GP-MALA - STAT 547D Project Proposal

Zachary Lau

February 21, 2026

Abstract

MCMC methods are particularly useful for estimating expectations in higher-dimensional parameter spaces where typical quadrature methods suffer from the curse of dimensionality. However, even these algorithms may not be practical when the likelihood is very expensive to evalute, such as when it requires solving complex systems of differential equations or expensive matrix inversions. One alternative which has been proposed is to use a Gaussian Process to model the log-likelihood or log posterior. Our method focusses on sampling directly from the joint posterior model implied by the Gaussian process model of the log-likelihood. In this project we propose a (to the best of the author's knowledge) novel method for sampling from the approximate model implied by the GP based on the Metropolis Adjusted Langevin Algorithm (MALA). Instead of using function evaluations, the method uses samples from the GP posterior to obtain likelihood values and gradients. We will (hopefully) show that this leads to unbiased estimates of posterior quanities of interest with respect to our proposed joint model, and compare its performance to simpler but potentially biased schemes.

1 Introduction and Background

1.1 Likelihood free and surrogate likelihood methods

The challenge of performing Bayesian inference with intractable or expensive likelihood functions has been approached from multiple directions in the past. Rasmussen [4] suggests making proposals using HMC applied to a surrogate model of the log posterior and then performing a Metropolis-Hastings accept-reject with an evaluations of the true likelihood. However, it is not clear that this scheme has the correct stationary distribution. Christen & Fox [2] introduced delayed acceptance MCMC which similarly makes proposals using a surrogate, but uses a pre-screening step to ensure that the chain targets the correct distribution. On the other hand, likelihood free approaches attempt to perform bayesian inference when the likelihood function is not available or tractable, but simulations can be made from the data generating process. One such framework is Approximate Bayesian Computation (ABC) [1]. More recently Gutmann & Corander [3] have considered applying Bayesian optimization to improve simulation efficiency. Finally Stuart et al. [6] discuss error bounds on the distance between the true posterior and a Gaussian Process approximation of the posterior in terms of Hellinger distance. Our work differs from that of Rasmussen and Christen & Fox in that we focus on sampling directly from the surrogate model without a correction step. Furthermore, it differs from ABC and likelihood free methods in that the surrogate model is based on true evaluations of the likelihood.

1.2 Gaussian Processes

Gaussian processes (GP) modelling is a type of non-parametric Bayesian model. It can be viewed as placing a Gaussian process prior on a function, i.e. observations $y_1, y_2, \ldots, y_n$ at any finite set of input points $x_1, x_2, \dots x_n$ are distributed according to a multivariate normal distribution. Via the Kolmogorov extension theorem, such a prior is fully specified by a mean function $\mu(x)$ and the covariance function $k(x, x')$. Given a set of training points $x_1, \dots, x_n$, inference at a set of test points $x'_1, \dots, x'_m$ is found by computing the conditional distribution of $y'_1, \dots, y'_m$ given $y_1, \dots, y_n$. The resulting poisterior distribution sees $[y'_1, \dots, y'_m]^\top \sim \mathcal{N}(\mu', K')$ where μ' and K' are given by the inference equations [5]

$$\mu' = \mu + k^*(K + \sigma^2 I)^{-1}\vec{y}$$
$$K' = k^{**} - k^*(K + \sigma^2 I)^{-1}\vec{y}$$

Where μ represents the prior mean, $k^{**}_{ij} = k(x'_i, x'_j)$ gives the prior covariance, $k^*_{ij} = k(x'_i, x_j)$ the cross-covariance and $\vec{y}_i = y_i$ the vector of observations, and σ^2 represents i.i.d. observational noise (even when there is no noise, a small amount of noise may be included for numerical stability of the inverse). An important feature of Gaussian process modelling is that in addition to giving estimates at new test points through the posterior mean, it also quantifies uncertainty through the posterior covarianace.

We will consider GP's from two perspectives. Firstly bayesian hyperparameter inference in GP's provides a good test case where the likelihood evaluation can be prohibitive. Secondly GP's can themselves be used as a surrogate model for the likelihood that quantifies uncertainty at unobserved locations.

1.3 Hyper-parameter inference for Gaussian Processes

Typically the covariance function used for GP inference belongs to a parametric family. For example, the squared-exponential family in one-dimension has the form $k(x, x') = \sigma_f^2 \exp\left(-\frac{1}{2}\left(\frac{x-x'}{\ell}\right)^2\right)$ with hyperparametgers σ_f^2 and ℓ. It is most common to choose these parameters given the data by maximizing the so-called log-marginal likelihood, i.e. the likelihood of the observations after marginalizing over the unobserved latent function. With observation noise, this function becomes

$$l(\theta) = -\frac{1}{2}y^\top(K + \sigma^2 I)^{-1}y - \frac{1}{2}\log|K + \sigma^2 I| - \frac{n}{2}\log(2\pi)$$

An alternative to using MLE is to place a prior over the hyperparameters, and then perform inference by averaging over the hyperparamter posterior. Our target then becomes

$$\log p(\theta|y) = \log p(\theta) - \frac{1}{2}y^\top(K + \sigma^2 I)^{-1}y - \frac{1}{2}\log|K + \sigma^2 I| - \frac{n}{2}\log(2\pi)$$

Some of the major challenges with sampling from such a distribution include multi-modality and the computational cost of computing matrix solves and determinants. The latter can be a particular challenge in the case of GP's with large numbers of observations, particularly since their cost scales as $O(n^3)$.

2 Algorithms

2.1 Set up and motivation

Firstly we consider the case where we have some Bayesian inference problem with a tractable but expensive likelihood function. Gradient information may or may not be available. The first premise of this work, who's validity may be explored in future work, is that in the expensive likelihood regime Bayesian inference should be separated into two parts

- *Likelihood model building*: judicious evaluations of the likelihood function to provide the most useful understanding of the likelihood function
- *Sampling*: Generation of samples from the posterior implied *by our model*

In this work we focus on the second phase, leaving to future work the evaluation of the first. Concretely, suppose that we have some pre-existing set of log-likelihood evaluations $l(\theta_1), \dots, l(\theta_m)$ (hopefully judiciously chosen). We construct a a GP model of the function l over the parameter space, call this model $\mathcal{M}$. Our work here explores the problem of sampling from θ given a model of its likelihood *that incorporates uncertainty.*

2.2 Mean estimation

The most obvious way to sample from the GP surrogate model is to use to use the posterior mean as the log likelihood. That is, if the posterior mean is given by $\mu(\theta)$ we take $l(\theta) = \mu(\theta)$ so that the likelihood is given by $\mathcal{L}(\theta) = \exp(\mu(\theta))$. Another method is is to consider the posterior mean of f itself. Noting that $f(\theta)$ is marginally log-normal it has expectation $\mathbb{E}f(\theta) = \exp\left(\mu(\theta) + \frac{1}{2}\sigma^2(\theta)\right)$, giving log-likelihood $l(\theta) = \mu(\theta) + \frac{1}{2}\sigma^2(\theta)$. This second estimate takes into account uncertainty, but only insofar as it affects the mean. The posteriors arising from such likelihoods can be targeted using off the shelf samplers like NUTS.

2.3 Uncorrected Langevin Dynamics

The third method we propose to analyze involves taking *samples* from the gradient instead of using a fixed surrogate model in order to reflect uncertainty in the surrogate model. Importantly, we note that the gradient of a GP and the function itself are jointly still a GP, so sampling amounts to sampling from a GP posterior. We restrict ourselves to uncorrected langevin dynanamics because this requires only a single sample of the gradient. Sampling a longer HMC trajectory would be significantly more expensive because each gradient sample would need to be taken conditional on the previous samples. With sample i requiring a roughly $O(i^2)$ block matrix update to the inverse, the total trajectory cost would be roughly $O(L^3)$. On the other hand, given a precomputed inverse that is shared amongst all points, a single Langevin step with a model built on m true evaluations requires only a single $O(m^2)$ matrix multiplication.

2.4 Theoretical optimum

As discussed in [6], the theoretical optimum distribution, in a sense, to sample from would be the *joint* distribution over f and θ, i.e. we the model defined by the following generative process.

$$\begin{aligned} \log f &\sim \mathcal{M} \\ \theta &\sim \frac{p_0(\theta)f(\theta)}{\mathcal{Z}(f)} \end{aligned}$$

Unfortunately, sampling from this joint distribution is a difficult problem that our proposed methods only approximate in the well informed case.

Firstly we consider our Langevin dynamics inspired sampler. This can be seen as an approximate Metropolis-within-Gibb's sampler where the marginal $f|\theta$ is approximated by the GP posterior. Indeed in the limit of small step sizes our update of θ preserves the target distribution. However, a true conditional update of f would need to preserve (see derivation in Appendix)

$$\begin{aligned} p(f|\theta) &= \frac{p(f,\theta)}{p(\theta)} \\ &\propto \frac{p(f)f(\theta)}{\mathcal{Z}(f)} \end{aligned}$$

On the other hand, if we approximate $\mathcal{Z}(f)$ as a constant, we recover the log mean sampler $\hat{l}(\theta) = \mu(\theta) + \frac{1}{2}\sigma^2(\theta)$. To see this note that integrating our joint distribution over f we find

$$\begin{aligned} p(\theta) &= \int p(f,\theta)df \\ &\propto \int p_0(\theta)f(\theta)p(f)df \\ &= p_0(\theta)\mathbb{E}f(\theta) \\ &= p_0(\theta)\exp\left(\mu(\theta) + \frac{1}{2}\sigma^2(\theta)\right) \end{aligned}$$

Thus none of our samplers targets the “correct” stationary distribution; finding a sampling mechanism that is able to account for the $\mathcal{Z}(f)$ factor remains an open problem. However, we note that heuristically given more evaluations of the true likelihood (and hence less variane in our model) these constant approximations will be approximately true as the influence of any one sample diminishes.

3 Experiments

We test each of the three methods

4 Discussion

In this project we provide a renewed exploration of the idea of using surrogate models of the likelihood function in Bayesian inference, particularly with a focus on Bayesian inference for Gaussian process hyperparameter learning. Our proposed algorithm uses samples of the gradient in an attempt to account for surrogate model uncertainty. While it has experimental and theoretical shortfalls, we hope this opens the door for future research on understanding how surrogate models which quanntify uncertainty can be used in Bayesian inference.

Specifically we envision two areas of future research. Firstly, in a similar vein to that discussed in this project, future work may include analysis of new algorithms and stochastic surrogate models where targetting the joint distribution over the model and the inference paramter becomes feasible. A second future vein which has not been explored by this work includes exploiting surrogate models' measures of uncertainty to guide future true likelihood evaluations in a manner similar to Bayesian optimization and active learning.

References

[1] Mark A. Beaumont. Approximate bayesian computation. *Annual Review of Statistics and Its Application*, 6:379–403, 2019.

[2] J. Andrés Christen and Colin Fox. Markov chain monte carlo using an approximation. *Journal of Computational and Graphical Statistics*, 14(4):795–810, 2005.

[3] Michael U. Gutmann and Jukka Corander. Bayesian optimization for likelihood-free inference of simulator-based statistical models. *Journal of Machine Learning Research*, 17(125):1–47, 2016.

[4] Carl Edward Rasmussen. Gaussian processes to speed up hybrid monte carlo for expensive bayesian integrals. In J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, editors, *Bayesian Statistics 7*, pages 651–659, Oxford, UK, 2003. Oxford University Press.

[5] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, USA, 2006.

[6] Andrew M. Stuart and Aretha L. Teckentrup. Posterior consistency for gaussian process approximations of bayesian posterior distributions. *Mathematics of Computation*, 87(310):721–753, 2018.

A Target conditional

Given the generative model, we get joint distribution

$$p(f,\theta) = p(f)\frac{p_0(\theta)f(\theta)}{\mathcal{Z}(f)}$$

Where the normalizing constant $\mathcal{Z}(f) = \int p_0(\theta)f(\theta)d\theta$. This gives the θ marginal of

$$\begin{aligned} p(\theta) &= \int p(f)\frac{p_0(\theta)f(\theta)}{\mathcal{Z}(f)}df \\ &= p_0(\theta)\int p(f)\frac{f(\theta)}{\mathcal{Z}(f)}df \end{aligned}$$

And then the conditional is

$$p(f|\theta) \propto \frac{p(f)f(\theta)}{\mathcal{Z}(f)}$$

Intuitively this skews f higher at the current value of θ, but this scaling is tempered by the subsequent increase in the normalizing constant. If we assume that $\mathcal{Z}(f)$ is relatively constant, then sampling $f(\theta)|\theta$ corresponds to sampling from $p(y) \propto y \log\mathcal{N}(\mu, \sigma^2)$ which corresponds to the tilted log normal $\log\mathcal{N}(\mu + \sigma^2, \sigma^2)$.

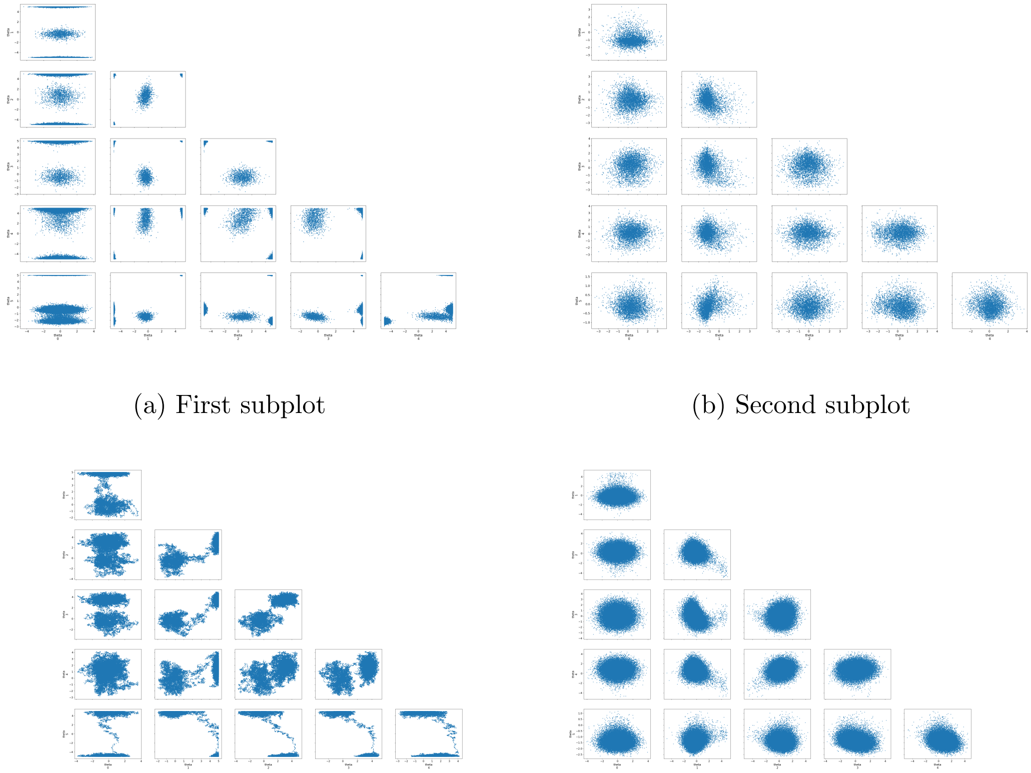
B Algorithms

Algorithm 1 shows the algoritm we use for uncorrected Langevin dynamics using GP sampled gradients. In implementation step size is scaled down proportional to the standard deviation of observed log-likelihoods.

Algorithm 1 Unadjusted Langevin for GP Likelihood

```
Require: Stepsize $\tau$, Iterations $N$
  Compute $K$ the covariance amongst train points
  $Ki \leftarrow (K + \sigma^2 I)^{-1}$
  Initialize $\theta \sim p_0(\theta)$
  for $i = 0$ to N do
      Sample $\nabla \log f(\theta)$ from GP model
      Sample $\xi \sim \mathcal{N}(0, 1)$
      $\theta \leftarrow \theta + \tau \nabla \log f(\theta) + \sqrt{2\tau}\,\xi$
      Save $\theta$
  end for
```

C Figures and Tables



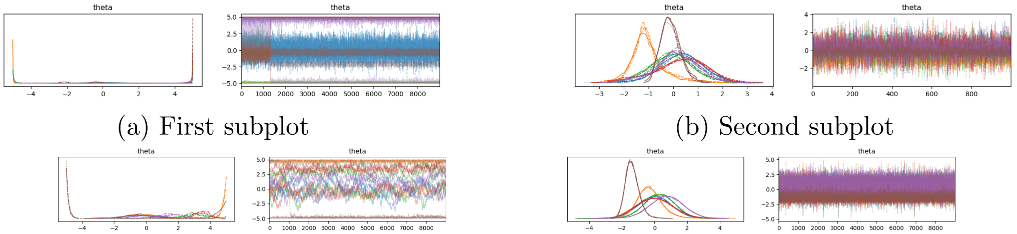
(a) First subplot

(b) Second subplot

(c) Second subplot

(d) Second subplot

Figure 1: Main figure caption



(a) First subplot

(b) Second subplot

(c) Second subplot

(d) Second subplot

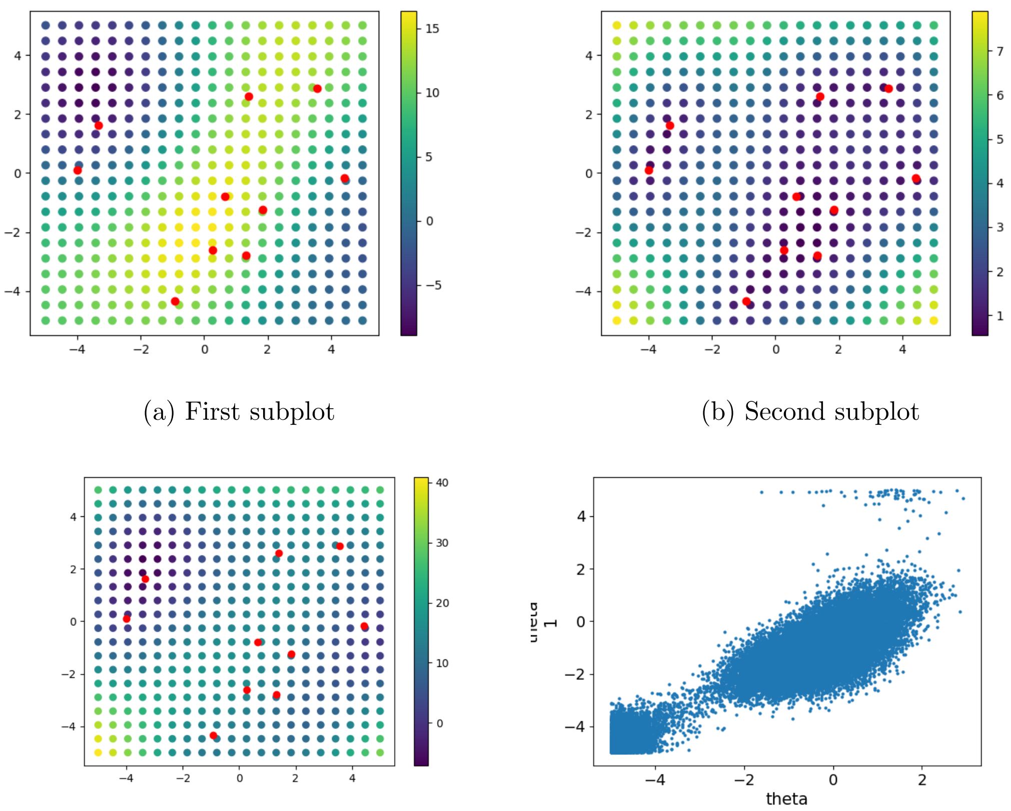
Figure 2: Main figure caption

Parameter	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
theta[0]	-0.01	1.00	-1.92	1.84	0.01	0.01	23649.0	20483.0	1.00
theta[1]	-2.33	4.30	-5.00	4.99	2.12	1.16	7.0	26.0	1.60
theta[2]	-0.17	4.84	-5.00	4.99	2.40	0.12	6.0	91.0	1.76
theta[3]	4.76	1.02	4.79	5.00	0.26	0.64	23.0	23.0	1.11
theta[4]	2.68	3.96	-4.96	5.00	1.88	1.24	7.0	23.0	1.53
theta[5]	0.56	2.66	-2.26	5.00	1.32	0.72	5.0	23.0	2.42

Table 1: Posterior summary statistics for adjusted mean target sampled with NUTS

Parameter	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
theta[0]	-0.10	1.03	-1.79	1.97	0.15	0.07	49.0	101.0	1.05
theta[1]	3.52	2.23	-0.85	5.00	1.07	0.61	7.0	33.0	1.53
theta[2]	2.03	1.77	-1.45	4.44	0.76	0.36	8.0	73.0	1.51
theta[3]	2.43	1.84	-1.21	4.47	0.84	0.46	7.0	32.0	1.52
theta[4]	1.16	1.22	-1.43	3.32	0.39	0.16	11.0	42.0	1.29
theta[5]	-2.27	4.22	-5.00	4.89	2.02	1.04	7.0	42.0	1.50

Table 2: Posterior summary statistics with GP sampled gradients



(a) First subplot

(b) Second subplot

(c) Second subplot

(d) Second subplot

Figure 3: Main figure caption