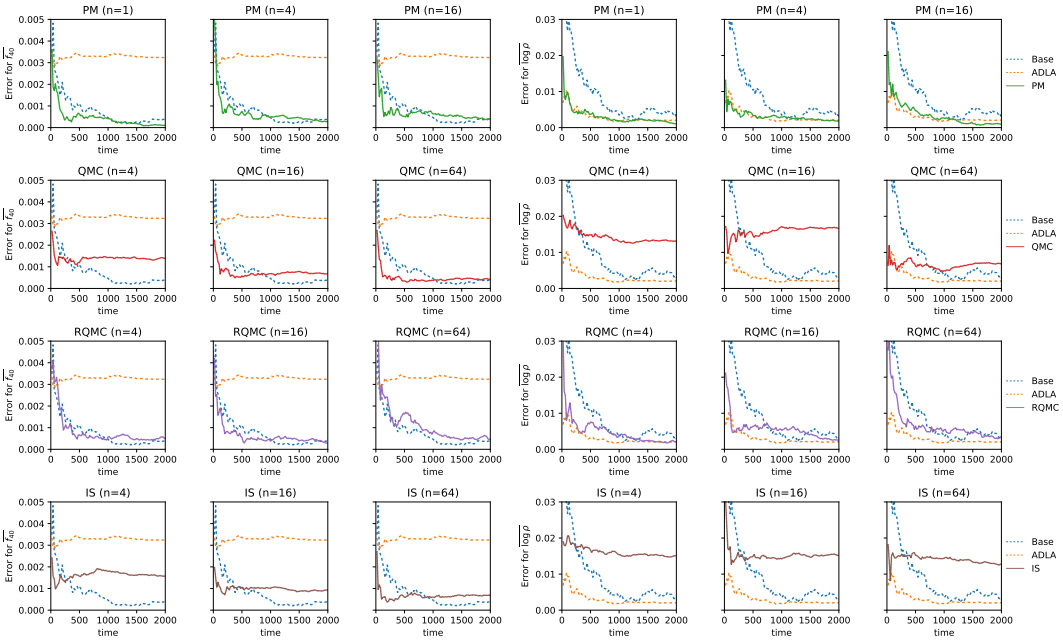


Figure 6: Error of estimating the means of parameters as a function of time in seconds for the synthesized Gaussian process with Poisson likelihood. Results are averaged from 5 independent runs. Ground-truth is estimated from NUTS on the unmarginalized model. The last row presents the results for $\hat{\pi}^{\mathrm{IS}}_{\epsilon_{1:n}}(\theta, y)$, and the first three rows are the same as Figure 1.

Method	n	Number of samples	Time (min)	Divergences
Base	-	100000	217 (1)	0 (0)
ADLA	-	100000	29 (2)	0 (0)
PM	1	100000	168 (2)	0 (0)
PM	4	100000	122 (0)	0 (0)
PM	16	100000	128 (3)	0 (0)
QMC	4	100000	46 (4)	0 (0)
QMC	16	100000	47 (6)	0 (0)
QMC	64	100000	53 (4)	0 (0)
RQMC	4	100000	88 (3)	0 (0)
RQMC	16	100000	126 (7)	0 (0)
RQMC	64	100000	224 (6)	0 (0)

Table 2: More details about the experiments on the synthesized Gaussian process model. Mean and standard deviation across 5 runs are reported.

interaction model, we estimate $\mathbb{E}[\log c_{\mathrm{aux}}]$ and $\mathbb{E}[\log \chi]$ and the results are in Figure 8. For these two estimations, our methods correct the error but converge slower than HMC. This is mainly because the two variables are easy to sample, so the benefits of marginalization are not significant. For the mixed-effects model, we show the estimation error for $\mathbb{E}[\beta_0]$ and $\mathbb{E}[\log T_1]$ in Figure 9. We find that, compared to HMC, ADLA and our methods have a higher error when estimating $\mathbb{E}[\beta_0]$ but a lower error when estimating $\mathbb{E}[\log T_1]$, although the magnitude of the first error is small. For the first error, it may be because there is still an error in the optimizer. For the second error, note that the raw HMC reports a significant number of divergent samples in Table 4, which biases the estimation.

Method	n	Number of samples	Time (min)	Divergences
Base	-	300000	892 (83)	4 (2)
ADLA	-	300000	431 (82)	0 (0)
QMC	4	100000	465 (166)	0 (0)
QMC	16	100000	358 (15)	0 (0)
QMC	64	100000	401 (3)	0 (0)
RQMC	4	100000	372 (12)	4 (6)
RQMC	16	100000	447 (22)	1 (1)
RQMC	64	100000	773 (153)	3 (4)

Table 3: More details about the experiments on the sparse kernel interaction model. Mean and standard deviation across 5 runs are reported.

Method	n	Number of samples	Time (min)	Divergences
Base	-	5000000	442 (42)	164838 (86482)
ADLA	-	4000000	328 (29)	10 (3)
QMC	4	2500000	424 (60)	5 (0)
QMC	16	2500000	543 (61)	8 (2)
QMC	64	1500000	465 (43)	4 (2)
RQMC	4	500000	429 (19)	2 (1)
RQMC	16	100000	490 (27)	0 (0)
RQMC	64	100000	794 (31)	0 (1)

Table 4: More details about the experiments on the mixed effects models. Mean and standard deviation across 5 runs are reported.

D.4 The posterior of RQMC-ADLA

In RQMC-ADLA, an issue is that $\hat{\pi}^{\mathrm{RQMC}}(\theta, U, y)$ is not continuous with U. Here we demonstrate this problem with a simple example model:

$$\theta \sim \mathrm{normal}(0, 1),\ z \sim \mathrm{normal}(\theta, 1),\ y \sim \mathrm{Cauchy}(z, 1), \tag{17}$$

where the observation is $y = 2$. We can compute $\hat{\pi}^{\mathrm{RQMC}}(U|\theta, y)$ with numerical integration. In Figure 10, we see that the discontinuity exists in this model. The number of discontinuous points is the same as n. Compared with the uncountable number of real numbers in $[0, 1]$, the probability of sampling at a discontinuous point remains 0 even with large n. Also, as we increase n, $\hat{\pi}^{\mathrm{RQMC}}(U|\theta, y)$ becomes more uniform, implying easier inference.

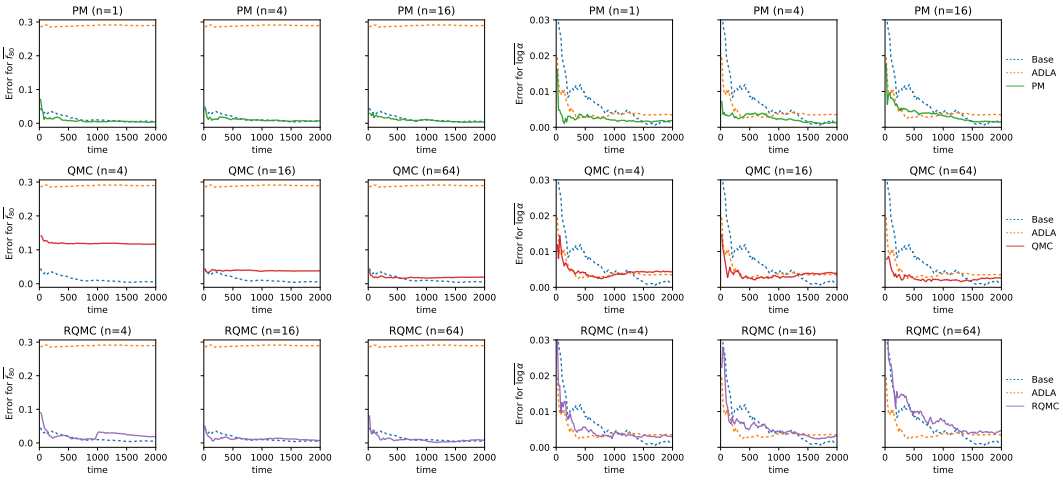


Figure 7: Error of estimating $\mathbb{E}[f_{80}]$ and $\mathbb{E}[\log \alpha]$ as a function of time in seconds for the synthesized Gaussian process with Poisson likelihood. Results are averaged from 5 independent runs.

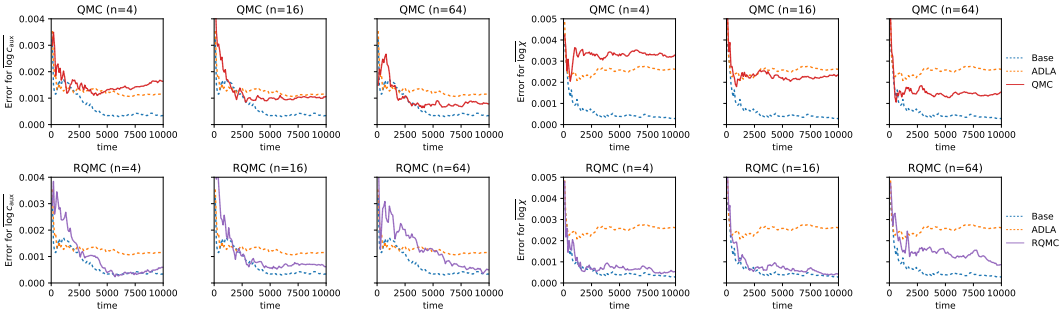


Figure 8: Error of estimating $\mathbb{E}[\log c_{\text{aux}}]$ and $\mathbb{E}[\log \chi]$ as a function of time in seconds for the sparse kernel interaction model. Results are averaged from 5 independent runs.

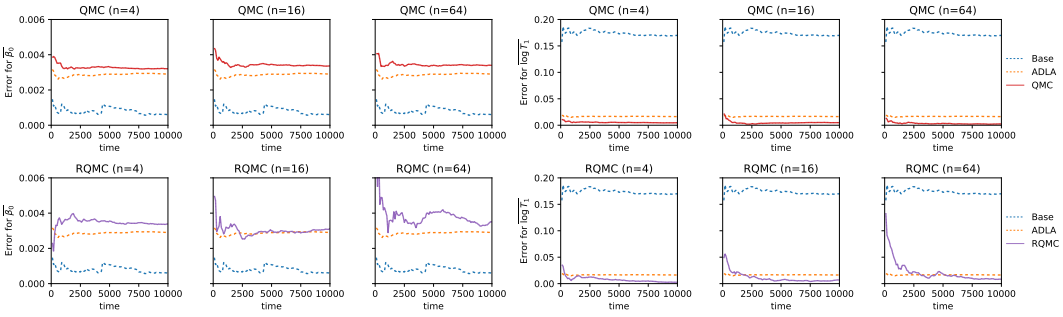


Figure 9: Error of estimating $\mathbb{E}[\beta_0]$ and $\mathbb{E}[\log T_1]$ as a function of time in seconds for the mixed-effects model. Results are averaged from 5 independent runs.

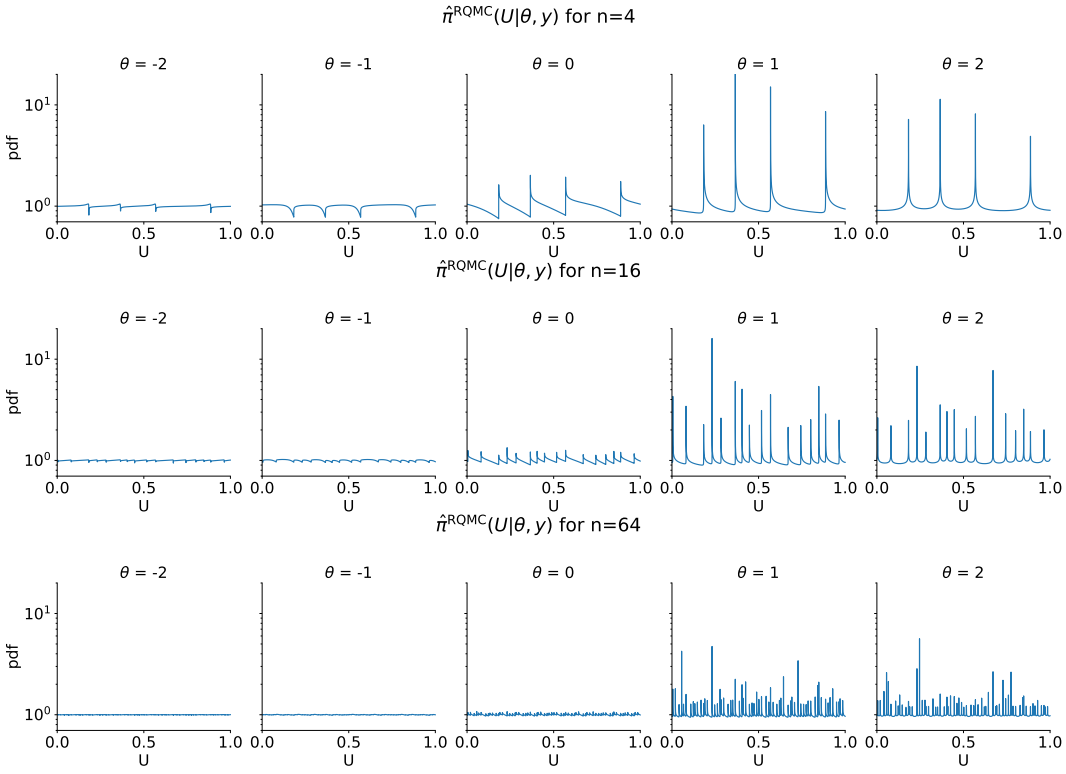


Figure 10: Distribution of $\hat{\pi}^{\mathrm{RQMC}}(U|\theta, y)$ with different θ and n for the simple example model, using 5,000 evaluation points in $[0, 1]$. The distribution is more uniform as we increase n. Also, $\hat{\pi}^{\mathrm{RQMC}}(U|\theta, y)$ is not continuous.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our contributions are novel methods for correcting the error of integrated Laplace approximation, which are well articulated in the abstract and introduction.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: We have Table 1 about the tradeoffs of different methods. A detailed limitation statement can be found in Section 9.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: We stated our theoretical results in the main text and proved them in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have provided the algorithmic details and hyperparameter settings that are enough for reproducing the results.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The link to the code is provided.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We have included all the details in our experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: For numerical results, we estimate mean and standard deviation of 5 independent runs.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We have specified the CPU used in our experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the Code of Ethics and conformed with it.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This paper presents methodological work whose goal is to advance the area of Bayesian inference. There may be potential societal consequences of our work when it is applied in specific domains, none of which we feel must be specifically highlighted here.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: We use public available data and models.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We have properly credited the data and models.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: This work does not release new assets.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.