**RESPONSE TO REFEREES OF THE MANUSCRIPT**
***FAST AUTOMATIC BAYESIAN CUBATURE USING LATTICE SAMPLING***

Thanks to the the referees and the editors for their very helpful comments. We have revised our manuscript substantially in response. Specific concerns are addressed point-by-point below in blue. We hope that all will find the revision satisfactory.

Reviewer 1: The paper is an interesting read on Bayesian cubature ...

The speedup of their algorithm is in the linear algebra part of the calculations. But the derivation in Sections 4.3 and 4.4 can be obtained much faster by realising that their definition of the covariance matrix, ... is a permutation of a circulant matrix from which all required properties follow. ... The current Sections 4.3 and 4.4 are therefore not needed. We have altered these sections substantially.

A second contribution is made in section 4.5 titled 'Overcoming cancellation error', but here it is not so clear to me if what is proposed will really help. See further. We believe that it does help, both experientially and theoretically.

- Abstract
  - The second sentence doesn't make sense to me. Re-written.
  - In the third sentence I would replace 'postulates' with 'considers'. Done.
- Section 1
  - The authors write 'We construct a reliable stopping criterion' but it is unclear what they mean with 'reliable'. This paragraph has been rewritten for clarity.
  - On page 2 they write 'If function evaluation is expensive, then $\$(f)$ might be similar in magnitude to $\log(n)$.': this makes no sense as evaluating $f$ depends on the dimension $d$. So this depends on how big $n$ is and if we want more accuracy then $n$ becomes bigger. We have added some clarification. If function evaluation from an expensive simulation takes 5 minutes for one value, then performing the FFT calculations is going to take less time for practical $n$, say up to a million.
- Section 2
  - '$f$ is drawn from a Gaussian process': it should be explained what this means, e.g., does it mean $f$ is the Gaussian function with unknown mean and covariance, or does it behave like this in expectation, or ... This sounds like a very strong assumption. Some further explanation has been provided. The assumption may be strong, but it is common in this literature. The fact that we require some parameters to be inferred from data should make that assumption more palatable.
  - It should maybe be mentioned here that there are hyper distributions for the hyper parameters. Done.
  - It is never said how the $x_i$ are sampled. Does (5a) hold for any distribution of samples $x_i$? Yes, any distinct nodes. This has been mentioned explicitly
  - In this section there are different quantities being defined: the matrix $C_{\boldsymbol{\theta}}$ which involves evaluating the covariance function, $c_0$ which involves integrating the covariance function and $\boldsymbol{c}$ which involves calculating integrals of the covariance function with one leg fixed to a sample point. Only at the end the authors say that they assume these things to be cheap. I think this should be spelled out immediately. Done.
  - It think is not a good idea to leave out the dependence on $\boldsymbol{\theta}$ for $C$. E.g. in (11) the $\boldsymbol{\theta}$ is not visible at all in the expression to be minimized for it. The $C$ without subscript also clashes that of lemma 1. $\boldsymbol{\theta}$ has been added as much as possible, and particularly into the equations where there is optimization. However, $\boldsymbol{\theta}$ has been omitted in places where it would clutter the derivation.

  - The lead in to lemma 1 makes it feel like this is a field specific lemma but in fact it is just the well known conditional Gaussian. Please write something like 'We need the following lemma on conditional normal distribution ...'. Done.
  - Do not use $C$ in this lemma. Notation has been changed.
  - The other derivations in the paper are all made in prose. It would be better to formally state some of them. E.g., the results on the top of page 3. The key results from Sections 2.2.1–2.2.3 have been collected in a theorem just before Section 2.2.1. The key results of Section 3 are also collected into a theorem. Theorem 3 now summarizes the results of Section 4.
  - From (6) to (7) the notation is different: $c^T C^{-1} 1$ vs $1^T C^{-1} c$, this might be corrected. Done.
- Section 2.2.1
  - It would help to see the dependence on $\boldsymbol{\theta}$ in (9), (10) and (11). We have added the dependence on $\boldsymbol{\theta}$ through the end of Section 2. However, we have omitted it later where it would make the derivations hard to follow.
  - It looks like (11) is independent of $m_{\text{MLE}}$ and $s^2_{\text{MLE}}$, how does it pop up? Otherwise we could just first optimize for $\boldsymbol{\theta}$, then for m and then for $s^2$. No, the optimization problem for $\boldsymbol{\theta}$ already has $m_{\text{MLE}}$ and $s^2_{\text{MLE}}$ inserted into the log-likelihood.
- Section 2.2.2
  - I cannot follow the derivation. The notation has been clarified. This is a tedious derivation and has been moved to the Appendix.
  - At the end of the section the authors say that this is not a good method. Why then list it? Maybe the final paragraph needs to be rewritten. We only cast doubt on the full Bayes approach for estimating $\boldsymbol{\theta}$, not for dealing with the other hyperparameters.
- Section 2.3. The authors stress here that the integrand must be a 'typical' draw from the assumed Gaussian process and not an outlier. That sounds scary. It may sound scary, but that is the common implicit assumption of probabilistic numerics approaches, such as this one. Other traditional methods for estimating the error of a cubature rule either assume that the integrand lies in the ball of some function space (how do you know the radius of that ball?) or estimate the error heuristically.
- Section 2.4. Please use $d'$ for the dimension of the original integral in (22) and immediately say that the dimension for the computation is $d = d' - 1$, 'as will now be explained ...'. So put $d'$ after the display (22) and in the display (23) and update the formula for $\mu$ at the bottom of the page and the display on the next page. After these changes you can keep the '$d = 2$' in the caption of Fig 2. Excellent suggestion. Done.
- Section 3
  - Maybe write 'previous section' instead of 'last section'. Done.
  - Maybe write 'covariance kernel' instead of just 'kernel'. I think the word kernel hasn't been used up till this point. Done.
- Section 3.1. Use $C_{\boldsymbol{\theta}}$ in (27) for clarity. Done.
- Section 3.3. Elaborate on what is meant with the second sentence. This section re-written and eliminates this sentence.
- Sections 4.2, 4.3 and 4.4. The definition (35) makes it that the covariance matrix $C$ is a permutation of a circulant matrix. .... These subsections have been substantially re-written in response to the referee's suggestion.
- Section 4.5.
  - The authors present an alternative form of calculating the matrix $C$ for covariance functions of the particular product form in Section 4.2 ... I think they are right that this is a good way of calculating Ccirc without doing the -1 at the end. However, it is not clear whether this will actually help. ... It is not clear to me if operating the FFT on the first (permuted) column of $\mathring{C}$ in this form is not going to increase the error on $\mathring{\lambda}_1$, so putting the numerical error elsewhere. A key quantity can lose accuracy due to catastrophic cancellation. This has been observed by us in practice. The proposed method avoids this. We do not see how we are introducing significantly more numerical error elsewhere. We are basically pulling the constant matrix one out of the Gram matrix. Then the FFT of the first column is not essentially $n$. We want the difference from $n$.

  - Please delete the word 'periodic' in proposition 1, this is covered by the shift-invariantness. Done.
- Section 5.1. The Baker's transform is introduced first and then the authors talk about more general transforms without saying so. For the Baker's transform the substitution written with $\Psi'$ is undefined. This paragraph should be restructured. Done.

## Reviewer 2: The authors study a stopping criterion ...

There is no theory available that covers the whole process (estimation and cubature with credible interval); instead, the authors present numerical experiments to demonstrate the performance of three variants of their new algorithm. We would disagree with the sweeping nature of the claim that "there is no theory". Our method is taking the established theory of Bayesian numerical analysis and using it to develop adaptive stopping criteria. In our revised manuscript we summarize our results in three theorems. We recognize some gaps in the theory.

- p. 3, Sec. 2.2 The application of Lemma 1 to automatic Bayesian cubature is not clear to me, since integrand data is accumulated sequentially, which does not correspond to applying a single linear mapping to all integrands (realizations) $f$. This is a valid criticism, which we address in Section 2.3.
- Sec. 2.2.1, 2.2.2 I suggest to include some references for the results that are presented in Sections 2.2.1 and 2.2.2. Our knowledge of the literature is limited, but much of what we read assumes a zero mean for the Gaussian process, which simplifies the derivations and leads to somewhat different results. We would welcome sources to cite that consider our setting.
- p. 5, Sec. 2.4 Why are Sobol points chosen for the Matern kernel? Is there some theoretical link? Added "Sobol' points are a typical space-filling design."
- In the example the integral (22) is considered over a compact set. What is the benefit of Genz's transformation, compared to a simple affine linear transformation, in this case? A sentence has been added. Genz's transformation is state-of-the-art.
- p. 6 Why Hermitian? Are you considering complex-valued kernels? It is convenient since we are working with FFT's of real vectors, which are complex vectors. Some explanation is provided at the beginning of Section 3.1.
- p. 7, top left Definition of $v_1^*$? ... where ... (lower case) Clarified. Fixed.
- Sec. 3.2–3.4 Does the operation count also include the computational cost for the minimization problem to estimate $\theta$? I suspect that this is not the case, since there is no assumption concerning the dependence of $C_\theta$ on $\theta$. The same question arises in the setting of Section 4. You are correct. This has been clarified.
- Sec. 4.1 and 4.2 As the authors state, larger $r$ implies a greater degree of smoothness of the kernel. This should lead to a greater amount of smoothness of the integrands, and thus to a fast convergence of suitable cubature formulas. It might be that this is not exploited by the integration lattices (plus periodization), as studied in the present paper. Please comment. Greater smoothness of the covariance kernel implies an *assumption* of greater smoothness of the integrands, which may or may not be justified. Lattice rules work well for all degrees of smoothness. This point is mentioned at the end of Section 4.2.
- p. 9, eqn. (37) $i, j = 1$. Done.
- p. 9, bottom, right Assumption (25c) ... The rewriting of this section has eliminated that sentence.
- p. 12, center left ... differs depends ... ? Fixed.
- Sec. 5.1 Is there some kind of a universal periodization that works for every value of $r$? One could choose an infinitely smooth periodization, but this would change the integrand to potentially make its implied covariance scaling, $s$, large. This is mentioned at the end of Section 5.1.
- Sec. 5.2 You take $r = 2$ or $r = 1$ for the periodization in these examples. Does this mean that $\theta = (2, \gamma)$ is considered for the shift-invariant kernel, or do you still consider $\theta = (r, \gamma)$ with $r \in \mathbb{N}$ as the unknown parameter? For technical simplicity we fix $r$. We mention that in Section 5.2.
- I suggest not to use $r$ also in the drift term in the option pricing example. Excellent suggestion. Replaced by $R$.

## Reviewer 3

- Section 2.1, line 9: I think "The function $C$" should be "The function $C_\theta$," since there is no function $C$. Yes, corrected.
- Right before (12): Not sure if "simplify to" is the right expression, since it does not look like a simplification of those formulas. This part has been re-written.
- Section 2.2.3, line 5: This is conditional on the data, but also a function of the parameters, I suppose? I think this should be clarified. Done.
- On line 7, read "is the sum". Fixed.
- Two lines before Equation (20): "confidence interval" and not "credibility interval" like the other ones? Fixed.
- Five lines after (21): $\text{err}_{\text{CI}}$ refers to what exactly? The value in (20)? Clarify. Re-written to clarify.
- In (24), are $C_1, \ldots, C_n$ the columns of C? Yes.
- It took me a while to figure out what the four different colors of the points in Figures 5 to 13 mean. There is a continuum of colors on the right, but only four specific colors for the points are used, so I think it would be useful to say explicitly what each of these four colors represent. In the text on page 12, it would be useful to say for example that $\epsilon = 10^{-5}$ corresponds to the green points. By the way, are there 100 green points, for the 100 independent shifts? We have modified our experiments to randomly choose from a continuum of error tolerances. We have also explained that each experiment correspondes to a different random shift.
- Section 5.2, line 8: correct "differs depends". Fixed.
- Page 14: The three figures look all the same! Any comment on that? They are similar, but not the same.
- In this example, does it really make sense to assume that the integrand comes from a Gaussian process? Theoretically, perhaps not, but with a nonzero mean, maybe this is not too much of a stretch. Bayesian cubature assuming non-Gaussian processes is much more challenging. We have addressed this at the end of Section 5.2
- Four lines below Figure 13: "A noticeable aspect from the plots is how much ....": It may be good to explain exactly how we see that in the plots. We have elaborated more.
- Reference [1]: "Griolami"? Mark may not recognize himself! Sorry. Fixed.

## Editor: We have received two very careful reviews of this paper ...

The revision should at a minimum point out the connection to the FFT. Done.

Because the goal is to reach people outside of PN some additional explanations are in order. We have tried.

1. The results in Section 2.2.2 were confusing. ... This has been re-written.
2. It is weird to use complex numbers for the eigenvectors of a real symmetric matrix ... While it may be strange, it is convenient since we are working with FFT's of real vectors, which are complex vectors. Some explanation is provided at the beginning of Section 3.1.
3. In(24), why is $V^H = nV^{-1}$ and not just $V^{-1}$? ... This normalization allows $\boldsymbol{V}_1 = \mathbf{1}$.
4. The argument at the end of section 4.3 involving is very slippery. ... Section 4 has been substantially re-written.

Some minor points

1. The $x_i$ in (4) would have to be distinct to force a strictly positive quadratic form. Clarified.
2. The notation for the full Bayes treatment is not consistent. ... We have tried to fix this.
3. First paragraph of 2.3: for the $\mu \to$ for $\mu$. Also intervals $\to$ interval. Fixed.
4. End of Section 2: the the Matérn $\to$ the Matérn Fixed.
5. In the big display near the middle of the left column on page 7 should $c^T$ be $c^H$ ? Otherwise how do we get $\tilde{c}_i^*$? It is correct as stands. A sentence of explanation has been added just before those equations.
6. In equation (30), I think t should have $n-1$ df, not $n_{j-1}$ df. Otherwise what is $n_j$? Corrected.
7. 1st sentence of 4.3: shift-invariance $\to$ shift-invariant Corrected.