Gaussian Processes to speed up MCMC with automatic exploratory-exploitation effect

Alessio Benavoli, Jason Wyse, Arthur White, SCSS, Trinity College Dublin, Ireland

Introduction

Probabilistic programming languages rely on general-purpose automatic MC sampling techniques for computing accurate approximations of the posterior. These techniques require repeated evaluations of the likelihood function, which can be computationally demanding in some applications. We consider the problem of sampling from a posterior

$$\pi(\boldsymbol{\theta}|D) \propto p(D|\boldsymbol{\theta})p(\boldsymbol{\theta}),$$

where D denotes data and $\boldsymbol{\theta} \in \Theta$ is a vector of unknown parameters, in the case where the likelihood $p(D|\boldsymbol{\theta})$ is costly to evaluate.
We discuss two-stage algorithms.

Adaptive Metropolis-Hastings

Stage 1: adaptive Metropolis-Hastings (MH) algorithm which employs an adaptively tuned Gaussian Process (GP) surrogate model at the first stage to filter out poor proposals.
Stage 2: If a proposal is not filtered out, at the second stage a full (expensive) log-likelihood evaluation is carried out and used to decide whether it is accepted as the next state. Introduction of the 1st stage saves computation on poor proposals.
Key contribution is in the form of the acceptance probability in the first stage obtained by marginalising out the GP function. This makes the acceptance ratio dependent on the variance of the GP, which naturally results in an exploration-exploitation trade-off similar to the one of Bayesian Optimisation

MALA

The second algorithm is a two-stage form of Metropolis adjusted Langevin algorithm (MALA). Here, we use GP as a surrogate for the log-likelihood function again, but in this case the GP is also used to approximate the gradient required for MALA updating, using a well known result that the gradient of a GP is also a GP. **See papers for details.**

The approximation we use is

$$\mathsf{LL}(D|\boldsymbol{\theta}) := \ln p(D|\boldsymbol{\theta}) \approx \widetilde{\mathsf{LL}}_t(D|\boldsymbol{\theta}) \sim \mathsf{GP}(\mu(\boldsymbol{\theta}|\mathscr{I}_t), k(\boldsymbol{\theta},\boldsymbol{\theta}^*|\mathscr{I}_t)) \quad (1)$$

where $\mathscr{I}_t$ denotes the set of t full evaluations of the log-likelihood by the current iteration, and $\boldsymbol{\theta}^*$ collectively denotes the parameter values at which these evaluations were made.

Notation and explicit definition of $\mathscr{I}_t$:

- $\mathscr{S}_k$ denotes the points sampled up to the iteration k of the algorithm;
- $\boldsymbol{\theta}^{(k)}$ denotes the most recent element in $\mathscr{S}_k$ and $\boldsymbol{\theta}^*$ denotes the proposed state
- $\mathscr{I}_t = \{(\boldsymbol{\theta}^{(i)}, \mathsf{LL}(D|\boldsymbol{\theta}^{(i)})) : \; i = 1,\dots,t\}$ denotes the $t \leq k$ exact likelihood evaluations performed up to iteration k.

We use a noise free GP as a surrogate model for the log-likelihood and denote by $\mathsf{GP}_k(\mu(\boldsymbol{\theta}|\mathscr{I}_t), k(\boldsymbol{\theta},\boldsymbol{\theta}|\mathscr{I}_t))$ the posterior GP at the iteration k conditioned on the collection $\mathscr{I}_t$. We use $\widetilde{\mathsf{LL}}_k(\boldsymbol{\theta})$ to denote the GP-distributed log-likelihood. We choose the parameters of the GP to satisfy the following exact interpolation property.

Assumption 1 *The prior mean and covariance function of the GP are selected to guarantee exact interpolation:*

$$\mu(\boldsymbol{\theta}^{(i)}|\mathscr{I}_t) = \mathsf{LL}(D|\boldsymbol{\theta}^{(i)}), \qquad k(\boldsymbol{\theta}^{(i)},\boldsymbol{\theta}|\mathscr{I}_t) = 0,$$

for all $\boldsymbol{\theta}^{(i)}$ with a corresponding entry in $\mathscr{I}_t$ and $\boldsymbol{\theta} \in \Theta$.

The two stages of the MH algorithm are as follows.

Stage 1

Use the predictive posterior GP (conditioned on the collection $\mathscr{I}_t$) to approximate the log-likelihood. Define the first stage acceptance probability:

$$\tilde{\alpha}^{(1)}(\boldsymbol{\theta}^{(k)},\boldsymbol{\theta}^*) = 1 \wedge \frac{\exp(\widetilde{\mathsf{LL}}_t(\boldsymbol{\theta}^*))p(\boldsymbol{\theta}^*)q(\boldsymbol{\theta}^{(k)}|\boldsymbol{\theta}^*)}{\exp(\widetilde{\mathsf{LL}}_t(\boldsymbol{\theta}^{(k)}))p(\boldsymbol{\theta}^{(k)})q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(k)})} \quad (2)$$

where $\widetilde{\mathsf{LL}}_t(\cdot) \sim \mathsf{GP}_k(\mu(\boldsymbol{\theta}|\mathscr{I}_t), k(\boldsymbol{\theta},\boldsymbol{\theta}|\mathscr{I}_t))$ and we use the shorthand notation $a \wedge b = \min(a,b)$. Note that, because of the exact interpolation property in Assumption 1, it results that $\widetilde{\mathsf{LL}}_t(\boldsymbol{\theta}^{(k)}) = \mathsf{LL}_t(\boldsymbol{\theta}^{(k)})$.

The acceptance probability $\tilde{\alpha}^{(1)}(\boldsymbol{\theta}^{(k)},\boldsymbol{\theta}^*)$ (respectively, $\tilde{\alpha}^{(1)}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^{(k)})$) depends on $\widetilde{\mathsf{LL}}_t(\boldsymbol{\theta}^*) - \widetilde{\mathsf{LL}}_t(\boldsymbol{\theta}^{(k)})$ (respectively, $-\widetilde{\mathsf{LL}}_t(\boldsymbol{\theta}^*) + \widetilde{\mathsf{LL}}_t(\boldsymbol{\theta}^{(k)})$) which is GP distributed.

Proposition 1 *The distribution of $e^{\widetilde{\mathsf{LL}}_t(\boldsymbol{\theta})}$ is Lognormal $(\mu(\boldsymbol{\theta}|\mathscr{I}_t), k(\boldsymbol{\theta},\boldsymbol{\theta}|\mathscr{I}_t))$, and its mean is*

$$e^{\mu(\boldsymbol{\theta}|\mathscr{I}_t)+\frac{1}{2}k(\boldsymbol{\theta},\boldsymbol{\theta}|\mathscr{I}_t)}. \quad (3)$$

By exploiting Proposition 1, we remove the dependence of the acceptance probability on $\widetilde{\mathsf{LL}}$ in (2) resulting in the acceptance probability:

$$\alpha^{(1)}(\boldsymbol{\theta}^{(k)},\boldsymbol{\theta}^*) = 1 \wedge \frac{e^{\mu(\boldsymbol{\theta}^*|\mathscr{I}_t)+\frac{1}{2}k(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*|\mathscr{I}_t)}p(\boldsymbol{\theta}^*)q(\boldsymbol{\theta}^{(k)}|\boldsymbol{\theta}^*)}{e^{\mu(\boldsymbol{\theta}^{(k)}|\mathscr{I}_t)}p(\boldsymbol{\theta}^{(k)})q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(k)})}, \quad (4)$$

where $\mu(\boldsymbol{\theta}^{(k)}|\mathscr{I}_t) = \mathsf{LL}(\boldsymbol{\theta}^{(k)})$ is the exact log-likelihood (by Assumption 1).

Given (4), in Stage 1, we accept $\boldsymbol{\theta}^*$ with probability $\alpha^{(1)}(\boldsymbol{\theta}^{(k)},\boldsymbol{\theta}^*)$, otherwise $\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)}$. This defines the following transition kernel at Stage 1:

$$Q_k^*(A|\boldsymbol{\theta}^{(k)}) = \int_A \alpha^{(1)}(\boldsymbol{\theta}^{(k)},\boldsymbol{\theta}^*)q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(k)})d\boldsymbol{\theta}^* + I_A(\boldsymbol{\theta})\int_\Theta (1-\alpha^{(1)}(\boldsymbol{\theta}^{(k)},\boldsymbol{\theta}^*))d\boldsymbol{\theta}^*. \quad (5)$$

One can show that the above transition kernel satisfies the detailed balance property for the approximated target distribution $e^{\mu(\boldsymbol{\theta}|\mathscr{I}_t)+\frac{1}{2}k(\boldsymbol{\theta},\boldsymbol{\theta}|\mathscr{I}_t)}p(\boldsymbol{\theta})$.

Stage 2.

We perform another MH acceptance step, evaluating the exact log-likelihood. Let $\boldsymbol{\theta}^*$ denote a point sampled from $q_k^*(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(k)}) := Q_k^*(d\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(k)})$. Note that, $\boldsymbol{\theta}^*$ is either equal to the point $\boldsymbol{\theta}^*$ sampled at Stage 1 or to $\boldsymbol{\theta}^{(k)}$ if $\boldsymbol{\theta}^*$ was rejected at Stage 1. So, with probability

$$\begin{aligned}\alpha^{(2)}(\boldsymbol{\theta}^{(k)},\boldsymbol{\theta}^*) &= 1 \wedge \frac{\exp(\mathsf{LL}(D|\boldsymbol{\theta}^*))p(\boldsymbol{\theta}^*)q_k^*(\boldsymbol{\theta}^{(k)}|\boldsymbol{\theta}^*)}{\exp(\mathsf{LL}(D|\boldsymbol{\theta}^{(k)}))p(\boldsymbol{\theta}^{(k)})q_k^*(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(k)})}\\ &= 1 \wedge \frac{\exp(\mathsf{LL}(D|\boldsymbol{\theta}^*))p(\boldsymbol{\theta}^*)q(\boldsymbol{\theta}^{(k)}|\boldsymbol{\theta}^*)\alpha^{(1)}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^{(k)})}{\exp(\mathsf{LL}(D|\boldsymbol{\theta}^{(k)}))p(\boldsymbol{\theta}^{(k)})q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(k)})\alpha^{(1)}(\boldsymbol{\theta}^{(k)},\boldsymbol{\theta}^*)},\end{aligned} \quad (6)$$

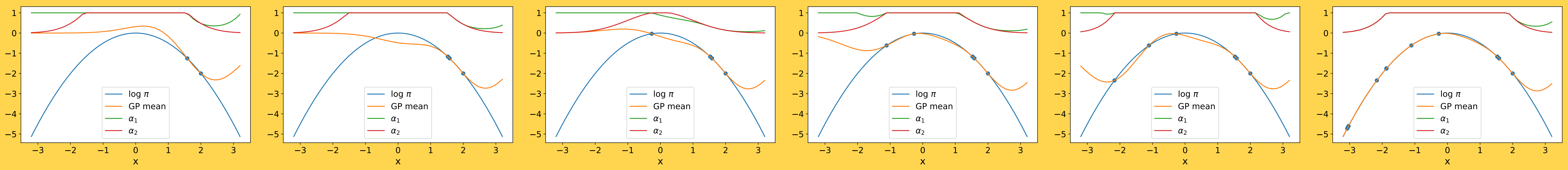
we accept $\boldsymbol{\theta}^*$, otherwise $\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)}$.

Convergence

$$\lim_{t\to\infty} \sup_{\boldsymbol{\theta}\in\Theta} ||P_t(\cdot|\boldsymbol{\theta}) - P_{t-1}(\cdot|\boldsymbol{\theta})|| = 0 \quad \text{in probability.} \quad (7)$$

Toy Example

We consider a 1D case with $\pi(\theta) \propto e^{-\frac{x^2}{2}}$. It can be noticed how α_1 converges to α_2 at the increase of the log-likelihood evaluations.



Simulations

We consider five target distributions.

T1: 2D posterior of the parameters a, b of the banana shape distribution (true value set to $a = 0.2, b = 2$);

T2: 3D posterior of the parameters a, b, σ of the nonlinear regression model $y = a\frac{x}{x+b} + \varepsilon$, $\varepsilon \sim N(0, \sigma^2)$ (true value set to $a = 0.14, b = 50, \sigma = 0.1$).

T3: 3D posterior of the parameters ℓ_1, ℓ_2, σ^2 of the SE kernel for a GP-classifier.

T4: 4D posterior of the parameters $\beta, \gamma, \sigma_1, \sigma_2$ of a Susceptible, Infected, Recovery (SIR) model.

T5: 5D posterior of the parameters $\beta_0, \dots, \beta_4$ of a parametric logistic regression problem.

We compare our two-stage algorithm with the standard implementations of MH and MALA. For each target problem and in each simulation, we generate 2500 samples (500 for burnin). We have deliberately selected a small number of samples to show that our approach converges quickly, which is important in computationally expensive applications.

		AR	ESS	ESJD	Eval%	SD
T1	MH	0.37	90	0.13	100	0.02
	GP-MH	0.36	113	0.13	**41**	0.02
	MALA	0.26	73	0.2	100	0.03
	GP-MALA	0.26	75	0.2	**35**	0.02
T2	MH	0.28	138	32.6	100	339
	GP-MH	0.27	133	31	**39**	339
	MALA	0.26	220	51	100	316
	GP-MALA	0.21	147	29	**43**	255
T3	MH	0.42	137	0.44	100	4.1
	GP-MH	0.42	135	0.38	**42**	3.5
	MALA	0.44	133	0.48	100	4.1
	GP-MALA	0.43	134	0.45	**45**	3.5
T4	MH	0.1	51	0.003	100	0.009
	GP-MH	0.1	45	0.003	**15**	0.009
T5	MH	0.29	98	0.002	100	0.006
	GP-MH	0.29	102	0.002	**35**	0.006
	MALA	0.67	339	0.009	100	0.006
	GP-MALA	0.67	368	0.009	**68**	0.006