Answer to the referees concerning the review of "Random time step probabilistic methods for uncertainty quantification in chaotic and geometric numerical integration"

Assyr Abdulle   Giacomo Garegnani

We thank the reviewers for their numerous comments and we are confident that clarifying certain aspects will help us to consistently ameliorate our work. We divided our answer in three sections. In Section 1, we discuss the modifications of Section 7 (Hamiltonian systems), since both referees had multiple comments about this part of our work. In Section 2 and Section 3 we answer separately to the comments of Reviewer #1 and #2, respectively, about the other topics treated by our work.

Some comments are not directly addressed in this report (e.g., language/syntax issues, typos, slight notation changes) but have nonetheless been considered in the rewriting of our paper.

1 Modifications of Section 7

Both Reviewer #1 and Reviewer #2 have multiple comments regarding symplecticity and the long-time conservation of Hamiltonians. Since we believe that this section may be the most relevant of our work, and that it is for sure the one presenting the most tedious calculations, we decided to let it undergo a thorough revision process and to dedicate it a separate section in this response. In the following, we highlight the major modifications.

1. **Answer to the following comment by Reviewer #1**

   The authors should demonstrate that the decomposition that occurs over (84), (85), and (86) is correct by showing the intermediate decompositions. For example, the decomposition could be formulated as a lemma using arbitrary quantities, e.g. a_j instead of η_j and $b_{j,k}$ instead of $(H_j^k - h^k)\Delta_{j,k}$, so that the reader can see and verify the correctness of the decomposition, without having to do so in terms of the visually more complicated $(H_j^k - h^k)\Delta_{j,k}$. The estimates in (87) and (88) should also be given in separate lemmas, each with their own proofs.

   The proof of Theorem 6 has been shortened to make it easier for the reader to follow our reasoning and the calculations. This is achieved by the introduction of two lemmas (Lemma 7 and Lemma 8), where the most technical steps are performed. In particular, Lemma 7 deals with the decomposition of the sum, and Lemma 8 with the bound of the remainder terms. Since the proofs of these two results involve tedious calculations, we moved them to the Appendix, so that an interested reader can still verify their rightfulness. We thank Reviewer #1 for having pointed out this lack of clarity, as we believe that now our manuscript is dealing with these technical steps in a much clearer manner.

2. **Answer to the following comments by Reviewer #2**

   After Eq. (63): is Q required to be a polynomial or ∇Q?
   Try to given an outline of the argument where κ comes from.
   Can you give a specific reference that explains the existence of $\widetilde{Q}$ in (70)?

   Concerning the first question, neither Q nor ∇Q are required to be polynomials, but assumptions on the smoothness of Q are needed for the analysis of long-time conservation to hold (Assumption

5). Methods for conserving *exactly* polynomial first integrals can be designed, and we explain the properties of their RTS-RK counterpart in Section 6. In Section 7 our goal is to observe the *good approximation* of Hamiltonian functions provided by symplectic integrators.

Concerning the second and third questions, we believe that these points are clearly exposed in [3, Chapter IX]. We added precise citations in the text in order to clarify this.

3. **Answer to the following comment by Reviewer #1**

> H_k is selected via a random mapping $\tau(y, h) = \tau(h) = h\Theta_k$, where Θ_k are opportunely scaled ...' – Do the authors mean that $H_k = \tau(y, h) = h\Theta_k$? If so, they should explicitly state this. I did not find 'is selected via' to be sufficiently clear.

and to the following comments by Reviewer #2

> Maybe hint at the argument why $\tau(y, h)$ exists and why Θ_k exists without violating Assumption 1 .
> The authors reference the paper by Skeel & Gear. Given the main reference through the rest of the text is [3], maybe the authors could use the notation thereof and point to specific sections of [3].
> How can I see from (66) that (65) is satisfied?

The mapping $\tau(y, h)$ and Θ_k trivially exist by defining $\Theta_k := H_k/h$, and $\tau(y, h) = h\Theta_k$. The fact we would like to highlight with Lemma 5 is that the numerical flow $\Psi_{\tau(y,h)}$ conserves the property of symplecticity when the choice of h is independent of y. This is not the case when h is chosen via standard adaptive techniques, which pick time steps as a function of error estimators, which are based on the solution. In our case, we have $\partial_y \tau(y, h) = 0$ and therefore (65) (with Definition 4) is trivially satisfied. Finally, we added a reference to [3, Section VIII.1], where these issues are treated thoroughly. We believe that with the modifications we made, the proof of Lemma 5 should be clear for a reader of Statistics and Computing.

4. **Answer to the following comment by Reviewer #2**

> Lemma 6: I found the arguments of the proof helpful for following the text. Consider putting it in the main body.

While we believe that building a geometry-aware probabilistic integrator may be the central result of our work, we would not want the reader to be distracted from the main messages by tedious computations. That is the reasoning behind our choice of placing the proof of Lemma 6 (and of the new Lemma 7 and Lemma 8) in the Appendix. Therefore, we decided to keep them in the Appendix in this revised version, and hope that this is satisfactory for both the referees.

5. **Answer to the following comment by Reviewer #1**

> Do the authors intend for (74) to be the definition of the η_j's? If so, then they should explicitly state this, reformulate (74), and use the ':=' notation if possible to indicate that a definition is being made, instead of an assertion of equality between two objects. – The authors should state the statistical properties of the η_j's, e.g. whether or not they are independent with respect to certain random variables, whether or not they are identically distributed, etc. They should provide justifications for these assertions. Doing so provides a useful service to readers who wish to verify that the proof of Theorem 6 is correct. – It is not clear why the η_j's satisfy $|\eta_j| \leq CH_j e^{-\kappa/H_j}$ almost surely. Provide an explanation.

and to the following comment by Reviewer #2

> How can I see that (74) can be written thusly?

Equation (74) appears now reversed in order to make it a proper definition of the variables η_j. Regarding the almost sure bound $|\eta_j| \leq CH_j e^{-\kappa/H_j}$, we found a mistake in equation (71), which in its previous version stated

$$|Q(\Psi_h(y)) - Q(y)| \leq Che^{-\kappa/h}.$$

The inequality above is indeed not correct, but it holds for the modified Hamiltonian as

$$|\widetilde{Q}(\Psi_h(y)) - \widetilde{Q}(y)| \leq Che^{-\kappa/h}.$$

For a reference of the latter, see [3, Theorem IX.7.6] and the proof of [3, Theorem IX.8.1]. In particular, this implies by definition of the modified Hamiltonian that for any $y \in \mathbb{R}^{2d}$ we have

$$\left| Q(y) - Q(\Psi_h(y)) + \sum_{k=q}^{N-1} h^k \big(Q_{k+1}(y) - Q_{k+1}(\Psi_h(y))\big) \right| \leq Che^{-\kappa/h}.$$

Replacing h with H_j, one gets the almost sure bound $|\eta_j| \leq CH_j e^{-\kappa/H_j}$. Indeed, the difference between the modified Hamiltonian $\widetilde{Q}$ and the local modified Hamiltonians $\widehat{Q}_j$ is their global vs. local nature. Hence, on a local level the two quantities behave in the same manner. In order to clarify this point, we added in the text a reference to (71) when stating the bound on the (random) local truncation error η_j.

Regarding the statistical properties of the η_j's, not much can be stated. In fact, we have that η_j and η_i for $i \neq j$, are not independent random variables. In fact, when following a trajectory, the variable η_j depends on Y_j (respectively η_i and Y_i), and since Y_i and Y_j are not independent, the variables η_j and η_i are not independent. Nonetheless, the almost sure bounds $|\eta_j| \leq CH_j e^{-\kappa/H_j}$ form a sequence of independent random variables, as the constant C is independent of y (see [3, Theorem IX.7.6]) and the random time steps are independent under Assumption 1. This means that we can compute, for example,

$$\mathbb{E}\,\eta_i \eta_j \leq C\,\mathbb{E}\,H_j e^{-\kappa/H_j} H_i e^{-\kappa/H_i} = C\,\mathbb{E}\,H_j e^{-\kappa/H_j}\,\mathbb{E}\,H_i e^{-\kappa/H_i}, \tag{1}$$

and then use the result provided by Lemma 6 to proceed with calculations. We employ this property in the proof of Lemma 8.

6. **Answer to the following comments by Reviewer #1**

   "Moreover, for any $r, s > 1$ such that $r + s < R$" $\rightarrow$ *"Then for any $r, s > 1$ such that $r + s < R$." This hypothesis should be stated explicitly. In addition, since the authors have not explicitly proof that value of C may change from line to line, inequality is strictly speaking false. problem can resolved by adding a factor 2 or stating at beginning proof.*

We updated our text to clarify this point, with a particular care on the value of the constants, which, in fact, were not correct in the previous version.

7. **Answer to the following comments by Reviewer #1**

   "...if the numerical solution y_n ...is close enough to the initial condition" [...] The constraint that the numerical solution "is close enough to the initial condition" is vague. Theorem IX.8.1. of reference [3] is not very helpful: it just specifies some compact subset K of the domain of analyticity without even saying what K is. The authors should provide a more informative and specific constraint on the numerical solution.

Rigorous arguments on the domain of validity of results such as Theorem 6 or [3, Theorem IX.8.1] ought to be found, for example, in [1]. We added this reference, but we believe that entering discussions in this direction would be out of the scope of our work.

8. **Answer to the following comments by Reviewer #1**

   Second (90): for true, we need $t_n \geq 1$
   First (93): two terms inside parentheses been justified before their appearance here. are requested state justify bounds used.
   Last help reader verify correct, write they use $t_n = nh$. an explicit bound $\sqrt{r}$ helps understand why $\mathcal{O}(e^{-\kappa(4mh)})$
   Why are the first two terms on the right-hand side of the first inequality of (94) bounded by $C_4 h^q$?

and to the following comment by Reviewer #2

I have not fully understood how this imposing of this terms should work in Eqs. (91) and (95).

We believe that the new version of the proof of Theorem 6, together with the proofs of Lemma 7 and Lemma 8, answer all the above questions.

9. **Answer to the following comments by Reviewer #1**

Remark 9: – From the point of view of probability theory, I do not think it is appropriate to refer to $p \to \infty$ as a 'deterministic limit'. It may be better to just write 'in the limit as $p \to \infty$'. – Explain in more detail why the coefficient M in Assumption 6 tends to 1 as $p \to \infty$.

We rephrased the Remark in light of this comment. In particular, the value M does not *tend* to 1, but it can be chosen arbitrarily close to 1 if $p \to \infty$.

10. **Answer to the following comments by Reviewer #1**

Remark 10: To a reader unfamiliar with symplectic integrators, it is not immediately clear what is the significance of this remark with respect to the preceding results in this paper. If a connection exists, then the authors should state and explain it clearly; otherwise the remark does not serve any meaningful purpose and ought to be removed.

and to the following comment by Reviewer #2

I have not been able to build up an intuition about Remark 10 in light of Theorem 6. How do these two statements relate to each other? What exactly am I to make of this as an user of your method?

We believe that this issue is clarified with the addition to Remark 10 of the sentence

Conversely, Theorem 6 proves that random step sizes do not spoil, under the assumptions specified above, the good long time properties of symplectic integrators with fixed step size.

11. **Answer to the following comment by Reviewer #1**

Remark 11: 'We introduce ... the remainder $\widehat{S}_1$.' – In the proof, the assumption is also used to simplify the terms in $\widehat{S}_2$.

For clarity, since $p \geq 3/2$ is now one assumption behind Lemma 8 and since we now avoided the splitting of S into $\widehat{S}_1$ and $\widehat{S}_2$, we rephrased this sentence as

As it can be noticed in the proof of Lemma 8, we introduce the assumption $p \geq 3/2$ in order to simplify the terms composing the remainder $S(\Delta, \eta)$.

2   Reviewer #1

1.   *"An additive random term could force the solution on the negative plane with a non-zero probability, which can become significantly big in case the magnitude of one component is small." – This sentence is unclear, because it is not clear what can become significantly big, and it is not precisely clear what the component is of.*

We modified the sentence above as

In particular, an additive random term could force the solution on the negative plane with a non-zero probability, and this probability could become non-negligibly big in case the magnitude of one component of the solution is small

2. *Definition 1: "... for any function Φ ... with all derivatives bounded uniformly on $\mathbb{R}^d$."*
   – From the point of view of probability theory, the class of "test functions" that one considers is the class of bounded, continuous functions. To avoid any confusion, the authors should state very clearly whether or not the test functions that they consider are bounded. I consider "all derivatives" to be imprecise, as certain readers may take this to mean that the test functions themselves are bounded, when in fact the authors consider unbounded test functions later, e.g. $\Phi(x) := x^\top x$. – For brevity and simplicity, perhaps the authors could define the appropriate function class in an equation and use a symbol to refer to the function class afterwards, instead of repeatedly writing "... with all derivatives bounded uniformly on $\mathbb{R}^d$".

   We introduced the symbol $\mathcal{C}_b^\infty(\mathbb{R}^d, \mathbb{R})$ for the functions in $\mathcal{C}^\infty(\mathbb{R}^d, \mathbb{R})$ with all derivatives bounded uniformly in $\mathbb{R}^d$. The function Φ that we consider in the numerical experiment is indeed unbounded on $\mathbb{R}^2$ ($d = 2$ being the dimension of the FitzHugh–Nagumo system). Nonetheless, the FitzHugh–Nagumo system is known to present periodic solutions lying in a bounded attractor $\mathcal{A}$, and if the numerical solver is stable the numerical solution will stay in a neighbourhood of this attractor, too. Therefore, the restriction of the function $\Phi \colon \mathcal{A} \to \mathbb{R}$, $x \mapsto x^\top x$ on the attractor is a valid test function.

3. *"and denote in the following ... the Markov generator on the step size h." I am aware that Conrad et al. use $\mathcal{L}^h$ for the infinitesimal generator of the Markov chain. However, I think this notation can be improved for the present article, for the following reasons: 1) $\mathcal{L}$ has already been used for the Lie derivative of the flow, and $\mathcal{L}$ and $\mathcal{L}^h$ are not related by taking "powers"; 2) I think the infinitesimal generator should depend on the distribution H_0 from which the random time steps are drawn, and this dependence is just as important as the dependence on the deterministic step size h. If the authors wish to emphasise the dependence of the generator on h, then I think they should also emphasise its H_0-dependence as well, at least for the first instance of the infinitesimal generator. If the authors wish to do so, they can can omit the H_0 dependence in subsequent instances to simplify notation, but I think it is advantageous to the authors to remind the reader that the present setting is different from that of Conrad et al.*

   In order to answer to the first part of this comment, we remark that the semi-group notation for the infinitesimal generator $e^{h\mathcal{L}}$ involves the Lie derivative of the flow $\mathcal{L}$, which in fact resolves this notation clash. We added a reference to [5, Section 4.3], where this topic is treated extensively. For the problem of "taking powers", we agree that the notation $\mathcal{L}^h$ could be misleading, and we therefore changed it to $\mathcal{L}_h$, as for the infinitesimal generator of the Markov chain $\mathcal{P}_h$. Conversely to the operator $\mathcal{L} = f \cdot \nabla$, the expression of $\mathcal{L}_h$ cannot be written explicitly, and the exponential notation $\mathcal{P}_h = e^{h\mathcal{L}_h}$ is employed for analogy with the exact solution.

   We agree that the infinitesimal generator of the Markov chain depends on the distribution on the step sizes, as the infinitesimal generator of the probabilistic method defined in [2] depends on the distribution of the additive random variables ξ_k. Nonetheless, the index h of $\mathcal{P}_h$ and $\mathcal{L}_h$ is chosen because these operators map an approximation of the solution at time t to an approximation of the solution at time $t + h$. In fact, even though our method proceeds by random steps, the value Y_k is still conceived as an approximation of $y(kh)$ (see in Section 2, just after (6), "where Y_k is still a random variable approximating $y(t_k)$ ..." and/or Remark 5). Therefore we believe that writing explicitly the dependence on H_0 is not only unnecessary, but even not entirely correct.

4. *"hence given $h > 0$ there exists an operator $\mathcal{P}_h$ such that " – As a service to the reader, provide a reference to a theorem or proposition that justifies the use of "hence". I do not expect all readers of this article to be familiar with homogeneous Markov chains.*

   and

   Also, as a service to the reader, provide a reference to a theorem or proposition that justifies (15). I do not expect all readers of this article to be familiar with homogeneous Markov chains.

We added a reference to [4, Section 2.3], where the theory of infinitesimal generator for homogeneous Markov chains is treated extensively.

5. *The authors should provide detailed, step-by-step, rigorously justified steps that proceed from (28) to the first inequality in (29). Do not use vague phrases such as 'Proceeding iteratively'. Note that 'proceeding iteratively' does not yield the first inequality in (29), even after using $w_0 = W_0$. Instead, it yields*

$$\sup_{u \in \mathbb{R}^d} |W_k(u) - w_k(u)| \leq C h^{min\{2p+1, q+1\}} \sum_{j=0}^{k-1} (1 + Lh)^j.$$

In fact, using Gronwall's inequality for nonnegative sequences yields

$$\sup_{u \in \mathbb{R}^d} |W_k(u) - w_k(u)| \leq C h^{min\{2p+1, q+1\}} \exp\left(k(1 + Lh)\right).$$

Since the upper bound of k is $N = T/h$, the exponential will increase to infinity faster than any power of h.

We thank the referee for pointing this imprecision out. We modified the result, that now we believe being correct. Moreover, let us remark that since $1 + x \leq e^x$, the "1+" term disappears in the exponential and we have

$$\sum_{j=0}^{k-1} (1 + Lh)^j \leq \sum_{j=0}^{k-1} e^{Lhj} \leq \sum_{j=0}^{k-1} e^{LT} \leq k e^{LT} \leq h^{-1} T e^{LT},$$

which yields

$$\sup_{u \in \mathbb{R}^d} |W_k(u) - w_k(u)| \leq C h^{\min\{2p, q\}} T e^{LT}.$$

6. *In Assumption 4, if $\mathcal{L}^h$ depends on the distribution H_0 (which I think it does), then this assumption should also state something about the distribution H_0, not just the vector field f. At the very least, I would expect "The function f is such that ..." to be changed to "The function f and the distribution H_0 are such that ...". Also, although it may be obvious to the authors, I think it should be stated explicitly as a service to the reader that the positive constant L may depend on f and H_0, but not on Φ or h.*

We modified Assumption 4 as follows

The function f and the distribution of the random time steps H_k, $k = 0, 1, \ldots$, are such that the operator $e^{h\mathcal{L}_h}$ satisfies for all functions $\Phi \in \mathcal{C}_b^\infty(\mathbb{R}^d, \mathbb{R})$ and a positive constant L,

$$\sup_{u \in \mathbb{R}^d} |e^{h\mathcal{L}_h} \Phi(u)| \leq (1 + Lh) \sup_{u \in \mathbb{R}^d} |\Phi(u)|, \tag{2}$$

where L may depend on f and on the distribution of the random time steps, but not on Φ or h.

7. *"In the sub-optimal case ... are balanced." – The authors could provide an explicit formula for M in terms of h, p and q in the case that p "if for any"*

A new formula has been added for treating this sub-optimal case.

8. *The expression of the Hellinger distance is incorrect, because the integral on the right-hand side is not taken with respect to the prior. See Definition 6.35 in Stuart's Acta Numerica paper.*

We adopted this expression for the Hellinger distance since we assumed that prior and posterior distributions admit a probability density function with respect to the Lebesgue measure. In this case, the Hellinger distance can be simplified and its expression is the one we provide. Since the section regarding Bayesian inverse problems delivers qualitative results, later confirmed by simple experiments, we decided to consider the case of finite-dimensional parameters. In this case, we feel that assuming that prior and posterior admit a probability density function is not limiting our analysis.

9. *"the standard random walk Metropolis–Hastings is usually employed" – Provide evidence for this by citing some literature.*

We corrected the sentence above with "the standard random walk Metropolis–Hastings can be employed". We believe that the standard implementation of Metropolis–Hastings is nowadays a common sampling tool, and that a list of references of papers employing this algorithm would add no value to our work.

10. *To make it explicitly clear how this section is connected to the preceding material (in particular, the introduction to Bayesian inverse problems), I recommend that the authors do the following: 1) Write 'Z = $\varphi_h(y_0^*) + \epsilon$' instead of 'd = $\varphi_h(y_0^*) + \epsilon$'. Upper case letters should be used for random variables and lower case letters should be used for deterministic values. 2) Specify the parameter space Θ, observable, and linear forward operator. 3) As far as possible, use the same symbols as used in (99), (100), (101).*

We modified the notation of this example by replacing $y_0 \leftarrow \vartheta$, $d \leftarrow Z$, $\bar{y}_0 \leftarrow \bar{\vartheta}$, $y_0^* \leftarrow \vartheta^*$ in order to get more uniformity with respect to the introduction to Bayesian inverse problems. Moreover, we added the sentence "In this case, the parameter space is $\Theta = \mathbb{R}$ and the forward operator $\mathcal{G}$ is defined by $\mathcal{G} \colon \mathbb{R} \to \mathbb{R}$, $\mathcal{G} \colon \vartheta \mapsto \vartheta e^{-h}$".

11. *Section 8.1: The symbol '→' and the phrase 'tends to' are used to suggest convergence. However, the authors do not specify the type of convergence. Thus, the statements should be considered only at an imprecise, heuristic level, and not at the level of rigorous mathematics. The authors should notify the reader of this.*

We added the sentence "In the following, we verify heuristically the convergence of the posterior distributions obtained with deterministic and probabilistic integrators with respect to a vanishing noise scale."

12. *It is difficult to distinguish the curves corresponding to different values of σ. I recommend that the authors reduce the range of values of y_0 in each subplot to make it easier for the reader to distinguish the curves for different values of σ*

We increased the sizes of the plots and "stretched" the horizontal axis to make the lines more distinguishable. Nonetheless, we think that keeping the same horizontal range for the four figures renders the idea of how more concentrated deterministic posteriors are. Since the additive noise posterior densities have non-negligible values over the range $\vartheta \in [-1, 3]$, we believe that decreasing the range more would reduce the visual and conceptual impact of this figure.

13. *Table 2: the sub-table for the explicit trapezoidal method should contain as many columns as the number of values of h that were used.*

Since for each one of the 5 values of p we consider 5 different values of h, this would create a table with 20 columns (n. values of $p \times$ (n. values of h - 1)). Therefore, we decided to write the average convergence slope over the set of step sizes for each value of p. We still believe that this choice renders our result more readable.

14. *Section 9: The authors should specify all the parameters that they use in their experiments. For example, in some subsections below the value of the parameter p is given, while in other subsections it is not.*

We reviewed the description of the numerical experiments and added the missing parameters.

15. *Some readers may assess that the variance of the Gaussian random variable is excessively small (on the order of 10^{-8}) relative to the initial condition. The authors should explain why they chose such a small variance.*

and

"Hence, initial conditions with a different energy level with respect to the observation are endowed with a high value of likelihood" – The authors should explicitly state the likelihood that they use and clearly explain how the energy level and the likelihood are connected.

We modified the introduction of the numerical experiment in order to clarify our choice of likelihood and the small noise as

> Noise is then set to be a Gaussian random variable $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2 I)$, where $\sigma_\varepsilon = 5 \cdot 10^{-4}$, and we fix a standard Gaussian prior on the initial condition, i.e., $\pi_0 = \mathcal{N}(0, I)$, so that the likelihood is given by (101). We choose the observational noise to have a small variance (i.e., of order $\mathcal{O}(10^{-8})$) as in this case classical solvers present the misleading overconfident behaviour explained in Section 8.

The observation about energy levels and likelihood is modified as

> Hence, initial conditions with a different energy level with respect to the observation are mapped by the approximate forward model to points which are close to the observations, and as a result the posterior distribution is concentrated far from the true value.

The meaning of the above sentence is that if the approximate forward model is not conserving the Hamiltonian, the energy state of the approximate solution at time t will be different than the one of the initial condition. Let us assume without loss of generality that the numerical integrator causes a positive drift in energy. In this case, an initial condition with a lower energy than the observations will be "seen" under the posterior as the most likely. Conversely, if the forward map conserves (i.e., approximately conserves) the Hamiltonian, *at least* the correctness on the energy for the solution of the inverse problem with respect to the observations can be trusted.

16. Regarding the proof of Lemma 6. *The authors need to bound the Mh term by a constant that does not depend on h.*

We added to the statement of the Theorem "and if $h \leq \bar{h} < \infty$" and bounded $Mh \leq M\bar{h}$. In the following results, we most often assume $\bar{h} = 1$.

17. *Proof of Lemma 6: The authors only consider the case where $H_j \geq h$. They should discuss the case where H_j "in the mean-square sense"*

The proof is actually valid for $H_j < h$ as well. We nonetheless modified it for clarity by introducing the notation

> In the following, we denote by $[\![a, b]\!]$ the interval $[\![a, b]\!] = [a, b]$ if $a < b$ and $[\![a, b]\!] = [b, a]$ if $a \geq b$.

The rest of the proof is unchanged, and we believe correct, by modifying the intervals $[h, H_j]$ to $[\![h, H_j]\!]$.

3 Reviewer #2

1. *Where did M^{-1} go from (54) to (55)? Cf. (51)?*

The Monte Carlo estimator is unbiased, therefore $\mathbb{E}(\widehat{Z}_{N,M} - Z) = \mathbb{E}(Z_N - Z)$ and then Theorem 1 can be applied consequently. We nonetheless explicitly highlighted this fact this in the revised version.

2. *System (122) might be easier to parse written as a second-order system.*

We agree with the reviewer on the fact that the Kepler system is oftentimes found in literature as a second-order system. Nonetheless, since our work treats first-order equations, we believe that if we first presented the Kepler system as a second-order system, we would have to write its first-order transformation too, which would make the numerical experiment heavier to digest for the reader.

References

[1] G. BENETTIN AND A. GIORGILLI, *On the Hamiltonian interpolation of near to the identity symplectic mappings with application to symplectic integration algorithms*, J. Statist. Phys., 74 (1994), pp. 1117–1143.

[2] P. R. CONRAD, M. GIROLAMI, S. SÄRKKÄ, A. STUART, AND K. ZYGALAKIS, *Statistical analysis of differential equations: introducing probability measures on numerical solutions*, Stat. Comput., 27 (2017), pp. 1065–1082.

[3] E. HAIRER, C. LUBICH, AND G. WANNER, *Geometric Numerical Integration. Structure-Preserving Algorithms for Ordinary Differential Equations*, Springer Series in Computational Mathematics 31, Springer-Verlag, Berlin, second ed., 2006.

[4] G. A. PAVLIOTIS, *Stochastic processes and applications: Diffusion processes, the Fokker-Planck and Langevin equations*, vol. 60 of Texts in Applied Mathematics, Springer, New York, 2014.

[5] G. A. PAVLIOTIS AND A. M. STUART, *Multiscale methods: Averaging and Homogenization*, vol. 53 of Texts in Applied Mathematics, Springer, New York, 2008.