Importance Gaussian Quadrature

Importance Gaussian Quadrature

Víctor Elvira, *Senior Member, IEEE*, Luca Martino, and Pau Closas, *Senior Member, IEEE*

***Abstract*—Importance sampling (IS) and numerical integration methods are usually employed for approximating moments of complicated target distributions. In its basic procedure, the IS methodology randomly draws samples from a proposal distribution and weights them accordingly, accounting for the mismatch between the target and proposal. In this work, we present a general framework of numerical integration techniques inspired by the IS methodology. The framework can also be seen as an incorporation of deterministic rules into IS methods, reducing the error of the estimators by several orders of magnitude in several problems of interest. The proposed approach extends the range of applicability of the Gaussian quadrature rules. For instance, the IS perspective allows us to use Gauss-Hermite rules in problems where the integrand is not involving a Gaussian distribution, and even more, when the integrand can only be evaluated up to a normalizing constant, as it is usually the case in Bayesian inference. The novel perspective makes use of recent advances on the multiple IS (MIS) and adaptive (AIS) literatures, and incorporates it to a wider numerical integration framework that combines several numerical integration rules that can be iteratively adapted. We analyze the convergence of the algorithms and provide some representative examples showing the superiority of the proposed approach in terms of performance.**

Keywords*—Importance sampling, quadrature rules, numerical integration, Bayesian inference.

I. Introduction

The number of applications where it is required to approximate intractable integrals is countless. There is a plethora of approximate methods in the wide of literature in engineering, statistics, and mathematics. These methods are often divided into two main families: the numerical integration (deterministic) methods and the Monte Carlo (random) methods.

Gaussian quadrature is a family of numerical integration methods based on a deterministic (and optimal, in some sense) choice of weighted points (or nodes) [1].[1] The approximation is then constructed through a weighted linear combination (according to the weights) of a nonlinear transformation of the points. This non-linearity, as well as the choice of the nodes and weights, depend on the specific integral to solve. The nodes are deterministically chosen in order to minimize the error in the approximation, which explains the high performance when they can be applied. Thus, when their application is possible, the corresponding algorithms have become benchmark techniques in their fields. As an example in signal processing, Gauss-Hermite rules have been successfully applied in a variety of applications of stochastic filtering, often with remarkable performance [3]. Particularly, the Quadrature Kalman filter (QKF) [4], [5] and its variants for high-dimensional systems [6], [7] showed improved performance over simulation-based methods when the Gaussian assumption on noise statistics holds. QKF falls in the category of sigma-point Kalman filters, where other variants can be found depending on the deterministic rule used to select and weight the nodes. For instance, one encounters also the popular Unscented Kalman filter (UKF) [8] or the Cubature Kalman filter (CKF) [9], both requiring less computational complexity than QKF while degrading its performance in the presence of high nonlinearities [10]. Moreover, quadrature methods have also been applied in the static framework in a multitude of applications in physics, econometric, and statistics at large [11], [12], [13]. However, the application of these methodologies is generally limited to Gaussian noise perturbations in the assumed probabilistic model [14].

The second family is constituted by the Monte Carlo algorithms, where the nodes are generated randomly (i.e., they are samples) [15], [16]. Arguably, the two main Monte Carlo subfamilies are Markov chain Monte Carlo (MCMC) and importance sampling (IS), and both of them are often used to approximate integrals that involved a specific target distribution. In the former, a Markov chain is constructed in a way that its stationary distribution exists and coincides with the target distribution after a burn-in period. IS simulates samples from a simpler proposal distribution and weights them properly to perform integral approximations. IS provides valid estimators without requiring a burn-in period while enjoys of solid theoretical guarantees such as consistency of the estimators and explicit convergence rates, [17], [18]. Due to their advantages and limitations, in the literature several authors have proposed novel schemes attempting to merge the benefits of both previous families, e.g., including deterministic procedures within the Monte Carlo techniques. This is the case of quasi Monte Carlo methods [19] and variance reduction methods [17, Chapter 8].

Contributions. In this work, we propose a theoretically-grounded framework based on quadrature rules. The IS-based interpretation allow us to propose novel quadrature methods, and pave the way to more sophisticated adaptive mechanisms in very generic settings. We develop the framework by explicitly using the Gauss-Hermite rule (i.e., for Gaussian distributions), but our perspective can be applied to a much

V. Elvira is with the School of Mathematics at the University of Edinburgh (UK), e-mail: victor.elvira@ed.ac.uk; L. Martino is with the Universidad Rey Juan Carlos of Madrid (Spain); P. Closas is with the Department of Electrical and Computer Engineering, Northeastern University, Boston, MA (USA), e-mail: closas@ece.neu.edu. The work of Víctor Elvira was supported by Agence Nationale de la Recherche of France under PISCES project (ANR-17-CE40-0031-01). (Corresponding author: Víctor Elvira.) P. Closas was partially supported by the National Science Foundation under Awards CNS-1815349 and ECCS-1845833.

[1]The term numerical integration is often considered synonym of numerical quadrature, or simply quadrature. Some authors prefer to use the term quadrature for one-dimensional integrands, using the term cubature for higher dimensions [2]. For the sake of brevity, in this paper we will use the term quadrature indistinctly regardless of the dimension.

wider class of quadrature rules for integration in a variety of sets. The basic method on which we develop the framework is referred to as *importance Gauss-Hermite* (IGH) method. We propose a novel estimator, inspired by the self-normalized IS estimator, that can be used when the target distribution can be evaluated only up to a normalizing constant. IGH extends the applicability of the Gauss-Hermite rules to a more generic class of integrals which involve other non-Gaussian distributions. This is done by the introduction of the so-called *proposal* density, which is Gaussian in the case of IGH, in a similar manner to the proposal in IS. We also provide error bounds for the approximations of the integrals in IGH, a related discussion regarding the optimal choice of the proposal function, and a through discussion about the computational complexity. Once the IS perspective is introduced, other more sophisticated schemes can be employed, including the use of several proposal pdfs, as in multiple IS (MIS) [20], or the adaptation of the proposals as in adaptive IS (AIS) [18]. Recent works have deeply studied the MIS framework, showing for instance that many weighing schemes are possible when more than one proposal are available [20], [21]. We propose two novel IGH-based schemes with multiple proposals and discuss their performance both from a theoretical point of view and via numerical simulations. Next, we provide some guidelines for the selection and the adaptation of the proposals in IGH. In particular, we propose two novel simple and high-performance adaptive IGH algorithms that are theoretically justified. Due to our re-interpretation of quadrature rules from a statistical point of view, we propose statistically inspired mechanisms to adjust the complexity, and a novel metric (named ESS-IGH) for self-assessing the new importance quadrature methodology.

Connections to the literature. In [22], the change of measure is proposed in the context of Gauss-Hermite quadrature, using a single Gaussian distribution. This introduced measure, that here we call proposal under our statistical perspective, is set to the Laplace approximation. The paper considers a unimodal integrand and assumes the maximum to be known. The relation of this simple change of measure with importance sampling is only mentioned in [23], although the methodology is not developed. The change of measure is also compared with the Laplace approximation [24] (see also [25] for a recent application). In a recent paper [26], the authors apply a change of measure in a more restricted setup (similarly to [22]), in order to approximate the marginal likelihood with quadrature rules in the context of Gaussian processes. In summary, the methodological power of this change of measure, has not been sufficiently explored in the literature, neither the statistical interpretation of quadrature rules. For instance, the weighted nodes in quadrature methods bear interesting parallelism with importance sampling. A better understanding of these connection will allow in the future for further significant methodological advances.

Structure of the paper. The rest of the paper is organized as follows. In Section II we present the problems and briefly discuss importance sampling and numerical integration methods. In Section III, we introduce the importance quadrature framework, particularizing for the case of Gauss-Hermite rules, and introducing the basic IGH method. We discuss the theoretical properties, the choice of the proposal, the computational complexity, and we provide two toy examples and a final discussion where we propose a method for sparse-grids in higher dimensions, and a metric to self-assessed importance quadrature methods. Section IV generalizes the IGH for multiple proposals, and we propose two quadrature methods based on two different interpretations coming from the MIS literature. We also discuss the theoretical properties of the methods. Section V introduces and adaptive version of IGH, and a discussion about further extensions of the framework. In Section VI we present three numerical examples: 1) a challenging multimodal target; 2) a signal processing example for inferring the parameters of an exoplanetary system; and 3) a Bayesian machine learning problem for estimating hyperparameters in a Gaussian process (GP). Finally, we conclude the paper with some remarks in Section VII.

II. Problem statement and background

Let us first define a r.v. $\mathbf{X} \in \mathcal{D} \subseteq \mathbb{R}^{d_x}$ with a probability density function (pdf) $\widetilde{\pi}(\mathbf{x})$. In many applications, the interest lies in computing integrals of the form

$$I = \int_{\mathcal{D}} f(\mathbf{x})\widetilde{\pi}(\mathbf{x})d\mathbf{x}, \tag{1}$$

where f can be any integrable function of $\mathbf{x}$ with respect to $\widetilde{\pi}(\mathbf{x})$. Unfortunately, in many practical scenarios, we cannot obtain an analytical solution for Eq. (1) and approximated methods need to be used instead. An illustrative example is the case of Bayesian inference, where the observed data as $\mathbf{y} \in \mathbb{R}^{d_y}$ parametrize the posterior pdf of the unknown vector $\mathbf{x} \in \mathbb{R}^{d_x}$ which is defined as

$$\widetilde{\pi}(\mathbf{x}|\mathbf{y}) = \frac{\ell(\mathbf{y}|\mathbf{x})p_0(\mathbf{x})}{Z(\mathbf{y})} \propto \pi(\mathbf{x}|\mathbf{y}) = \ell(\mathbf{y}|\mathbf{x})p_0(\mathbf{x}), \tag{2}$$

where $\ell(\mathbf{y}|\mathbf{x})$ is the likelihood function, $p_0(\mathbf{x})$ is the prior pdf, and $Z(\mathbf{y})$ is the normalization factor. This example is even more complicated, since $Z(\mathbf{y})$ is also unknown, and then $\widetilde{\pi}(\mathbf{x}|\mathbf{y})$ can be evaluated only up to a normalizing constant. From now on, we remove the dependence on $\mathbf{y}$ to simplify the notation.

In the following, we review the basics of importance sampling (IS) and deterministic numerical integration with Gaussian distributions.

A. Importance sampling (IS)

The basic implementation of IS can be readily understood by first rewriting Eq. (1) as

$$\begin{aligned} I &= \int_{\mathcal{D}} f(\mathbf{x})\widetilde{\pi}(\mathbf{x})d\mathbf{x} \\ &= \int_{\mathcal{D}} \frac{f(\mathbf{x})\widetilde{\pi}(\mathbf{x})}{q(\mathbf{x})} q(\mathbf{x})d\mathbf{x}, \end{aligned} \tag{3}$$

where $q(\mathbf{x})$ is the so-called *proposal* pdf with non-zero value for all $\mathbf{x}$ where the integrand is non-zero. The integral in Eq. (3) can be approximated via IS by first simulating a set of N samples $\{\mathbf{x}_n\}_{n=1}^N$ from a proposal pdf, $q(\mathbf{x})$, with heavier

tails than $|f(\mathbf{x})|\pi(\mathbf{x})$. Then, each sample is associated an importance weight given by

$$w_n = \frac{\pi(\mathbf{x}_n)}{q(\mathbf{x}_n)}, \quad n = 1, \dots, N. \tag{4}$$

Finally, an unbiased and consistent estimator (with increasing N) can be built as

$$\widehat{I}_{\text{UIS}} = \frac{1}{NZ} \sum_{n=1}^{N} w_n f(\mathbf{x}_n), \tag{5}$$

which is often denoted as the *unnormalized* importance sampling (UIS) estimator. In many applications, Z is unknown and the UIS cannot be directly applied. Instead, using the same samples and weights, the integral in Eq. (1) can be approximated with the self-normalized IS (SNIS) estimator as

$$\widetilde{I}_{\text{SNIS}} = \sum_{n=1}^{N} \bar{w}_n f(\mathbf{x}_n), \tag{6}$$

where $\bar{w}_i = \frac{w_i}{\sum_{j=1}^{N} w_j}$ are the normalized weights. Note that the SNIS estimator can be obtained by plugging the unbiased estimate $\widehat{Z} = \frac{1}{N}\sum_{j=1}^{N} w_j$ instead of Z in Eq. (5) [15]. The variance of UIS and SNIS estimators is related to the discrepancy between $\pi(\mathbf{x})|f(\mathbf{x})|$ and $q(\mathbf{x})$, and hence adaptive schemes are usually implemented in order to iteratively improve the efficiency of the method [18].

B. Numerical Integration based on Gaussian quadrature

A vast literature in the numerical integration is available, and the specific rules and their justification go beyond the scope of this paper (see for instance in [1] a review of simple quadrature rules). Here we focus in Gaussian quadrature methods, where a set of weighted nodes are carefully chosen. Common Gaussian quadrature rules are the Gauss-Legendre quadrature for integrals in the bounded domain $[-1, 1]$ and the Gauss-Hermite (GH) quadrature for integrals involving Gaussian distributions. Moreover, other variants are available, including the Gauss-Kronrod quadrature and Gauss-Patterson quadrature. In multidimensional integration, many other rules exist as well, especially with the aim of avoiding an exponential growth of the number of points with the dimension (sparse quadrature rules) [17, Chapter 8]. Some of the most popular approach are the so-called product rule cubature, constructed by directly extending a quadrature rule [27], or the Smolyak cubature, which is known to be more efficient in the selection of points by exploiting sparsity [28]. The use of sparse grids in multi-dimensional examples allows for computationally efficient integration techniques [29]. For some further details, see Appendix E and Table III. In this work, for simplicity, we focus on the GH rule. However, all the schemes and concepts presented in this work can be easily extended to other Gaussian quadrature rules. Since we mainly focus on the GH rule, now we review methods that approximate integrals over Gaussian distributions. Let us consider the integral of the form

$$I = \int_{\mathcal{D}} h(\mathbf{x}) \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x} \,, \tag{7}$$

where $\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ represents a Gaussian pdf with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, and h is a (possibly non-linear) function of the unknown variable $\mathbf{x}$. This integral, which computes a specific moment of a Gaussian distribution $I = \mathbb{E}[h(\mathbf{x})]$, can be efficiently computed leveraging the aforementioned deterministic rules.

Those deterministic methods approximate the integrals with a set of weighted samples/points. We refer the interested reader to [30], [31]. More specifically, the set of deterministic samples and weights are defined as $\mathcal{S} = \{\mathbf{x}_n, v_n\}_{n=1}^{N}$. Here we focus on the Gauss-Hermite quadrature rules without loss of generality with the aim of being specific, although we point out that the choice of points and weights in $\mathcal{S}$ for approximating the integral in Eq. (1) is not unique. The resulting Gauss-Hermite estimator of the integral is given by

$$I \approx \widehat{I}_{\text{GH}} = \sum_{n=1}^{N} v_n h(\mathbf{x}_n). \tag{8}$$

In GH quadrature, the points $\mathbf{x}_n$ are roots of the Hermite polynomial, and the weights v_n are also function of such polynomial (see the last row of Table III for an explicit expression and Appendix E for more details). It is worth noting that in GH, $N = \alpha^{d_x}$ points are selected, where α corresponds to the number of unique values per dimension (i.e., the resulting points form a multidimensional grid). Therefore, the complexity grows exponentially with the dimension of $\mathbf{x}$ although this issue can be alleviated using other deterministic rules with lower complexity rates such as cubature or unscented rules (requiring $N = 2\alpha$ and $N = 2\alpha + 1$ nodes, respectively) [10].

In the case of quadrature rules, exact integration in Eq.(8) occurs when $h(\mathbf{x})$ is a polynomial of order less or equal than $2\alpha - 1$. Conversely, there is an integration error when the function has a degree higher than $2\alpha - 1$. For the unidimensional case, $d_x = 1$, the error associated to the Gauss-Hermite quadrature rule is related to the remainder of the Taylor expansion of $h(x)$ [32], [5]

$$e = \frac{\alpha! h^{(2\alpha)}(\varepsilon)}{(2\alpha)!}, \tag{9}$$

where $h^{(2\alpha)}(x)$ is the 2α-th derivative of $h(\cdot)$ and ε is in the *neighborhood* of x. This error analysis can be extended to the multidimensional case, considering that the restriction on the degree should apply per dimension. At this point, we would like to notice that (9) can be bounded as

$$e \leq \frac{\alpha! ||h^{(2\alpha)}||_\infty}{(2\alpha)!}, \tag{10}$$

where $||\cdot||_\infty$ is the supremum operator. Hence, for any $h(\cdot)$ where the supremum of the 2α-th derivative grows slower than $\frac{(2\alpha)!}{\alpha!}$, we can guarantee that the upper bound of the error decreases when we increase the number of quadrature points. Note that in all cases, reducing $||h^{(2\alpha)}||_\infty$, implies decreasing the upper bound of the error. In Appendix D, we provide a result showing that when α grows, then the bound on the error tends to zero such that $e \to 0$.

III. Importance Quadrature Schemes

In the following, we develop a novel quadrature-based framework that approximates the integral of Eq. (1) for a generic non-Gaussian distribution $\widetilde{\pi}(\mathbf{x}) = \frac{\pi(\mathbf{x})}{Z}$. To that end, we aim at applying deterministic integration rules under an importance sampling perspective by introducing one or several *proposal* densities. This connection between quadrature methods and IS allows us to develop further non-trivial extensions of quadrature methods, the extension to the case of multiple proposals, the extension of existing adaptive IS procedures, and the development of new adaptive methodologies. We recall that specific importance quadrature methods can be implemented depending on the integration domain $\mathcal{D}$. In the following, and without loss of generality, we focus on $\mathcal{D} = \mathbb{R}^{d_x}$ and on the Gauss-Hermite quadrature rules.

A. Basic importance Gauss-Hermite (IGH) method

Let us rewrite the targeted integral in Eq. (3) as

$$I = \int_{\mathcal{D}} h(\mathbf{x}) q(\mathbf{x}) d\mathbf{x}, \tag{11}$$

where

$$h(\mathbf{x}) = f(\mathbf{x}) \frac{\widetilde{\pi}(\mathbf{x})}{q(\mathbf{x})} \ , \tag{12}$$

and $q(\mathbf{x})$ is the introduced *proposal* pdf with $q(\mathbf{x}) > 0$ for all values where $f(\mathbf{x})\frac{\widetilde{\pi}(\mathbf{x})}{q(\mathbf{x})}$ is non-zero.[2] Note that this rearrangement is the same as the usual IS trick of Eq. (3). We now choose a Gaussian *proposal* $q(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$, which allows us to re-interpret I as the expectation of $h(\mathbf{x})$ under the distribution $q(\mathbf{x})$, as in Eq. (7). The weighted samples are deterministically chosen with the Gauss-Hermite rules discussed in Section II-B, reason why we called the method *importance Gauss-Hermite* (IGH) method. Following this double interpretation (from IS and quadrature perspectives), we have an extra degree of freedom in the choice of the parameters of the Gaussian *proposal* pdf $q(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Let us summarize the basic IGH method in Algorithm 1, which will serve as a basis for further extensions below. In Step 1, N deterministic points, $\{\mathbf{x}_n\}_{n=1}^N$, and their associated quadrature weights $\{v_n\}_{n=1}^N$, are chosen according to the Gauss-Hermite rule. In Step 2, we compute the importance weights according to the standard expression of Eq. (4). Interestingly, the IGH weights, $\{w'_n\}_{n=1}^N$, are computed as the product of the quadrature and the IS weight, in Eq. (14). Note that the weights are multiplied by a factor of N, so they can be used at the estimator of Z in Eq. (17). The unnormalized IGH estimator is given in Eq. (15) in Step 4 (only if Z is known) while the self-normalized estimator is given in (16) of Step 5.

B. Two toy examples

We present two illustrative toy examples that provide useful insights about the behavior of IGH and the importance of the choice of the proposal, motivating the next sections.

[2]Note that we use the terminology of IS for the proposal $q(\mathbf{x})$ although the samples are not simulated.

Algorithm 1 Basic Importance Gauss-Hermite (IGH) algorithm

Input: N, $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$

1: Select N points $\mathbf{x}_n$ and the associated quadrature weights v_n, for $n = 1, \ldots, N$, considering a Gaussian pdf $q(\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

2: Account for the mismatch between $\pi(\mathbf{x})$ and $q(\mathbf{x})$ by calculating the importance weights as

$$w_n = \frac{\pi(\mathbf{x}_n)}{q(\mathbf{x}_n)}, \qquad n = 1, \ldots, N \ . \tag{13}$$

3: Compute the quadrature importance weights as

$$w'_n = w_n v_n N \ , \tag{14}$$

i.e., the product of the importance weight and the quadrature weight.

4: The unnormalized estimator is built as

$$\widehat{I}_{\text{IGH}} = \frac{1}{ZN} \sum_{n=1}^{N} w'_n f(\mathbf{x}_n) \tag{15}$$

if Z is known.

5: The self-normalized estimator is built as

$$\widetilde{I}_{\text{IGH}} = \sum_{n=1}^{N} \bar{w}'_n f(\mathbf{x}_n) \ , \tag{16}$$

where $\bar{w}'_n = \frac{w'_n}{\sum_{j=1}^N w'_j}$. The normalizing constant Z can be approximated as

$$\widehat{Z}_{\text{IGH}} = \frac{1}{N} \sum_{n=1}^{N} w'_n \ . \tag{17}$$

Output: $\{\mathbf{x}_n, w'_n\}_{n=1}^N$

*1) **Toy example 1.** Approximation of the central moments of a modified Nakagami distribution:* The goal is to obtain the central moments of a modified Nakagami distribution given by

$$\widetilde{\pi}(x; \mu, \sigma^2, r) = \frac{|x|^r}{Z_{\sigma^2, r}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) , \tag{18}$$

where $x \in \mathbb{R}$ and $Z_{\sigma^2,r} = \int |x|^r \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx$. Note that for some values of the distribution parameters (μ, σ^2, r), $\widetilde{\pi}$ is a symmetric version of the Nakagami distribution. We approximate now the first 5 even moments, $p \in \{2, 4, 6, 8, 10\}$, with IGH (note that all odd moments are zero due to the symmetry of the pdf). Let us choose the IGH proposal $q(x) = \mathcal{N}(x; \mu, \sigma^2)$, from which we select the N deterministic weighted points $\{x_n, v_n\}_{n=1}^N$. The unnormalized IGH estimator reduces to

$$\widehat{I}_{\text{IGH}} = \frac{1}{Z_{\sigma^2,r} N} \sum_{n=1}^{N} w'_n h(x_n) \tag{19}$$

$$= \frac{1}{Z_{\sigma^2,r} N} \sum_{n=1}^{N} v_n \frac{\pi(x_n)}{q(x_n)} f(x_n) \tag{20}$$

$$= \frac{1}{Z_{\sigma^2,r} N} \sum_{n=1}^{N} v_n x_n^{r+p} \ . \tag{21}$$

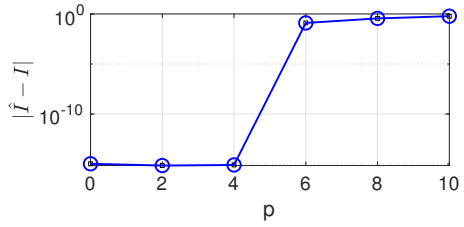


Fig. 1. **Toy example 1.** Target $\widetilde{\pi}(\mathbf{x};\mu,\sigma^2,r)=\frac{|x|^r}{Z_{\sigma^2,r}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$, with $r=4$, $\mu=0$, $\sigma=1$, and $q(x)=\mathcal{N}(x;0,1)$. Function $f(x)=x^p$.

Note that we have chosen the Gaussian proposal such that the exponential term of the target cancels out with the proposal at the IS weight. Note also that $h(x)=\frac{\pi(x)}{q(x)}f(x)$. According to (10), and since $d_x=1$ and $\alpha=N$, the error bound of (21) is

$$|\widehat{I}-I|\leq\frac{N!||h^{(2N)}||_\infty}{(2N)!}=\frac{N!||\left(x^{r+p}\right)^{(2N)}||_\infty}{(2N)!}\,. \qquad (22)$$

Hence, if $2N>r+p$ then the numerator of Eq. (22) is zero, and the estimator has zero error, i.e., $|\widehat{I}_{\text{IGH}}-I|=0$ if the order of the moment satisfies $p\leq 2N-r-1$. Fig. 1 shows the relative absolute error of $\widehat{I}_{\text{IGH}}$ when the number of samples is $N=5$ and the parameter of the target is $r=4$. From Eq. (22), we know that all moments $p\leq 2N-r-1=5$, must be approximated with zero error. Indeed, the figure shows a tiny error of 10^{-15} for all $p<5$, due to the finite computer precision. For, $p>5$ however, the error becomes significant. Note that in this case, the upper bound of Eq. (22) is no longer valid since the bound goes to infinity. Finally, note that the selected IGH proposal is particularly good for the considered target distribution. The use of the Laplace approximation as proposal would not necessarily achieve a successful performance in this problem (for $N=5$, the relative error is around 10^{-5}).

2) ***Toy example 2.*** *Optimal proposal in IS and IGH:* Let us consider a unidimensional Gaussian target $\widetilde{\pi}(x)=\mathcal{N}(x;1,1)$ and we aim at estimating the mean of the target, i.e., $f(x)=x$, in such a way that we know the solution for this toy problem ($I=0$). We apply IS and IGH with the same proposal $q(x)=\mathcal{N}(x;1,\sigma^2)$. We evaluate the performance of the estimators for different values of σ, using $N=5$ samples/points in both algorithms. Moreover, we use a version of IS where, instead of sampling, we obtain the points using randomized quasi Monte Carlo (QMC) [33]. We name this naive approach *importance QMC* (IQMC). In particular, we obtain the points from the Halton sequence [34] (skipping the first point), and use the Rosenblatt transform so their marginal distribution is the desired normal distribution. We use a randomized version using the Cranley-Patterson rotation [17, Chapter 17]. The results are averaged over 200 independent runs. Figure 2(a) shows the the (mean squared error of the unnormalized estimators in IS (dotted blue), IGH (dashed red), and IQMC (solid black), when $\sigma\in[0.85,5]$. We also display the squared error of IGH when the Laplace approximation is set as a proposal (dotted gray). Similarly, Figure 2(c) shows the (mean) squared error of the self-normalized estimators, and Figure 2(c) displays the (mean) squared error of the normalizing constant estimator. In all figures, the blue circle represents the minimum MSE in IS.

In all IS-based estimators, the minimum MSE is achieved with a $\sigma\in[1,2]$, but the minimum is not achieved at the same value for the three estimators (see [17, Chapter 9] for a discussion). The squared error in the IGH estimators are in general several orders of magnitude below the variance of the corresponding IS estimators. Moreover, the minimum error is achieved for a $\sigma=1$ in the three QIS estimators, which coincides with the standard deviation of the target distribution. Note that IQMC always outperforms IS, but it is still far from the performance of IGH with its optimal proposal. Finally, note that using the Laplace approximation as proposal in IGH provides an adequate performance (but not optimal).

In this same setup, now we fix $\sigma=1.5$ and we approximate the normalizing constant with IS, IQMC, and IGH. Note that the choice of σ is particularly good for IS, as shown in Fig. 2(c). In Figure 3, we show the evolution of the (mean) squared error in IS, IQMC, and IGH for several values of $N\in[3,\ 20]$. We also display the IGH with the Laplace approximation as proposal. We see that the convergence rate in this toy example is much faster in IGH than in IS or IQMC.

C. Analysis of the basic IGH and discussion

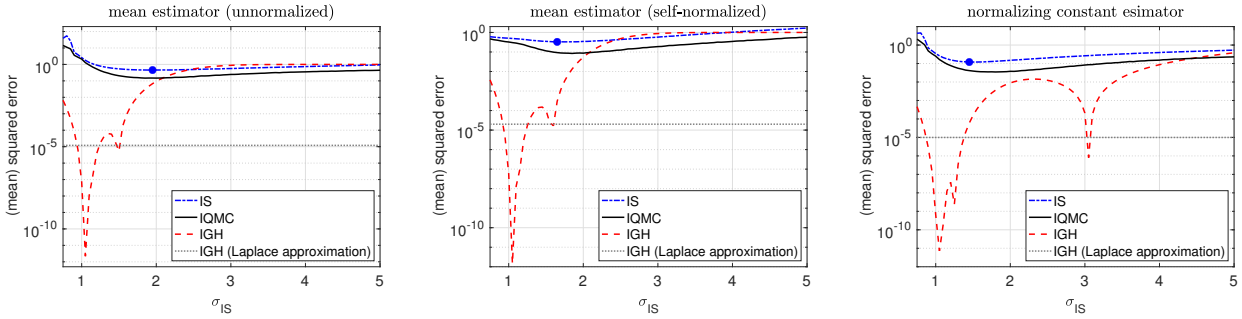
It is interesting to note that in IS the proposal needs to have heavier tails than the integrand, i.e. $h(\mathbf{x})=f(\mathbf{x})\frac{\pi(\mathbf{x})}{q(\mathbf{x})}$ must be bounded. In contrast, in the Toy Example 1, when $p\leq 2N-r-1$, the integrand is $|x|^p\mathcal{N}(x;0,1)$ while the IGH proposal $\mathcal{N}(x;0,1)$ has lighter tails. Let us interpret this from two points of view. On the one hand, regarding Eq. (10), the proposal must be chosen in a way that $h(x)$ is not necessarily bounded, but its 2α-th derivative is, so the error of the IGH estimator is also bounded. In general, if we aim at a perfect integration, then we need to find a $q(\mathbf{x})$ such that the 2α-th derivative of $h(x)$ is zero. On the other hand, in an i.i.d. (random) simulation, the samples are concentrated proportionally to the pdf, while in Gaussian quadrature, from a statistical perspective, the tails are over-represented in terms of nodes (but with an associated weight that is smaller when the distance from the mean to the node grows). For this reason, IGH can still obtain good results with a narrow $q(\mathbf{x})$. This suggests that a Laplace approximation of the integrand, as used for instance in [22], while providing a good performance in some settings, it is not necessarily the best choice (this is also supported by the two toy examples). The results of IGH from the toy examples 1 and 2 are indeed promising, when compared to IS methods. The superior performance comes with some challenges that need to be addressed in order to make IGH a universal methodology that can be used in practical problems.

Theoretical guarantees. Let us first address the convergence of the basic IGH method.

Theorem 1: The unnormalized, $\widehat{I}_{\text{IGH}}$, self-normalized $\widetilde{I}_{\text{IGH}}$, and normalizing constant, $\widehat{Z}_{\text{IGH}}$, estimators in IGH converge to I when $N\to\infty$.

Proof. See Appendix A. □

Remark 1: We recall the re-arrangement of (11) is only valid if $q(\mathbf{x})$ has probability mass for all points where $h(\mathbf{x})\neq$



(a) Mean estimator (unnormalized). (b) Mean estimator (self-normalized). (c) Normalizing constant estimator.

Fig. 2. **Toy example 2.** Target $\widetilde{\pi}(x) = \mathcal{N}(x; 1, 1)$ and $q(x) = \mathcal{N}(x; 0, \sigma^2)$, with $\sigma \in \{0.7, 5\}$. We display the (mean) squared error of the following estimators: (a) Unnormalized estimator of the target mean. (b) Self-normalized estimator of the target mean. (c) Estimator of the normalizing constant. The blue circle represents the minimum MSE in IS. The dashed gray line represents the performance of IGH when the proposal is set to the Laplace approximation (hence with a fixed σ).

0, similarly to what happens in IS. Clearly, this is the case in IGH since $q(\mathbf{x})$ is Gaussian.

The consistency of the estimators ensure the validity of the methodology, but it does not necessarily imply that the approach is efficient for any proposal $q(\mathbf{x})$. Similarly to IS, the performance of IGH depends on the appropriate choice of a proposal density. Note that the bounds in the approximation error given in Section II-B apply directly here. We recall that in IS, the optimal proposal that provides a zero-variance UIS estimator is the one proportional to the integrand of the targeted equation as described above in Section II-A. Interestingly, this result is connected to the optimal proposal in IGH.

Proposition 1: Let us consider a Gaussian proposal $q(\mathbf{x}; \boldsymbol{\theta})$ where $\boldsymbol{\theta}$ contains both the mean and the covariance matrix, and a function f which is non-negative for all values where $\pi(\mathbf{x}) > 0$. Let us suppose that the optimal IS proposal $q^*(\mathbf{x}; \boldsymbol{\theta}^*) = \frac{f(\mathbf{x})\pi(\mathbf{x})}{\int f(\mathbf{x})\pi(\mathbf{x})d\mathbf{x}}$ is Gaussian. Then, the same proposal $q^*(\mathbf{x}; \boldsymbol{\theta}^*)$ used in IGH provides a zero-error unnormalized estimator.

Proof: By plugging $q^*(\mathbf{x}; \boldsymbol{\theta}^*)$ in Eq. (12), then $h^*(\mathbf{x}) = \int f(\mathbf{x})\pi(\mathbf{x})d\mathbf{x} = I$, i.e., a constant. Since the Gauss-Hermite rules integrate polynomials perfectly up to order 2α, the error in this case is zero even with $N = 1$ point. □

Remark 2: If the optimal IS and IGH optimal proposal does not exist in the parametric form $q(\mathbf{x}; \boldsymbol{\theta})$, then the proposal that minimizes the variance of the UIS estimator does not necessarily coincide with the proposal that minimizes the error of the IGH estimator as we show in the second toy example. Note that an extension of the previous proposition can be found in the case that f is non-positive in the support of $\widetilde{\pi}(\mathbf{x})$. The case where f takes both positive and negative signs requires the use of multiple proposals (and two samples to obtain a zero-variance IS estimator). More details can be found in [17, Chapter 13.4]

In real-world problems, it is unlikely that the optimal proposal belongs to the Gaussian family, and hence $h(\mathbf{x})$ is usually not a constant (nor a polynomial) because of the ratio of densities. Therefore, the unnormalized IGH estimator can ensure no error in the estimation of the first 2α terms of the Taylor expansion of $h(\mathbf{x})$, while integration errors will come from the higher-order terms.

In the following, we present two toy examples. The first example shows a case where the proposal is chosen in such a way $h(\mathbf{x})$ is a low-order polynomial, so perfect integration is possible. The second example discusses the best proposal in IS and IGH when perfect integration is not possible, showing that the optimal proposal in IGH is not necessarily the same as in IS.

Computational complexity. We first discuss the computational complexity of IGH and related methods for fixed number of points/samples N, and then we briefly discuss the selection of N. The complexity of deterministic and stochastic integration methods depends on the number of points N at which the target function $h(\cdot)$ needs to be evaluated. For instance, in the standard Bayesian inference framework, every point requires the evaluation of all available data, which may be computationally expensive. Recall that the computational cost of drawing a multi-dimensional sample from a Gaussian distribution is $\mathcal{O}(d_x^2)$ [35]. Additionally, the evaluation of a multivariate Gaussian pdf is $\mathcal{O}(d_x^3)$. In Algorithm 1, we observe that, since the quadrature points are deterministic, they can be stored and only linear scaling and translation (to adjust for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$) is necessary. As such, the $\mathcal{O}(d_x^2)$ term does not apply in IGH. In contrast, since $h(\cdot)$ involves evaluating $q(\cdot)$, the complexity in IGH is dominated by this operation as $\mathcal{C}_{\text{IGH}} = \mathcal{O}(Nd_x^3)$ under the assumption that the complexity of evaluating $q(\cdot)$ is similar to that of $f(\cdot)$ and $\widetilde{\pi}(\cdot)$. Analogously, under Gaussian proposal pdf and same number of points N, the IS method has similar complexity $\mathcal{C}_{\text{IS}} = \mathcal{O}(Nd_x^3)$ since the complexity of drawing from $q(\cdot)$ is negligible compared to evaluating from $q(\cdot)$ in the asymptotic analysis (i.e., $\mathcal{O}(d_x^2 + d_x^3) = \mathcal{O}(d_x^3)$).

IS and Gaussian quadrature algorithms (including the novel IGH framework) require a number of points/samples N that scales exponentially with the dimension, suffering from a similar curse of dimensionality. In connection to this, some quadrature rules (e.g., Gauss-Hermite) generate a number of points of the form $N = \alpha^{d_x}$, where $\alpha \in \mathbb{N}^+$ is the number of points per dimension (i.e., not all arbitrary choices $N \in \mathbb{N}^+$ are possible). This can be cumbersome for some problems,

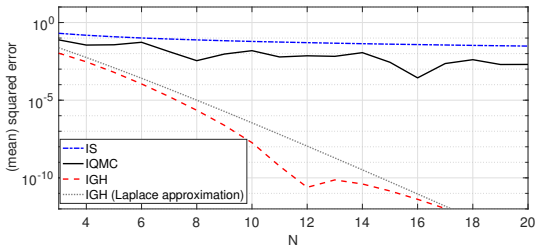


Fig. 3. Evolution of the (mean) squared error in the approximation of Z in IS and IGH when increasing the number of samples/points N.

e.g., when it is possible to select a proposal in such a way $h(\mathbf{x})$ becomes very smooth and one needs a small N (see previous section). In this case, quadrature sparse grids could be used instead [29]. However some drawbacks may appear, e.g., in the Smolyak's quadrature rule some weights can be negative, which implies that the numerical approximation can be negative even if the integrand is strictly positive ([29, Section 3.4] and [17, Chapter 7.8]). This can be a problem for importance quadrature techniques when the self-normalized estimator is used, since the normalization of the weights loses our statistical interpretation, and more practically, it can yield to negative estimations of the normalization constant or to even-order moments (see [29, Section 3.4] for more details). We propose here an alternative to lighten the IGH-based methods by resampling $N' < N$ points with replacement from the pool of N points, with probability equal to its associated quadrature weight. It is easy to show that the new (random) estimator, where the quadrature weights of the resampled points are set to $1/N'$, is unbiased w.r.t. to the costly $\widehat{I}_{\text{IGH}}$ estimator, and converges to it when N' grows (see the example in Section VI-A). Note that this does not reduce the number of points required in order to have a certain level of accuracy, but allows to chose an arbitrary number of points $N' \in \mathbb{N}^+$ where the target will be evaluated (unlike most quadrature rules). Many other similar strategies could be devised, although this goes beyond the scope of this paper.

Self-assessed IGH. Another important issue is the self-assessment of particle-based methods. In IS, the usual measure is an approximation of the effective sample size (ESS). See a discussion about this metric in [36]. We believe that in IGH, another similar metric should be used instead. Following, a recent work about proper metrics in weighted particle-based methods [37], we propose an ESS-like metric for IGH as

$$\text{ESS-IGH} = \frac{N}{\frac{N-1}{\sum_{i\neq j^*} v_i^2 + \left(1-v_{j^*}\right)^2}\left(\sum_{n=1}^{N}(\bar{w}'_n - v_n)^2\right) + 1}. \tag{23}$$

where $j^* = \arg\min_j v_j$. See the derivation and more details in Appendix F. Note that ESS-IGH fulfills the five desired properties described in [37]. For instance, the maximum ESS-IGH $= N$ is only reached in the best scenario when all importance weights w_n are the same (and hence the target is identical to the proposal). Also, the minimum ESS-IGH $= 1$ only occurs in the worst scenario, when only one weight is different from zero, and the associated node receives the minimum quadrature point. Note that, unlike the ESS in IS, the worst-case scenario happens when the point is the furthest point from the mean of the IGH proposal (which has the smallest v_j). In our statistical IGH perspective, this intuition also fits with in this extreme case: not only there is only one effective point, but that the relevant target mass is in the tail of the implicit quadrature proposal (which justifies to receive the minimum ESS-IGH $= 1$). We find this an interesting property, since unlike ESS which is invariant to the node/sample which takes the maximum weight, the ESS-IGH is more penalized when the unique non-zero weighted node is further in the tail.

Automatic IGH. At this point, we would like a sophisticated method that: 1) selects the parameters (mean and covariance) of the proposal density in a way that the integral has minimum error; 2) can operate in situations where the target pdf has multiple modes; 3) can use more than one proposal in order to provide extra flexibility for tackling non-standard distributions; and 4) can adapt to a plethora of complicated problems in an automatic manner. Addressing these challenges is the purpose of the next sections.

IV. Multiple Importance Gauss-Hermite Method

The novel perspective of the basic IGH method can be extended to the case where it is beneficial to use several proposal densities, $\{q_m(\mathbf{x})\}_{m=1}^M$. In the IS literature, it is widely accepted that using several proposals (or a mixture of them) can improve the performance of the method [38], [39], [20]. The justification lays on the fact that the efficiency of IS improves when the mismatch between $|f(\mathbf{x})|\bar{\pi}(\mathbf{x})$ and $q(\mathbf{x})$ decreases. A mixture of proposals is then more flexible in order to reconstruct the targeted integrand.

The extension of IGH from single to multiple proposals is not straightforward as we will show below. In order to establish the basis of this extension, let us first propose a generic multiple IGH (M-IGH) method in Alg. 2. The algorithm receives the parameters of the M proposals, and the number of weighted points per proposal, N. Although N can be different for each proposal, N_m, in this paper we will consider that $N_m = N, \forall m$ for simplicity of notation and the explanation. In Step 1, the N points and associated weights per proposal are chosen. Step 2 computes the importance weights according to some generic weighting scheme $w(\mathbf{x}) = \frac{\pi(\mathbf{x})}{\varphi_m(\mathbf{x})}$, where $\varphi_m(\mathbf{x})$ is a function that can be different for each proposal (see below for more details about the choice of $\varphi_m(\mathbf{x})$). In Step 3, the importance quadrature weights are computed. The unnormalized M-IGH estimator is computed in Eq. (26) of Step 4, and the self-normalized M-IGH estimator in Eq. (27) of Step 5. Note that again an estimator of the normalizing constant is also available in Eq. (28).

Similarly to what happens in MIS [20], there are several possible re-arrangements of the targeted integral that, introducing the set of M proposals, allow for an integral approximation. In the case of IGH, we can extend the basic re-arrangement in Eqs. (11)-(12) in different ways that will lead to different weighting schemes and interpretations. As we show below, these re-arrangements translate into different implementations

of Alg. 2, and in particular, in specific choices of the φ_m in the weights of Eq. (25).

Algorithm 2 Generic Multiple Importance Gauss-Hermite (M-IGH) method

Input: N, $\{\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m\}_{m=1}^M$

1: Select N points $\mathbf{x}_{m,n}$ and the associated quadrature weights v_n, for $n = 1, \dots, N$, associated to each Gaussian pdf $q_m(\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$.

2: Compute the importance weights as

$$w_{m,n} = \frac{\pi(\mathbf{x}_{m,n})}{\varphi_m(\mathbf{x}_{m,n})}, \quad m = 1, \dots, M; \quad n = 1, \dots, N \ . \tag{24}$$

3: Compute the importance quadrature weights as

$$w'_{m,n} = w_{m,n} v_n N \ , \tag{25}$$

that is, the product of the importance weight and the quadrature weight.

4: The unnormalized estimator is built as

$$\widehat{I}_{\text{M-IGH}} = \frac{1}{ZMN} \sum_{m=1}^{M} \sum_{n=1}^{N} w'_{m,n} f(\mathbf{x}_{m,n}) \tag{26}$$

if Z is known.

5: The self-normalized estimator is built as

$$\widetilde{I}_{\text{M-IGH}} = \sum_{m=1}^{M} \sum_{n=1}^{N} \bar{w}'_{m,n} f(\mathbf{x}_{m,n}) \ , \tag{27}$$

where $\bar{w}'_{m,n} = \frac{w'_{m,n}}{\sum_{i=1}^{M} \sum_{j=1}^{N} w'_{i,j}}$. The normalizing constant Z can be approximated as

$$\widehat{Z}_{\text{M-IGH}} = \frac{1}{MN} \sum_{m=1}^{M} \sum_{n=1}^{N} w'_{m,n} \ . \tag{28}$$

Output: $\{\mathbf{x}_{m,n}, w'_{m,n}\}_{m=1,n=1}^{M,N}$

A. Standard multiple IGH (SM-IGH)

This approach is a particular case of Alg. 2, where the importance weight in Eq. (24) for each point $\mathbf{x}_{m,n}$ is computed as $w_{m,n} = \frac{\pi(\mathbf{x}_{m,n})}{q_m(\mathbf{x}_{m,n})}$, i.e., $\varphi_m(\mathbf{x}) = q_m(\mathbf{x})$. Its derivation follows the re-arrangement of the targeted integral, similar to (11)–(12), but now involving the M proposal distributions. It is possible to rewrite I as

$$I = \frac{1}{M} \sum_{m=1}^{M} \int \frac{f(\mathbf{x})\widetilde{\pi}(\mathbf{x})}{q_m(\mathbf{x})} q_m(\mathbf{x}) d\mathbf{x} \tag{29}$$

$$= \frac{1}{M} \sum_{m=1}^{M} \int h_m(\mathbf{x}) q_m(\mathbf{x}) d\mathbf{x} \ , \tag{30}$$

where $h_m(\mathbf{x}) = \frac{f(\mathbf{x})\widetilde{\pi}(\mathbf{x})}{q_m(\mathbf{x})}$. Note that it is possible to approximate the M integrals in (30) by performing M independent IGH algorithms as in previous section, and the unnormalized estimator of Eq. (27) is simply the average of the M parallel estimators. The self-normalized estimator of Eq. (27) however involves the normalization of all MN weights. Interestingly, the re-arrangement of (29) is inspired in the standard multiple MIS scheme (SM-MIS), denoted `N1` scheme in the generalized MIS framework of [20]. For this reason, we denote this algorithm as *standard multiple IGH* (SM-IGH). In the SM-MIS scheme, each sample has an importance weight where only the proposal that was used to simulate the sample appears in the denominator. Note that in [20] it is shown that the MIS scheme `N1` provides a worse performance (largest variance) for the unnormalized estimator in comparison with other MIS schemes (see also [40, Section 4.1.1.]). This poor peformance in MIS is not necessarily translated into a bad performance of the SM-IGH scheme, as we discuss below. However, both SM-MIS and SM-IGH share the construction of the importance weight as $w_{m,n} = \frac{\pi(\mathbf{x}_{m,n})}{q_m(\mathbf{x}_{m,n})}$. The importance weight can be seen as a ratio that measures the mismatch between the target distribution and the denominator of the weight. Therefore, in SM-IGH when $\widetilde{\pi}$ has a complicated form that cannot be mimicked with a Gaussian proposal, no matter how many proposals are employed and how their parameters are selected, the mismatch of $\widetilde{\pi}$ with respect to each proposal will be high. In other words, a given Gaussian $q_m(\mathbf{x})$, regardless of the choice of its parameters, will be unable to mimic the target, yielding $h_m(\mathbf{x})$ very different from a low-order polynomial. In next section, we propose an alternative scheme to overcome this limitation. The following theorem proves the convergence of SM-IGH with N.

Theorem 2: The unnormalized and self-normalized SM-IGH estimators converge to I when $N \to \infty$.

Proof. See Appendix B.

B. Deterministic mixture IGH (DM-IGH)

We present a second variant of Alg. 2 with $\varphi_m(\mathbf{x}) = \frac{1}{M} \sum_{j=1}^{M} q_j(\mathbf{x})$, i.e., the same denominator for all samples of all proposals, which is based on an alternative re-arrangement. Let us first define $\psi(\mathbf{x}) \equiv \frac{1}{M} \sum_{m=1}^{M} q_m(\mathbf{x})$, the mixture of all (Gaussian) proposals. The alternative re-arrangement of I that involves $\psi(\mathbf{x})$ is given by

$$I = \int \frac{f(\mathbf{x})\widetilde{\pi}(\mathbf{x})}{\psi(\mathbf{x})} \psi(\mathbf{x}) d\mathbf{x}$$

$$= \int \frac{f(\mathbf{x})\widetilde{\pi}(\mathbf{x})}{\psi(\mathbf{x})} \frac{1}{M} \sum_{m=1}^{M} q_m(\mathbf{x}) d\mathbf{x} \tag{31}$$

$$= \frac{1}{M} \sum_{m=1}^{M} \int h(\mathbf{x}) q_m(\mathbf{x}) d\mathbf{x} \ , \tag{32}$$

where now the same function

$$h(\mathbf{x}) = \frac{f(\mathbf{x})\widetilde{\pi}(\mathbf{x})}{\psi(\mathbf{x})} = \frac{f(\mathbf{x})\widetilde{\pi}(\mathbf{x})}{\frac{1}{M} \sum_{m=1}^{M} q_m(\mathbf{x})} \ , \tag{33}$$

is present in all M integrals. This re-arrangement is inspired by the deterministic mixture MIS (DM-MIS) scheme, denoted as `N3` in [20], where it is proved to provide the smallest variance in the UIS estimator among of all known MIS schemes. Several reasons explain the good behavior of the

DM-MIS scheme (see the discussion in [40, Section 4.1.1]). Similarly, in the DM-IGH, the M integrands sharing the same function $h(\mathbf{x})$ that contains the mixture $\psi(\mathbf{x})$ with all proposals on its denominator.[3] We recall that $\widetilde{\pi}$ can be skewed, multimodal, or with different tails than a Gaussian, and while the Gaussian restriction in the proposals is limiting, under mild assumptions, any distribution can be approximated by a mixture of Gaussians [41], [42]. In the case of DM-IGH, and following similar arguments in Section III-C, if the M Gaussians are selected in such a way $h(\mathbf{x})$ can be approximated by a low-order polynomial, then all M integrals in Eq. (32) will be approximated with low error, and the DM-IGH will be accurate. Note that DM-IGH requires $O(NM^2)$ operations compared to $O(NM)$ operations in SM-IGH. We now prove the convergence of the DM-IGH method.

Theorem 3: The unnormalized and self-normalized DM-IGH estimators converge to I when $N \to \infty$.

Proof. See Appendix C.

Corollary 1: As a result of Theorem 3, one can form a partition of proposals and apply the DM-IGH method in each partition, combining then the estimators similarly to the case in MIS [21], [43], [44].

V. Selection and adaptation of the proposal

The proposed IGH methodology and its variants requires the selection of the mean and covariances of the (potentially multiple) proposal distributions. As in IS, an adequate selection of those parameters is crucial in obtaining the desired results from IGH. In this section, we provide two adaptive extensions to the IGH methodology such that the inference process can be automated and performed adaptively with little practitioner interaction.

A. Adaptive multiple IGH (AM-IGH)

We propose a first adaptive IGH algorithm that iteratively adapts the proposals through moment matching mechanisms (see [18] for a description of moment-matching strategies in adaptive IS). We describe the new method in Alg. 3 naming it as *adaptive multiple IGH* (AM-IGH). The algorithm runs for T iterations[4], adapting the parameters of the proposal $q^{(t)}(\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)})$ at each iteration t. The importance weights are computed in (34), where the generic function in the denominator $\varphi^{(t,\tau)}$ is discussed below. Note also that at each iteration, the importance weights corresponding to previous $t-1$ iterations might be also recomputed for a reason that will be apparent below. Then, the quadrature importance weights are computed in Eq. (35), which are then normalized in Eq. (36). Finally, the proposal is adapted through moment matching using the set of all Nt (re)-weighted points. In particular, we match the first and second moments of the target, which allows for the update of the mean and covariance matrix of the proposal.

[3]Note that we are forcing the Gaussians in the mixture to be equally weighted, but it would be straightforward to extend the scheme to the case where the mixture is $\psi_{\boldsymbol{\beta}}(\mathbf{x}) = \sum_{m=1}^{M} \beta_m q_m(\mathbf{x})$ instead.

[4]The term *multiple* comes from the fact that after T iterations, T different proposals have been used (see [45] for more details).

The generic Alg. 3 can be particularized for different choices of the function $\varphi^{(t,\tau)}(\mathbf{x})$ in the denominator of the weights. One reasonable choice is to use $\varphi^{(t,\tau)}(\mathbf{x}) = q_\tau(\mathbf{x})$, i.e., applying the proposal that was used to choose the points that are being weighted. In this case, it is not necessary to reweight the points of the previous $t-1$ iterations, i.e., only N weight calculations are done at each iteration. Another possible choice is $\varphi^{(t,\tau)}(\mathbf{x}) = \frac{1}{t}\sum_i^t q^{(i)}(\mathbf{x}_n^{(\tau)})$. Hence, all the sequence of proposals is used in the mixture of the denominator. However, in order to balance the presence of a proposal in the weight of future points, the past points must also be reweighted to incorporate the future proposals. Therefore, at each iteration t, not only the N new points receive a weight, but also the past $N(t-1)$ points need to be reweighted. This has a clear connection with the DM-IGH of Section IV-B. By plugging this choice in Alg. 3, the method has certain parallelism with the celebrated IS-based AMIS algorithm [45] that obtains a high performance in a plethora of applications (see [18] for more details). One limitation of this weighting scheme is that the cost in proposal evaluations grows quadratically with T (while it is linear when the choice $\varphi^{(t,\tau)}(\mathbf{x}) = q_\tau(\mathbf{x})$). Another limitation is that the consistency of the AMIS algorithm has not yet been proved (or the lack of it). Recently, a new method for alleviating the computational complexity of AMIS was proposed, also improving the stability of the algorithm [46]. The method choses iteratively and automatically a mixture $\varphi^{(t,\tau)}(\mathbf{x})$ with a reduced number of components. A similar mechanism can also be used in the proposed AM-IGH framework.

Algorithm 3 Adaptive Multiple Importance Gauss-Hermite (AM-IGH) method

Input: N, T, $\boldsymbol{\mu}^{(1)}$ $\boldsymbol{\Sigma}^{(1)}$

1: **for** $t = 1, \ldots, T$ **do**

2: Select N points $\mathbf{x}_n^{(t)}$ and the associated quadrature weights v_n, for $n = 1, \ldots, N$, associated to the Gaussian pdf $q^{(t)}(\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)})$.

3: Compute (and recompute) the (previous) importance weights as

$$w_n^{(\tau)} = \frac{\pi(\mathbf{x}_n^{(\tau)})}{\varphi^{(t,\tau)}(\mathbf{x}_n^{(\tau)})}, \quad n = 1, \ldots, N; \quad \tau = 1, \ldots, t. \tag{34}$$

4: Compute the importance quadrature weights as

$$w_n^{'(\tau)} = w_n^{(\tau)} v_n N, \quad n = 1, \ldots, N; \quad \tau = 1, \ldots, t\,, \tag{35}$$

that is, the product of the importance weight and the quadrature weight.

5: Compute the normalized importance weights as

$$\bar{w}_n^{'(\tau)} = \frac{w_n^{'(\tau)}}{\sum_{i=1}^{t}\sum_{k=1}^{N} w_k^{'(i)}}\,, \quad \tau = 1, \ldots, t\,. \tag{36}$$

6: Estimate the mean and the covariance of the target with the set of available weighted Nt points, and set $\boldsymbol{\mu}^{(t+1)}$ $\boldsymbol{\Sigma}^{(t+1)}$ to those values.

TABLE I. SUMMARY OF IGH METHODS IN TERMS OF PROPOSAL AND TARGET EVALUATIONS PER POINT.

	standard IGH	SM-IGH	DM-IGH	AM-IGH	M-PIGH
proposal eval.	1	1	M	TM^*	M
target eval.	1	1	1	1	1
multiple prop.	no	yes	yes	yes*	yes
adaptive	no	no	no	yes	yes

B. Multiple Population IGH (M-PIGH)

In many scenarios, the target distribution is multimodal and/or with a shape that cannot be well approximated by a Gaussian distribution. This is well known in the AIS literature, where most of methods employ several proposal densities in order to approximate the target distribution with a mixture the adapted proposals. Examples of the adaptation of multiple proposals in IS can be found in [47], [48], [45], [49], [40] among many others.

Here, we propose a second adaptive scheme, called *multiple population IGH* (M-PIGH), whose adaptation relies fully in the deterministic rules re-interpreting the adaptivity mechanism of M-PMC [48], an AIS algorithm (hence fully based in Monte Carlo). In summary, the original M-PMC iteratively adapts a mixture proposal of kernels (including the parameters of the kernels and their weight in the mixture) in a stochastic EM-based manner in order to minimize the KL-divergence between the mixture proposal and the target distribution. In M-PIGH, we select quadrature points and weights instead of sampling from the kernels. In order to not over-complicate the novel algorithm with a variable number of points per kernel, we do not adapt the weight of each kernel in the mixture proposal (hence, we do adapt the mean and covariances of the Gaussian kernels). For sake of brevity, we briefly describe the algorithm. M-PIGH adapts a mixture with M equally-weighted Gaussian kernels, that are initialized with some mean and covariance matrices. For T iterations, M-PIGH selects the points and quadrature weights as in IGH, compute the importance weights using the whole mixture in the denominator (implementing the DM-IGH approach) and builds the usual IGH estimators. The means and covariances of next iterations are computed through the Rao-Blackwellized version of the moment matching proposed in [50] and later implemented in [48]. A multimodal numerical example is presented in Section VI-B, where we compare the proposed M-PIGH and the original M-PMC. Finally, Table summarizes computational complexity of all proposed algorithms. Moreover, it displays whether the algorithms use multiple proposals and whether they are adaptive. Note that the AM-IGH algorithm can be implemented with $M = 1$ proposal, but also with $M \geq 1$

VI. SIMULATION RESULTS

In the first example, we build a posterior distribution and test the AM-IGH in a challenging signal-processing example. In the second example, we test the M-PIGH in a multimodal scenario. In the third example, we consider a Bayesian machine learning example where the hyperparameters of a Gaussian process are learned.

A. Inference in a exoplanetary model

In this section, we consider an astrophysics problem that deals with an exoplanetary system [51], [52]. We consider a simplified model of a Keplerian orbit and the radial velocity of a host star where the observations are given by

$$y_r(t_d) = v + k\left[\cos\left(\frac{2\pi}{p}t_d + \omega\right) + e\cos(\omega)\right] + \xi\ , \quad (37)$$

where t_d, with $d = 1, \ldots, D$, represent the time instants, $y_r(t_d)$ is the r-th observation obtained at the t_d-th instant, with $r = 1, \ldots, R$, V is the mean radial velocity, k is an amplitude, p is the period, ω is longitude of periastron, e the eccentricity of the orbit and $\xi \sim \mathcal{N}(0, \sigma_o^2)$ models the variance of the observation noise, σ_o^2 being known. Note that $t_1, t_2, \ldots, t_D$ are (known) time instants where the observations are acquired. In this example, we consider that the five parameters of the system (v, k, p, e, ω) are unknown, i.e., we aim at inferring the random variable $\mathbf{X} = [V, K, P, E, \Omega]^\top$ in dimension $d_x = 5$. In this Bayesian inference problem, we consider uniform priors as follows: $p(V) = \mathcal{U}[-15, 15]$, $p(K) = \mathcal{U}[0, 50]$, $p(P) = \mathcal{U}[0, 365]$, $p(E) = \mathcal{U}[0, 2\pi]$, and $p(\Omega) = \mathcal{U}[0, 1]$.

For this example, we implement the AM-IGH method with $N = 10^5$ samples/points per iteration, and $T = 20$ iterations. We simulate the model with the values $\mathbf{X} = [3, 2, 200, \pi, 0.2]^\top$, $D = 40$ time instants, and $\sigma_o^2 = 2$ for the observation noise. We made several tests for different values of R and since the results were coherent and did not provide any new insights, we discuss here those with $R = 1$. We approximate the first moment of the posterior distribution of $\mathbf{X}$ given the set of data. Fig. 4 shows the MSE in the estimate of the mean of the posterior distribution building the estimators with the samples at each iteration t. In both AM-IGH and AMIS (for comparison), the means of the Gaussian proposals have been initialized randomly in the square $[-1.5, 6] \times [1, 4] \times [100, 400] \times [\frac{\pi}{2}, 2\pi] \times [0.1, 0.4]$, and averaged over 100 independent runs. First, we observe that AM-IGH converges faster to a stable point. Second, AMIS has still not converged to a stable proposal distribution at the last iteration. We recall that the cost of AMIS in proposal evaluations is quadratic with T, which becomes a limitation when many iterations are needed to find a good proposal. It is worth noting that the achieved MSE of AM-IGH is several orders of magnitude below that of AMIS. Moreover, we have implemented two versions of AM-IGH that used the resampling strategy proposed in Section III-C, using $N' = N/2$ and $N' = N/5$ resampled nodes. Interestingly both algorithms converge faster than AM-IGH ($N' = N/2$ is the fastest), but AM-IGH obtains a better performance after few iterations (with $N' = N/2$ being better than $N' = N/5$). The interpretation is simple: in all cases, the best performance is attained after some iterations, and the performance is limited by the number of nodes at the given iteration. We also display the solution given by the Laplace approximation [22], finding the mode through a costly simulated annealing [53] with $E = 2 \cdot 10^6$ target evaluations). For unimodal distributions, a good initialization of AM-IGH may be this Laplace approximation, including the use of the Hessian as in [22], although it is hard to display

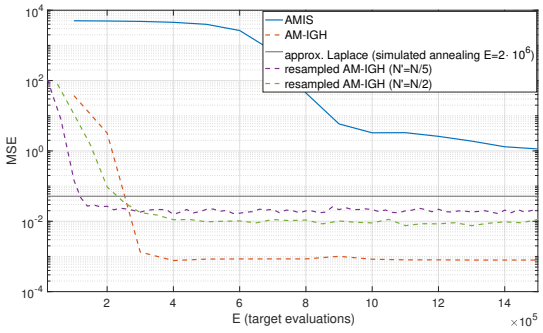


Fig. 4. **Ex. 1.** MSE in the estimate of the mean of the posterior distribution in the exoplanetary model (averaged over the $d_x = 5$ dimensions).

a fair comparison since finding the mode is a tough problem when the target is non-concave in the logarithmic domain (as in the considered problem).

B. Multimodal distribution

In this example we aim at approximating moments of a multimodal distribution given by the mixture

$$\pi(\mathbf{x}) = \frac{1}{5}\sum_{i=1}^{5}\mathcal{N}(\mathbf{x};\boldsymbol{\nu}_i,\mathbf{C}_i), \quad \mathbf{x}\in\mathbb{R}^2, \tag{38}$$

with the following mean vectors and covariance matrices: $\nu_1 = [-10,-10]^\top$, $\boldsymbol{\nu}_2 = [0,16]^\top$, $\boldsymbol{\nu}_3 = [13,8]^\top$, $\boldsymbol{\nu}_4 = [-9,7]^\top$, $\boldsymbol{\nu}_5 = [14,-14]^\top$, $\mathbf{C}_1 = [2,\ 0.6; 0.6,\ 1]$, $\mathbf{C}_2 = [2,\ -0.4; -0.4,\ 2]$, $\mathbf{C}_3 = [2,\ 0.8; 0.8,\ 2]$, $\mathbf{C}_4 = [3,\ 0; 0,\ 0.5]$, and $\mathbf{C}_5 = [2,\ -0.1; -0.1,\ 2]$.

In this numerical example, due to the multi-modality, we implement M-PIGH, the novel adaptive quadrature method presented in Section V-B. Unlike in [54], here the adaptive mechanism is also based on IGH. Table II shows the MSE in the estimation of the mean of the target and the normalizing constant, with both the (stochastic) M-PMC algorithm and the (deterministic) M-PIGH algorithm. In order to compare their behavior, we initialized randomly the location parameters of the kernels within the $[-4,4]\times[-4,4]$ square, i.e., without covering any modes of the target, in order to better evaluate the adaptivity of the algorithms. For both, M-PMC and M-PIGH we use an adaptive mixture with $M = 25$ proposals/kernels, $N = 25$ samples/points per proposal and iteration, for $T \in \{5, 10, 20\}$. We try three different initializations for the scale parameters of the proposals, with $\boldsymbol{\Sigma}_m^{(1)} = \sigma_1^2\mathbf{I}$ with $\sigma_1 \in \{1, 2, 5\}$. The results are averaged over 100 random initializations. In all cases, we compare both algorithms with equal number of target evaluations. We see that M-PIGH outperforms M-PMC in all setups, obtaining in some cases an improvement of more than one order of magnitude. For a small scale parameter initialization $\sigma_1 = 1$, both algorithms have trouble improving their estimate, although M-PIGH is able to significantly improve while M-PMC does not. Larger initial scale parameters benefit both algorithms. We also see that when the number of iterations T is increased, M-PIGH decreases the MSE in a larger factor than the M-PMC: the quadrature rules are not only useful for a better approximation but also for a faster adaptation.

C. Learning Hyperparameters in Gaussian processes with automatic relevance determination

Gaussian processes (GPs) are Bayesian state-of-the-art methods for function approximation and regression [55], where selecting the covariance function and learning its hyperparameters is the key to attain significant performance. Here we present an example in the context of estimating the hyperparameters in the *automatic relevance determination (ARD* covariance functions [56, Chapter 6]. The observations are P data pairs $\{y_j, \mathbf{z}_j\}_{j=1}^P$, with $y_j \in \mathbb{R}$ and $\mathbf{z}_j = [z_{j,1}, \ldots, z_{j,L}]^\top \in \mathbb{R}^L$, where L is the dimension of the input features. We denote the joint output vector as $\mathbf{y} = [y_1, \ldots, y_P]^\top$. The goal is to infer an unknown function f which links the variables y and $\mathbf{z}$ as

$$y = f(\mathbf{z}) + e, \tag{39}$$

where $e \sim N(e; 0, \sigma^2)$. The function $f(\mathbf{z})$ is considered to be a realization of a GP [55], $f(\mathbf{z}) \sim \mathcal{GP}(\mu(\mathbf{z}), \kappa(\mathbf{z}, \mathbf{r}))$, where $\mu(\mathbf{z}) = 0$, $\mathbf{z}, \mathbf{r} \in \mathbb{R}^L$ with kernel function

$$\kappa(\mathbf{z}, \mathbf{r}) = \exp\left(-\sum_{\ell=1}^{L}\frac{(z_\ell - r_\ell)^2}{2\delta_\ell^2}\right). \tag{40}$$

The hyper-parameters $\delta_\ell > 0$ corresponding to each input dimension are stacked in $\boldsymbol{\delta} = \delta_{1:L} = [\delta_1, \ldots, \delta_L]$. We consider the problem of learning the posterior of all hyper-parameters of the model, given by

$$\boldsymbol{\theta} \quad = \quad [\theta_{1:L} = \delta_{1:L}, \theta_{L+1} = \sigma] = [\boldsymbol{\delta}, \sigma] \in \mathbb{R}^{L+1},$$

i.e., all the parameters of the kernel function in Eq. (40) and the standard deviation σ of the observation noise. We assume a prior $p(\boldsymbol{\theta}) = \prod_{\ell=1}^{L+1}\frac{1}{\theta_\ell^\beta}\mathbb{I}_{\theta_\ell}$ where $\beta = 1.3$ and $\mathbb{I}_v = 1$ if $v > 0$, and $\mathbb{I}_v = 0$ if $v \leq 0$. Note that we are focusing on learning the marginal posterior of $p(\boldsymbol{\theta}|\mathbf{y})$, which can be obtained from $p(\boldsymbol{\theta}, f|\mathbf{y})$, taking into account that $p(f|\boldsymbol{\theta}, \mathbf{y})$ is tractable (see [57] for more details about this example).

In Fig. 5 we consider the case with $L = 3$, so the target is in $\mathbb{R}^4$, and set a ground truth of $\boldsymbol{\delta}^* = [1, 3, 1]$, $\sigma^* = \frac{1}{2}$ (recall that $\boldsymbol{\theta}^* = [\boldsymbol{\delta}^*, \sigma^*]$). We have generated $P = 500$ pairs of data, $\{y_j, \mathbf{z}_j\}_{j=1}^P$, drawing $\mathbf{z}_j \sim \mathcal{U}([0, 10]^L)$ and $\mathbf{y}_j$ according to the model in Eq. (39). We implement the AM-IGH algorithm of Table 3 with $\varphi^{(t,\tau)}(\mathbf{x}) = q^{(t)}(\mathbf{x})$, and with $\varphi^{(t,\tau)}(\mathbf{x}) = \frac{1}{t}\sum_j^t q^{(j)}(\mathbf{x})$, with $\tau = 1, ..., t$, which we denote AM-IGH (DM) in the plot. We also compare with the MC-based method AMIS [45], which incorporates similar weights, and a QMC version of AMIS, named LAIQMC, where the random sampling is substituted by randomized QMC samples (see the details in the second toy example). We consider also the LAIS algorithm [58], with $M = 1$ proposal, and an IGH version of it, denoted as LA-IGH (also with $M = 1$ proposal). Moreover, we introduce a variation of LAIS where, the lower layer implements a randomized QMC version (as explained above). We name this version as LAIQMC. In all LAIS-based algorithms, we use the variant

TABLE II. **Ex. 2** MSE IN THE ESTIMATION OF THE MEAN AND THE NORMALIZING CONSTANT OF THE M-PMC (AIS METHOD) AND THE M-PIGH (NOVEL ADAPTIVE QUADRATURE METHOD).

		$T=5$			$T=10$			$T=20$		
		$\sigma_1=1$	$\sigma_1=3$	$\sigma_1=5$	$\sigma_1=1$	$\sigma_1=3$	$\sigma_1=5$	$\sigma_1=1$	$\sigma_1=3$	$\sigma_1=5$
MSE (mean estimate)	M-PMC	46.4	55.6	11.7	67.7	57.9	8.25	72.8	63.1	7.59
	M-PIGH	18.8	6.94	3.12	9.56	5.13	1.3	8.3	4.21	0.245
MSE (Z estimate)	M-PMC	1.04	0.681	0.0989	0.824	0.63	0.0299	0.729	0.571	0.026
	M-PIGH	0.34	0.058	0.034	0.2	0.0385	0.0137	0.141	0.0257	0.00607

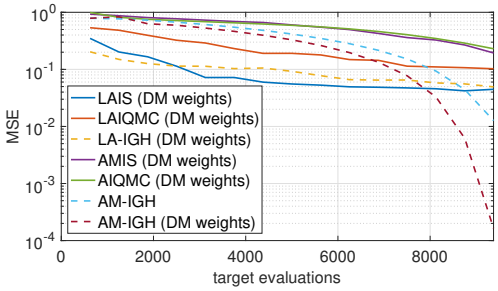


Fig. 5. **Ex. 3.** Mean squared error of several learning algorithms in the estimation of the mean of the posterior distribution of the hyperparameters of a Gaussian process. All algorithms are compared as a function of the total number of target evaluations (i.e., samples or points) $E = NMT$. For each algorithm and configuration, the number of iterations T is set in such a way $E = 10^4$.

with temporal DM weights (the whole sequence of proposals appear as a mixture in all weights), as described for AM-IGH above. More precisely, both LAIS and LA-IGH run a Metropolis-Hastings (MH) chain of length T, generating the sequence $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_T$ of location parameters for the IS and IGH methods, respectively. We use a variance $\sigma^2 = 0.4$ in all algorithms, with the curves being similar for other choices. In all algorithms, we set $N = 625$ samples/nodes per iteration. All considered algorithms are iterative, and the comparison is done in terms of target evaluations. We compute the posterior mean (ground truth) with an exhaustive and very costly Monte Carlo approximation so we can compare the methods (see [57] for more details). We see that LAIS and LA-IGH exhibit a better performance for a low number of target evaluations (few iterations). The proposed LA-IGH always outperforms LAIS in all the setups we have tested. The proposed AM-IGH and AM-IGH-DM algorithm largely outperform all competitors when the number of iterations is increased. Interestingly, the cheaper version AM-IGH (only one proposal evaluation per sample) is still very competitive w.r.t. the other alternatives, although it provides two orders of magnitude larger error than the AM-IGH-DM.

VII. CONCLUSIONS

In this paper, we have introduced a generic framework for numerical integration, extending its range of application due to (a) the introduction of a novel importance sampling (IS) perspective, and (b) the incorporation of several ideas from the IS literature. The framework can also be interpreted as an incorporation of deterministic rules into IS methods, reducing the error of the estimators by several orders of magnitude. The potential of the proposed methodology was shown on three numerical examples, as well as two toy examples used in the motivation of the method. This IS perspective allows the use of quadrature rules (in particular, this work focused on Gauss-Hermite rules, although it can be easily applied to other types of Gaussian quadrature rules) in problems where the integrand does not fulfill the standard requirements in numerical integration. Moreover, the new IS-based framework can also be used when the normalizing constant is unknown, extending its applicability to Bayesian inference. The methodology is completed with a set of extensions, including the use of mixtures of proposals and adaptive approaches to automatically adjust the parameters. Finally, the methodology comes with convergence guarantees and error bounds, which are validated in the discussed examples showing MSE results orders of magnitude below state-of-the-art importance sampling methods.

APPENDIX A
PROOF OF THEOREM 1

First, note that $\widehat{I}_{\text{IGH}}$ can be rewritten as in (8) if $q(\mathbf{x})$ is non-zero for all $\mathbf{x}$ where $\widetilde{\pi}(\mathbf{x})$ is non-zero. Then, due to the quadrature arguments reviewed in Section II-B, $\widehat{I}_{\text{IGH}}$ converges to I. The convergence of the self-normalized estimator $\widetilde{I}_{\text{IGH}}$ is also guaranteed due to similar arguments in IS [15, Section 3.3.2]. Note that (16) can be rewritten as $\widetilde{I}_{\text{IGH}} = \frac{\widehat{I}_{\text{IGH}} Z}{\widehat{Z}_{\text{IGH}}}$. Both $\widehat{I}_{\text{IGH}}$, the unnormalized estimator in (15), and $\widehat{Z}_{\text{IGH}}$, the normalizing constant estimator in (17), converge to the desired quantities when N goes to infinity (note that $\widehat{Z}_{\text{SM-IGH}}$ is a particular case of $\widehat{I}_{\text{SM-IGH}}$ with $f(\mathbf{x}) = 1$). Then, since both the numerator and denominator converge, and since $Z \neq 0$ by construction, we have that $\widetilde{I} \to I$ when N goes to infinity. □

APPENDIX B
PROOF OF THEOREM 2

We first write unnormalized SM-IGH estimator by substituting $\varphi_m(\mathbf{x}) = q_m(\mathbf{x})$ in Eq. (24):

$$\widehat{I}_{\text{SM-IGH}} = \frac{1}{ZMN} \sum_{m=1}^{M} \sum_{n=1}^{N} w'_{m,n} f(\mathbf{x}_{m,n}) = \frac{1}{ZMN} \sum_{m=1}^{M} \sum_{n=1}^{N} v_n \frac{\pi(\mathbf{x}_{m,n})}{q_m(\mathbf{x}_{m,n})} f(\mathbf{x}_{m,n}) \,. \quad (41)$$

Due to the properties of the Gauss-Hermite integration,

when N goes to infinity,

$$\lim_{N\to\infty} \widehat{I}_{\text{SM-IGH}} = \frac{1}{ZM}\sum_{m=1}^{M}\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N} v_n \frac{\pi(\mathbf{x}_{m,n})}{q_m(\mathbf{x}_{m,n})} f(\mathbf{x}_{m,n})$$
$$= \frac{1}{MZ}\sum_{m=1}^{M}\int \frac{\pi(\mathbf{x})}{q_m(\mathbf{x})} f(\mathbf{x}) q_m(\mathbf{x}) d\mathbf{x}$$
$$= \frac{1}{Z}\int \pi(\mathbf{x}) f(\mathbf{x}) d\mathbf{x} = I. \quad (42)$$

Since $\widehat{Z}_{\text{SM-IGH}}$ also converges with M (we recall that it is a particular case of $\widehat{I}_{\text{SM-IGH}}$ with $f(\mathbf{x}) = 1$). Due to the same arguments of Section III-C, and the convergence of both $\widehat{I}_{\text{SM-IGH}}$ and $\widehat{Z}_{\text{SM-IGH}}$ then self-normalized $\widetilde{I}_{\text{SM-IGH}}$ also converges with N. □

APPENDIX C
PROOF OF THEOREM 3

Let us first write explicitly the unnormalized DM-IGH estimator as

$$\widehat{I}_{\text{DM-IGH}} = \frac{1}{ZMN}\sum_{m=1}^{M}\sum_{n=1}^{N} w'_{m,n} f(\mathbf{x}_{m,n})$$
$$= \frac{1}{ZMN}\sum_{m=1}^{M}\sum_{n=1}^{N} v_n \frac{\pi(\mathbf{x}_{m,n})}{\frac{1}{M}\sum_{j=1}^{M} q_j(\mathbf{x}_{m,n})} f(\mathbf{x}_{m,n})$$

Again, following quadrature arguments, when N goes to infinity,

$$\lim_{N\to\infty} \widehat{I}_{\text{DM-IGH}} =$$
$$= \frac{1}{ZM}\sum_{m=1}^{M}\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N} v_n \frac{\pi(\mathbf{x}_{m,n})}{\frac{1}{M}\sum_{j=1}^{M} q_j(\mathbf{x}_{m,n})} f(\mathbf{x}_{m,n})$$
$$= \frac{1}{MZ}\sum_{m=1}^{M}\int \frac{\pi(\mathbf{x})}{\frac{1}{M}\sum_{j=1}^{M} q_j(\mathbf{x})} f(\mathbf{x}) q_m(\mathbf{x}) d\mathbf{x}$$
$$= \frac{1}{Z}\int \frac{\pi(\mathbf{x})}{\frac{1}{M}\sum_{j=1}^{M} q_j(\mathbf{x})} f(\mathbf{x}) \frac{1}{M}\sum_{m=1}^{M} q_m(\mathbf{x}) d\mathbf{x}$$
$$= \frac{1}{Z}\int \pi(\mathbf{x}) f(\mathbf{x}) d\mathbf{x} = I. \quad (43)$$

Similarly, since $\widehat{Z}_{\text{DM-IGH}}$ also converges with M because of the same reasons as in the IGH and SM-IGH methods, then self-normalized $\widetilde{I}_{\text{DM-IGH}}$ also converges to I when N goes to infinity. □

APPENDIX D
BOUND ON THE QUADRATURE ERROR

We aim at upper bounding the error as in Eq. (10) and showing that asymptotically, as $\alpha \to \infty$, the error vanishes. Notice that when the function can be approximated with a polynomial of degree $2\alpha - 1$, the error is zero. Therefore, in the following analysis we are interested in situations where the nonlinearity is such that $p \geq 2\alpha$, where p is the order of the nonlinearity. We make use of useful results regarding bounds of the supremum of a function's derivative [59], [60]. Let $f_p(x)$ a polynomial of order p such that, in the open interval (a, b), its supremum is bounded by a constant M_0, i.e., $\sup|f_p(x)| = M_0$, then the following inequality holds for the first derivative

$$|f_p^{(1)}(x)| \leq \frac{2M_0\, n^2}{b-a} = M_1, \quad (44)$$

and, in general, for the i-th derivative we have that

$$|f_p^{(i)}(x)| \leq K(i,p)\frac{M_0}{(b-a)^i} = M_i, \quad (45)$$

where

$$K(i,p) = \frac{2^i p^2(p^2-1)\cdots(p^2-(i-1))}{1\cdot 3\cdot 5\cdots(2i-1)} \quad (46)$$
$$= \frac{p}{p+i} 2^{2i}\cdot i! \begin{pmatrix} p+i \\ 2i \end{pmatrix}, \quad (47)$$

and equality only holds for Chebyshev polynomials. We aim at showing that $M_i \geq M_{i+1}$, meaning that the supremum of the derivative $i+1$ is bounded from above by the supremum of the i-th derivative. Using the above expressions $\frac{M_{i+1}}{M_i} = \frac{1}{b-a}\frac{p^2-i}{i(i+1/2)}$, which, for large i, tends to 0 such that $\frac{M_{i+1}}{M_i} \leq 1$ is satisfied asymptotically. This result is supported by a d'Alembert's ratio test analysis, which states that if the limit of the ratio is such that $\lim_{i\to\infty}\left|\frac{M_{i+1}}{M_i}\right| < 1$, then the series converges absolutely.

APPENDIX E
GAUSSIAN QUADRATURE RULES

For the sake of simplicity and without loss of generality, let us consider $d_x = 1$, i.e., $x \in \mathbb{R}$. A quadrature formula $\widehat{I} = \sum_{n=1}^{\alpha} v_n h(x_n)$ is an approximation of integral of type $I = \int_{\mathcal{D}} h(x) q(x) dx$ in Eq. (11), i.e.,

$$I = \int_{\mathcal{D}} h(x) q(x) dx \approx \widehat{I} = \sum_{n=1}^{\alpha} v_n h(x_n)\ . \quad (48)$$

The function $q(x)$ plays the role of a weighting function (i.e., a density) and it is not required to be normalized, i.e., we only need to assume that $\int_{\mathcal{D}} q(x)dx < \infty$, i.e., $q(x)$ is an unnormalized density [1], [2]. Given the function $q(x)$, in order to properly select these 2α unknown values (all the weights v_n's and all the nodes x_n's), we can consider a nonlinear system of 2α equations matching the first 2α non-central moments, i.e.,

$$\sum_{n=1}^{N} v_n x_n^r = \int_{\mathcal{D}} x^r q(x) dx, \quad \text{for} \quad r = 0, \ldots, 2\alpha - 1\ , \quad (49)$$

where v_n's and x_n's play the role of unknown and the integrals $\int_{\mathcal{D}} x^r q(x)dx$ (i.e., r-th moment of $q(x)$) should be a known value. Therefore, if the first 2α non-central moments $\int_{\mathcal{D}} x^r q(x)dx$ are available, the non-linear system is well-defined. However, since this system of equations is highly nonlinear, generally the solution is not available [1], [2]. Some specific choices of density $q(x)$ admit a closed-form expression. Table III shows some relevant examples.

TABLE III. GAUSSIAN QUADRATURE RULES

Gaussian Quadrature Rule	**weighted function** $q(x)$	**Domain** $\mathcal{D}$	**Nodes** x_n	**Weights** v_n
Legendre	$q(x) \propto 1$	$[-1, 1]$	roots of Legendre polynomials $P_\alpha(x)$	$v_n = \frac{2}{(1-x_n)[P'_\alpha(x_n)]^2}$
Chebyshev-Gauss	$q(x) \propto \frac{1}{\sqrt{1-x^2}}$	$(-1, 1)$	$x_n = \cos\left(\frac{2n-1}{2\alpha}\pi\right)$	$v_n = \frac{\pi}{\alpha}$
Chebyshev-Gauss-2	$q(x) \propto \sqrt{1-x^2}$	$[-1, 1]$	$x_n = \cos\left(\frac{n}{\alpha+1}\pi\right)$	$v_n = \frac{\pi}{\alpha+1}\sin\left(\frac{n}{\alpha+1}\pi\right)$
Gauss-Laguerre	$q(x) \propto \exp(-x)$	$[0, \infty)$	roots of Laguerre polynomials $L_\alpha(x)$	$v_n = \frac{x_n}{(\alpha+1)^2[L_{\alpha+1}(x_n)]^2}$
Gauss-Hermite	$q(x) \propto \exp(-x^2)$	$(-\infty, \infty)$	roots of Hermite polynomials $H_\alpha(x)$	$v_n = \frac{2^{\alpha-1}\alpha!\sqrt{\pi}}{\alpha^2[H_{\alpha-1}(x_n)]^2}$

APPENDIX F
ESS-IGH

Let us define the Euclidean distance between the two pmfs that define the IGH approximation $\{\bar{w}'_n\}_{n=1}^N$ and the quadrature approximation $\{v_n\}_{n=1}^N$

$$L_2 = \sqrt{\sum_{n=1}^{N}(\bar{w}'_n - v_n)^2} \tag{50}$$

$$= \sqrt{\sum_{n=1}^{N} v_n^2 \left(\frac{w_n}{\sum_{j=1}^N w_j v_j} - 1\right)^2} \tag{51}$$

$$= \sqrt{\sum_{n=1}^{N} v_n^2 \left(\frac{w_n}{\hat{Z}} - 1\right)^2}. \tag{52}$$

When all the importance weights, w_n are equal, $L_2 = 0$. Following the arguments in [37], the maximum in (52) happens when only one importance weight w_n is different from zero. But unlike in IS, here the position of the single non-zero weight plays a role (note that here the nodes are no longer i.i.d. as in IS, and hence they are not exchangeable). Let us denote

$$j^* = \arg\min_j v_j,$$

and hence v_{j^*} is the minimum quadrature weight. Then, the maximum L_2 is

$$L_2^* = \sqrt{\sum_{i\neq j^*}(0 - v_i)^2 + (1 - v_{j^*})^2}, \tag{53}$$

$$= \sqrt{\sum_{i\neq j^*} v_i^2 + (1 - v_{j^*})^2}, \tag{54}$$

i.e., the worst-case is determined by the case where the unique non-zero weight is the one associated to the minimum quadrature weight. In Section III-C we given an intuition why this result is relevant.

Next, we can build a metric ESS-IGH that fulfilled the five desired properties stated in [37], e.g., we would like that ESS-IGH takes its maximum when all the importance weights are the same (which corresponds to the target being identical to the proposal), and its minimum when one extreme point takes the only non-zero weight. We impose the structure ESS-IGH $= \frac{1}{aL_2^2+b}$, choosing a and b in such a way ESS-IGH $= 1$ in the worst scenario ($L_2 = \sqrt{\sum_{i\neq j^*} v_i^2 + (1 - v_{j^*})^2}$), and ESS-IGH $= N$ in the best case ($L_2 = 0$), which yields $a = \frac{N-1}{NL_2^{*2}}$ and $b = \frac{1}{N}$. Hence,

$$\text{ESS-IGH} = \frac{N}{\frac{N-1}{L_2^{2*}}\left(\sum_{n=1}^N v_n^2\left(\frac{w_n}{\hat{Z}} - 1\right)^2\right) + 1} \tag{55}$$

$$= \frac{N}{\frac{N-1}{L_2^{2*}}\left(\sum_{n=1}^N (\bar{w}'_n - v_n)^2\right) + 1}. \tag{56}$$

REFERENCES

[1] J. Stoer and R. Bulirsch, *Introduction to numerical analysis*, vol. 12, Springer Science & Business Media, 2013.

[2] D. Ballreich, *Deterministic Numerical Integration*, pp. 47–91, Springer International Publishing, Cham, 2017.

[3] P. Stano et al., "Parametric Bayesian filters for nonlinear stochastic dynamical systems: a survey," *IEEE Trans. on Cybernetics*, vol. 43, no. 6, pp. 1607 – 1624, Dec. 2013.

[4] K. Ito and K. Xiong, "Gaussian filters for nonlinear filtering problems," *IEEE Trans. on Automatic Control*, vol. 45, no. 5, pp. 910–927, May 2000.

[5] I. Arasaratnam, S. Haykin, and R. J. Elliot, "Discrete-time nonlinear filtering algorithms using Gauss-Hermite quadrature," *Proc. of the IEEE*, vol. 95, no. 5, pp. 953–977, 2007.

[6] P. Closas, C. Fernández-Prades, and J. Vilà-Valls, "Multiple Quadrature Kalman filtering," *IEEE Trans. Signal Processing*, vol. 60, no. 12, pp. 6125–6137, Dec. 2012.

[7] J. Vilà-Valls, P. Closas, and Á. García-Fernández, "Uncertainty exchange through multiple quadrature Kalman filtering," *IEEE Signal Processing Letters*, vol. 23, no. 12, pp. 1825–1829, 2016.

[8] J. Uhlmann S. J. Julier and H. F. Durrant-Whyte, "A new method for the non linear transformation of means and covariances in filters and estimators," *IEEE Trans. Automatic Control*, vol. 3, pp. 477–482, March 2000.

[9] I. Arasaratnam and S. Haykin, "Cubature Kalman filters," *IEEE Trans. Automatic Control*, vol. 54, no. 6, pp. 1254–1269, June 2009.

[10] P. Closas, J. Vilà-Valls, and C. Fernández-Prades, "Computational complexity reduction techniques for quadrature Kalman filters," in *Proc. of the CAMSAP'15*, Cancun, Mexico, Dec. 2015.

[11] B. Shizgal, "A Gaussian quadrature procedure for use in the solution of the Boltzmann equation and related problems," *Journal of Computational Physics*, vol. 41, no. 2, pp. 309–328, 1981.

[12] W. Gautschi and G. V. Milovanović, "Gaussian quadrature involving Einstein and Fermi functions with an application to summation of series," *Mathematics of Computation*, vol. 44, no. 169, pp. 177–190, 1985.

[13] L. Liu and X. Huang, "The use of Gaussian quadrature for estimation in frailty proportional hazards models," *Statistics in Medicine*, vol. 27, no. 14, pp. 2665–2683, 2008.

[14] S. Sarkka, *Bayesian Filtering and Smoothing*, 3 edition, 2013.

[15] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*, Springer, 2004.

[16] J. S. Liu, *Monte Carlo Strategies in Scientific Computing*, Springer, 2004.

[17] A. B. Owen, *Monte Carlo theory, methods and examples*, http://statweb.stanford.edu/∼owen/mc/, 2013.

[18] M. F. Bugallo, V. Elvira, L. Martino, D. Luengo, J. Míguez, and P. M. Djuric, "Adaptive importance sampling: The past, the present, and the future," *IEEE Signal Processing Magazine*, vol. 34, no. 4, pp. 60–79, 2017.

[19] R. E. Caflisch, "Monte Carlo and quasi-Monte Carlo methods," *Acta numerica*, vol. 7, pp. 1–49, 1998.

[20] V. Elvira, L. Martino, D. Luengo, and M. F. Bugallo, "Generalized multiple importance sampling," *Statistical Science*, vol. 34, no. 1, pp. 129–155, 2019.

[21] V. Elvira, L. Martino, D. Luengo, and M. F. Bugallo, "Efficient multiple importance sampling estimators," *Signal Processing Letters, IEEE*, vol. 22, no. 10, pp. 1757–1761, 2015.

[22] Q. Liu and Donald A. Pierce, "A note on Gauss-Hermite quadrature," *Biometrika*, vol. 81, no. 3, pp. 624–629, 1994.

[23] J. C. Pinheiro and D. M. Bates, "Approximations to the log-likelihood function in the nonlinear mixed-effects model," *Journal of Computational and Graphical Statistics*, vol. 4, no. 1, pp. 12–35, 1995.

[24] A. Azevedo-Filho and R. D. Shachter, "Laplace's method approximations for probabilistic inference in belief networks with continuous variables," in *Uncertainty Proceedings 1994*, pp. 28–36. Elsevier, 1994.

[25] P. Kabaila and N. Ranathunga, "On adaptive Gauss-Hermite quadrature for estimation in glmms," in *Research School on Statistics and Data Science*. Springer, 2019, pp. 130–139.

[26] Á. F. García-Fernández, F. Tronarp, and S. Särkkä, "Gaussian process classification using posterior linearization," *IEEE Signal Processing Letters*, vol. 26, no. 5, pp. 735–739, 2019.

[27] H. Engels, "Numerical quadrature and cubature," 1980.

[28] S. A. Smolyak, "Quadrature and interpolation formulas for tensor products of certain classes of functions," in *Doklady Akademii Nauk*. Russian Academy of Sciences, 1963, vol. 148, pp. 1042–1045.

[29] F. Heiss and V. Winschel, "Likelihood approximation by numerical integration on sparse grids," *Journal of Econometrics*, vol. 144, no. 1, pp. 62–80, 2008.

[30] G. H. Golub and J. H. Welsch, "Calculation of Gauss quadrature rules," *Mathematics of Computation*, vol. 23, no. 106, pp. 221–230, 1969.

[31] P. J. Davis and P. Rabinowitz, *Methods of Numerical integration*, Courier Corporation, 2007.

[32] W. Gautschi, "A survey of Gauss-Christoffel quadrature formulae," in *EB Christoffel*, pp. 72–147. Springer, 1981.

[33] A. B. Owen, "Quasi-monte carlo sampling," *Monte Carlo Ray Tracing: Siggraph*, vol. 1, pp. 69–88, 2003.

[34] H. Niederreiter, *Random Number Generation and Quasi-Monte Carlo Methods*, Society for Industrial Mathematics, 1992.

[35] D. B. Thomas and W. Luk, "Multivariate Gaussian random number generation targeting reconfigurable hardware," *ACM Transactions on Reconfigurable Technology and Systems*, vol. 1, no. 2, June 2008.

[36] V. Elvira, L. Martino, and C. P. Robert, "Rethinking the effective sample size," *arXiv preprint arXiv:1809.04129*, 2018.

[37] L. Martino, V. Elvira, and F. Louzada, "Effective sample size for importance sampling based on discrepancy measures," *Signal Processing*, vol. 131, pp. 386 – 401, 2017.

[38] E. Veach and L. Guibas, "Optimally combining sampling techniques for Monte Carlo rendering," *In SIGGRAPH 1995 Proceedings*, pp. 419–428, 1995.

[39] A. B. Owen and Y. Zhou, "Safe and effective importance sampling," *Journal of the American Statistical Association*, vol. 95, no. 449, pp. 135–143, 2000.

[40] V. Elvira, L. Martino, D. Luengo, and M. F. Bugallo, "Improving Population Monte Carlo: Alternative weighting and resampling schemes," *Signal Processing*, vol. 131, no. 12, pp. 77–91, 2017.

[41] W. Feller, *An Introduction to Probability and Its Applications*, vol. II of *Wiley Publication in Mathematical Statistics*, Wiley India Pvt. Limited, 1966.

[42] H. W. Sorenson and D. L. Alspach, "Recursive Bayesian estimation using Gaussian sums," *Automatica*, vol. 7, pp. 465–479, 1971.

[43] V. Elvira, L. Martino, D. Luengo, and M. F. Bugallo, "Heretical multiple importance sampling," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1474–1478, 2016.

[44] V. Elvira, L. Martino, D. Luengo, and M. F. Bugallo, "Multiple importance sampling with overlapping sets of proposals," *IEEE Workshop on Statistical Signal Processing (SSP)*, 2016.

[45] J. M. Cornuet, J. M. Marin, A. Mira, and C. P. Robert, "Adaptive multiple importance sampling," *Scandinavian Journal of Statistics*, vol. 39, no. 4, pp. 798–812, December 2012.

[46] V. Elvira Y. El-Laham, L. Martino and M. F. Bugallo, "Efficient adaptive multiple importance sampling," in *2019 27th European Signal Processing Conference (EUSIPCO)*. IEEE, 2019, pp. 1–5.

[47] O. Cappé, A. Guillin, J. M. Marin, and C. P. Robert, "Population Monte Carlo," *Journal of Computational and Graphical Statistics*, vol. 13, no. 4, pp. 907–929, 2004.

[48] O. Cappé, R. Douc, A. Guillin, J. M. Marin, and C. P. Robert, "Adaptive importance sampling in general mixture classes," *Statistics and Computing*, vol. 18, pp. 447–459, 2008.

[49] L. Martino, V. Elvira, D. Luengo, and J. Corander, "An adaptive population importance sampler: Learning from uncertainty," *IEEE Trans. Signal Process.*, vol. 63, no. 16, pp. 4422–4437, 2015.

[50] G. R. Douc, J. M. Marin, and C. P. Robert, "Minimum variance importance sampling via population Monte Carlo," *ESAIM: Probability and Statistics*, vol. 11, pp. 427–447, 2007.

[51] S. T Balan and O. Lahav, "Exofit: orbital parameters of extrasolar planets from radial velocities," *Monthly Notices of the Royal Astronomical Society*, vol. 394, no. 4, pp. 1936–1944, 2009.

[52] F. Feroz, S. T. Balan, and M. P. Hobson, "Detecting extrasolar planets from stellar radial velocities using bayesian evidence," *Monthly Notices of the Royal Astronomical Society*, vol. 415, no. 4, pp. 3462–3472, 2011.

[53] S. Kirkpatrick, C. D. Gelatt, and M P. Vecchi, "Optimization by simulated annealing," *science*, vol. 220, no. 4598, pp. 671–680, 1983.

[54] V. Elvira, P. Closas, and L. Martino, "Gauss-Hermite Quadrature for non-Gaussian Inference via an Importance Sampling Interpretation," in *27th European Signal Processing Conference (EUSIPCO)*, 2019, pp. 1637–1641.

[55] C. E. Rasmussen and C. K. Williams, *Gaussian processes for machine learning*, MIT press Cambridge, MA, 2006.

[56] C. M. Bishop, *Pattern recognition and machine learning*, Springer, 2006.

[57] L. Martino, V. Elvira, and G. Camps-Valls, "The recycling gibbs sampler for efficient learning," *Digital Signal Processing*, vol. 74, pp. 1–13, 2018.

[58] L. Martino, V. Elvira, D. Luengo, and J. Corander, "Layered adaptive importance sampling," *Statistics and Computing*, vol. 27, no. 3, pp. 599–623, 2017.

[59] O. Ore, "On functions with bounded derivatives," *Trans. of the American Mathematical Society*, vol. 43, no. 2, pp. 321–326, 1938.

[60] A. C. Schaeffer, "Inequalities of A. Markoff and S. Bernstein for polynomials and related functions," *Bulletin of the American Mathematical Society*, vol. 47, no. 8, pp. 565–579, 1941.