SIAM/ASA J. UNCERTAINTY QUANTIFICATION
Vol. 9, No. 3, pp. 1094–1131

A Higher Order Unscented Transform*

Deanna C. Easley† and Tyrus Berry†

Abstract. We develop a new approach for estimating the expected values of nonlinear functions applied to multivariate random variables with arbitrary distributions. Rather than assuming a particular distribution, we assume that we are only given the first four moments of the distribution. The goal is to efficiently represent the distribution using a small number of quadrature nodes which are called σ-points. What we mean by this is choosing nodes and weights in order to match the specified moments of the distribution. The classical scaled unscented transform (SUT) matches the mean and covariance of a distribution. In this paper, we introduce the higher order unscented transform (HOUT), which also matches any given skewness and kurtosis tensors. It turns out that the key to matching the higher moments is the tensor CANDECOMP/PARAFAC (CP) decomposition. While the minimal CP decomposition is NP-complete, we present a practical algorithm for computing a nonminimal CP decomposition and prove convergence in linear time. We then show how to combine the CP decompositions of the moments in order to form the σ-points and weights of the HOUT. By passing the σ-points through a nonlinear function and applying our quadrature rule we can estimate the moments of the output distribution. We prove that the HOUT is exact on arbitrary polynomials up to fourth order and derive error bounds in terms of the regularity of the function and the decay of the probability. Finally, we numerically compare the HOUT to the SUT on nonlinear functions applied to non-Gaussian random variables including an application to forecasting and uncertainty quantification for chaotic dynamics.

Key words. unscented transform, skewness, kurtosis, tensors, CP decomposition, Kalman filter

AMS subject classifications. 65D32, 41A55, 65R10, 15A69, 62H12

DOI. 10.1137/20M135546X

1. Introduction.

A fundamental problem in uncertainty quantification is to approximate the expectation of a function $f : \mathbb{R}^d \to \mathbb{R}$ applied to a random variable X sampled from a probability measure dp on $\mathbb{R}^d$, namely

$$\mathbb{E}[f(X)] = \int f(x)\, dp. \tag{1.1}$$

Even when the distribution is known this can be a challenging computation in high dimensions, and the problem is often compounded by uncertain or incomplete knowledge of f and dp. Moreover, in most problems of interest f has an extremely complex form. For example, f may encapsulate the solution of a differential equation and the computation of some feature of interest on the solution. So we may not be able to assume that f is known in an explicit form

*Received by the editors July 27, 2020; accepted for publication (in revised form) May 17, 2021; published electronically August 20, 2021.
https://doi.org/10.1137/20M135546X
Funding: Both authors were supported by NSF grant DMS-2006808, and the second author was also supported by NSF grant DMS-1854204.

†Department of Mathematical Sciences, George Mason University, Fairfax, VA 22030 USA (deasley2@gmu.edu, tberry@gmu.edu).

but instead that f or an approximation to f is available only as a black-box computational scheme which can take inputs x and produce outputs $f(x)$. Similarly, the type of partial knowledge of the probability measure can vary widely. We may have an explicit expression for a density function $p(x) = dp/dx$ (if it even exists), or we may only have some samples of dp or estimates of some of the moments.

The method developed in this manuscript will assume that the first four moments of the probability measure, dp, exist and can be accurately estimated. Our method will not use any additional knowledge of the probability beyond these moments. Moreover, we will not require any explicit knowledge of f, so our method is applicable if f is a black-box. In order to derive error bounds we will require some regularity assumptions on f and additional decay assumptions on the probability measure at infinity. While our error bounds depend on the error of approximating f by a polynomial, our method does not require us to actually find such an approximation and will only require evaluating f on a small number of test points.

The problem of approximating (1.1) can be approached as a problem of numerical quadrature (also known as cubature when x has dimensionality greater than one; we will use the term quadrature for both). A quadrature is an approximation of the form

$$\mathbb{E}[f(X)] \approx \sum_{i=1}^{N} w_i f(x_i), \tag{1.2}$$

where x_i are called nodes and w_i are called weights. The goal is to find a small number of nodes and weights that accurately represent the probability measure for a large space of functions $f \in \mathcal{C}$. A common strategy in quadrature methods is to choose nodes and weights so that the above approximation is actually an equality for all f in some finite dimensional subspace $\tilde{\mathcal{C}} \subset \mathcal{C}$ (such as a space of polynomials up to a fixed degree). For f outside of $\tilde{\mathcal{C}}$ we can then attempt to bound the error in the approximation (1.2) if we can control the error between f and its projection into $\tilde{\mathcal{C}}$. When f is sufficiently smooth and dp is concentrated in a small region, then it is reasonable to approximate f using the space of polynomials up to a fixed degree. Under these assumptions, we can bound the error between f and a low degree polynomial via interpolation error bounds.

Ensuring that (1.2) holds with equality for all polynomials up to degree k is equivalent to satisfying the so-called moment equations,

$$m_{j_1,\dots,j_n} = \mathbb{E}\left[X_1^{j_1} X_2^{j_2} \cdots X_n^{j_n}\right] = \sum_{i=1}^{N} w_i x_1^{j_1} x_2^{j_2} \cdots x_n^{j_n} \tag{1.3}$$

for all $j_1 + j_2 + \cdots + j_n \le k$, since polynomials of the form $x_1^{j_1}, \dots x_n^{j^n}$ form a basis of the space of all polynomials of degree less than or equal to k. In other words we are asking that the empirical moments of the nodes x_i, weighted by discrete probabilities w_i, exactly agree with the true moments $m_{j_1,\dots,j_n}$ of the distribution. When $k = 2$ the moment equations specify that weighted nodes must match the mean vector and covariance matrix of the true distribution, and this is achieved with the so-called scaled unscented ensemble (SUT) [18] (see section 2.1 for an overview).

The quadrature approach is an alternative to stochastic quadrature methods such as Monte Carlo quadrature which is commonly used in particle filtering. Stochastic quadratures use random variables X_i to build quadrature rules such that

$$\mathbb{E}[f(X)] \approx \mathbb{E}\left[\sum_{i=1}^{N} w_i f(X_i)\right]. \tag{1.4}$$

However, the computed value $\sum_{i=1}^{N} w_i f(X_i)$ will be stochastic. This means that in addition to possible approximation error in (1.4), we also have an error due to the variance of the random variable $\sum_{i=1}^{N} w_i f(X_i)$. While it is often easier to design stochastic quadrature methods where the approximation error in (1.4) is small or even zero, for many problems controlling the variance error requires a large number of random variables X_i and hence a large number of function evaluations. When f is very expensive to compute, it may be more efficient to use a small deterministic ensemble and accept the quadrature error in (1.2) in order to avoid the large ensemble size that would be required to control the variance in a stochastic quadrature.

The problem (1.1) is often part of a larger problem such as filtering [20], particle filtering [32], adaptive filtering [4], smoothing [30], parameter estimation [33, 34, 11], and even model-free filtering [12]. In all these applications it can be beneficial to have deterministic approximation of (1.1) to improve the stability of the overall algorithm. For example, filters built on random ensembles can fail catastrophically since they can generate realizations that would normally have very low probability but lead to perverse behavior [13, 1]. Similarly, a gradient-based optimization method for parameter estimation will need to carefully account for any stochasticity in the objective function, so replacing a stochastic quadrature with a deterministic quadrature can be desirable in certain applications.

The highly successful unscented Kalman filter (UKF) [20] is based on the SUT, as are many of the other methods mentioned above. A closely related technique called cubature Kalman filters (CKF) [2] follow a similar strategy and are typically designed to achieve a high degree of exactness under a Gaussian assumption on the distribution. Another potentially deterministic method would be quadrature based on sparse grids [15, 27]; however, designing such a quadrature typically requires detailed knowledge of the probability distribution. Similarly, polynomial chaos expansions [35, 25] require explicit knowledge of the function f and the distribution. Our method is an alternative quadrature that only requires us to know the first four moments of the distribution. Moreover, the nodes of our quadrature will be adapted to the moments of the distribution. A potential future application of the method developed here would be to a higher order UKF which tracks four moments, and this was one of the inspirations behind this work. However, the current work only generalizes the forecast step of the UKF to four moments, and generalizing the assimilation step of the UKF is a significant remaining challenge.

In this paper, we develop a higher order unscented transform (HOUT) based on tensor decomposition of the first four moments of a distribution. Whereas the UKF (and implicitly most CKFs) only requires the rank decomposition of the covariance matrix, the HOUT requires the CANDECOMP/PARAFAC (CP) decomposition of higher order tensors such as the skewness and kurtosis. The CP decomposition of a k-tensor is defined by vectors v_i such that

$$T = \sum_{i=1}^{p} v_i^{\otimes k}, \tag{1.5}$$

i.e., the CP decomposition decomposes a tensor as the summation of rank-1 tensors, $v_i^{\otimes k}$ for $i = 1, \ldots, p$. The minimal value of p such that the above decomposition exists is called the rank of T. For detailed definitions of tensors, tensor product ($\otimes$), and tensor decomposition, see section 2.2. For the sake of giving an overview, we assume these definitions for now. For details on tensor product and CP decomposition see Definitions 2.6 and 2.8. For a more detailed introduction to tensors we suggest [10, 24].

Ideally, we would like an exact CP decomposition (1.5) with the minimum possible number of vectors; however, this turns out to be an NP-complete problem [9, 14]. Instead, we will use an effective algorithm for obtaining an approximate CP decomposition up to an arbitrary tolerance. The algorithm was originally suggested by [22], and it works by repeatedly subtracting the best rank-1 approximation to a tensor until the norm of the residual is less than any desired tolerance. Many methods have been developed based on this idea (see [8] and citations therein) and in [7] it was proven to converge but without any convergence rate. In section 3 we give the first proof that this algorithm converges linearly and we derive an upper bound on the convergence rate. While the approximate decomposition typically requires many more vectors than the minimal CP decomposition, it avoids the NP-completeness of that problem and gives us an effective algorithm.

In section 2 we briefly review the SUT and some tensor facts and notation including the higher order power method (HOPM) [5] that we will use for finding tensor eigenvectors. Based on the HOPM, we prove the convergence of the approximate CP decomposition algorithm in section 3. This proof also requires some new inequalities relating the maximum eigenvalue of a tensor to the entries of the tensor, and these inequalities are likely to be of independent interest. In section 4 we introduce the HOUT, which generalizes the SUT in order to match arbitrary skewness and kurtosis tensors. The HOUT gives a quadrature rule with degree of exactness four that is applicable to arbitrary distributions. For a preview of the nodes of the HOUT, see Figure 1, where we consider data sampled from two-dimensional distributions with nontrivial skewness and kurtosis tensors. For each distribution we show the HOUT nodes that are designed so that the first four moments computed with this small number of nodes

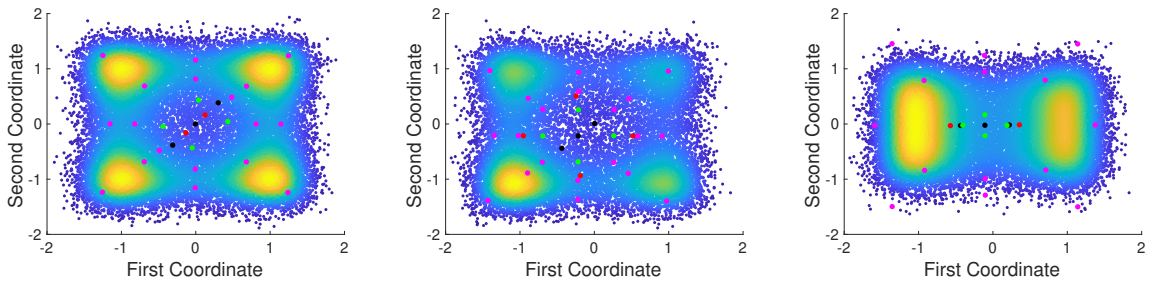


Figure 1. *The 4-moment σ-points of the HOUT we developed of* (a) *a quadramodal distribution,* (b) *a skewed quadramodal distribution, and* (c) *a skewed bimodal distribution. The black points are the σ-points that correspond to the mean, the green points are the σ-points corresponding to the covariance, the points in red are the σ-points corresponding to the skewness, and the magenta points are the σ-points that correspond to the kurtosis.*

match the true moments of the distribution up to the specified tolerance. In section 5 we derive error bounds under appropriate regularity assumptions on f and decay assumptions on the probability. Finally, we demonstrate the HOUT on various non-Gaussian multivariate random variables on complex nonlinear transformations in section 6 and briefly conclude in section 7.

2. Background.

We start by reviewing the SUT in section 2.1 which has degree of exactness two. We then briefly introduce our tensor notation in section 2.2 and tensor-vector products and tensor norms in section 2.3. Finally, in section 2.4 we review tensor eigenvectors and eigenvalues and the HOPM [5] for finding them.

2.1. Scaled unscented transform.

The SUT was introduced by Julier and Uhlmann in [18] and further developed in [21, 19, 16, 17, 20]. The fundamental goal of this paper is to generalize their method to higher order moments. This work was started in [16], which worked on matching the skewness, and below we show that CP decompositions are the key to generalizing their approach.

The SUT uses the mean and covariance of a distribution to choose quadrature nodes and weights such that the quadrature rule has degree of exactness 2. Degree of exactness k means that a quadrature rule is exact for computing the expectation of polynomials up to degree k. The fundamental insight of Julier and Uhlmann is that achieving degree of exactness 2 is equivalent to matching the first two moments of the distribution. Moreover, they showed that this can be efficiently accomplished using a matrix square root of the covariance matrix.

Definition 2.1 (ith column of the symmetric matrix square root of A). *Let A be a $d \times d$ matrix. We define the i*th column of the symmetric matrix square root of A, *denoted $\sqrt{A}_i$, by*

$$\sum_{i=1}^{d} \sqrt{A}_i^{\otimes 2} = \sum_{i=1}^{d} \sqrt{A}_i \sqrt{A}_i^{\top} = A.$$

The notation $v^{\otimes k}$ will be defined below. Note that the following definition can use any matrix square root but we have found empirically that the unique symmetric matrix square root has the best performance. The negative of any matrix square root is also a matrix square root. The following definition perturbs the mean μ by both a matrix square root and its negative to create an ensemble of $2d+1$ points.

Definition 2.2 (the scaled unscented transform [18]). *Let dp be a probability measure with mean $\mu \in \mathbb{R}^d$ and the covariance $C \in \mathbb{R}^{d\times d}$. Then for some $\beta \in \mathbb{R}$ the* σ-points *are defined by*

$$\sigma_i = \begin{cases} \mu & \text{if } \ i = 0, \\ \mu + \beta\sqrt{C}_i & \text{if } \ i = 1, \dots, d, \\ \mu - \beta\sqrt{C}_{i-d} & \text{if } \ i = d+1, \dots, 2d \end{cases}$$

and the corresponding weights *are defined by*

$$w_i = \begin{cases} 1 - \frac{d}{\beta^2} & \text{if } \ i = 0, \\ \frac{1}{2\beta^2} & \text{if } \ i = 1, \dots, 2d. \end{cases}$$

We note that the choice of β can have a significant impact on the effectiveness of the transform.

Remark 2.3. The absolute condition number of the SUT is bounded above by

$$\sum_{i=0}^{2d} |w_i| = \left|1 - \frac{d}{\beta^2}\right| + \frac{d}{\beta^2}.$$

If $\beta \geq \sqrt{d}$, then $\sum_{i=0}^{2d} |w_i| = 1$. If $\beta < \sqrt{d}$, then $\sum_{i=0}^{2d} |w_i| = \frac{2d}{\beta^2} - 1$.

The following theorem says that the SUT matches the first two moments, μ and C.

Theorem 2.4 (empirical mean and empirical covariance [18]). *For an arbitrary β, we have*

$$\mu = \mathbb{E}[X] = \sum_{i=0}^{2d} w_i \sigma_i \qquad \text{and} \qquad C = \mathbb{E}[(X-\mu)(X-\mu)^\top] = \sum_{i=0}^{2d} w_i (\sigma_i - \mu)(\sigma_i - \mu)^\top,$$

and if $q : \mathbb{R}^d \to \mathbb{R}$ is a polynomial of degree at most 2, we have, $\mathbb{E}[q(X)] = \sum_{i=0}^{2d} w_i q(\sigma_i)$.

We should note that if the distribution has zero skewness, such as a Gaussian distribution, then the symmetry of the nodes yields degree of exactness 3, and in the specific case of a Gaussian distribution the choice $\beta = \sqrt{3}$ achieves degree of exactness 4 [18, 17, 20]. The choice $\beta = \sqrt{d}$ is often called the *unscented transform* and sets $w_0 = 0$ so that only $2d$ of the σ-points are required. The ability of the SUT to match the first four moments of the Gaussian distribution has led some to associate the SUT with a Gaussian assumption; however, this is not required and degree of exactness 2 is achieved for arbitrary distributions. Our goal is to generalize the unscented transform to higher moments, which are tensors.

2.2. Tensors. Tensors are essentially multidimensional matrices, which will be used to conveniently express the notions of covariance, skewness, and kurtosis in a similar fashion.

Definition 2.5 (k-order tensor). *For positive integers d and k, a tensor T belonging to $\mathbb{R}^{d^k}$ is called a k-*order tensor *or simply a k-*tensor.

In particular, a vector in $\mathbb{R}^d$ can be viewed as a first order tensor and a $d \times d$ matrix as a second order tensor. Let $x \in \mathbb{R}^d$. We note that the outer product $xx^\top$ yields a $d \times d$ matrix whose ij-entry can be represented as

$$(xx^\top)_{ij} = x_i x_j = (x \otimes x)_{ij} = (x^{\otimes 2})_{ij}.$$

We generalize this process of forming higher order tensors from vectors with the following definition.

Definition 2.6 (kth-order tensor product). *Let $v \in \mathbb{R}^d$ and k be a positive integer. Then the kth-*order tensor product *is a k-tensor denoted*

$$v^{\otimes k} = \underbrace{v \otimes v \otimes \cdots \otimes v}_{k \text{ times}},$$

and the elements are given by $(v^{\otimes k})_{i_1,\ldots,i_k} = v_{i_1} \ldots v_{i_k}$.

Definition 2.6 immediately connects tensor products to the moments of a distribution since we can represent the covariance as $C = \mathbb{E}[(X-\mu)^{\otimes 2}] = \int (x-\mu)^{\otimes 2}\, dp(x)$ so that the skewness S and kurtosis K can be defined as

$$S = \int (x-\mu)^{\otimes 3}\, dp(x), \qquad K = \int (x-\mu)^{\otimes 4}\, dp(x),$$

so that, for example,

$$S_{ijk} = \int \left((x-\mu)^{\otimes 3}\right)_{ijk}\, dp(x) = \int (x-\mu)_i (x-\mu)_j (x-\mu)_k\, dp(x).$$

The following definition generalizes the notion of a rank-1 matrix to tensors.

Definition 2.7 (rank-1 tensor). *Let $T \in \mathbb{R}^{d^k}$. Then T is called a* rank-1 tensor *if there exists a $v \in \mathbb{R}^d$ such that*

$$v^{\otimes k} = T.$$

For tensors that are not rank-1, one may seek to decompose the tensor as a sum of rank-1 tensors.

Definition 2.8 (CP decomposition). *The vectors $v_1, \ldots, v_p$ form a* CP decomposition *of a tensor T if*

$$T = \sum_{\ell=1}^{p} v_\ell^{\otimes k}$$

and the minimum p for which such a decomposition exists is called the rank *of the tensor T.*

This notion of rank agrees with the classical notion of matrix rank in the case of second order tensors but many of the properties of matrix rank do not generalize to higher order tensors [24, 9, 14, 31, 23].

2.3. Tensor multiplication and tensor norms. In this section we introduce the necessary definitions and notation along with some preliminary results that will be needed below. The proofs of the lemmas along with a more detailed introduction to tensor multiplication can be found in Appendix A (see also [24]).

Definition 2.9 (n-mode product of a tensor). *The* n-mode product *of a k-order tensor $T \in \mathbb{R}^{d^k}$ with a vector $v \in \mathbb{R}^d$, denoted by $T \times_n v$, is defined elementwise as*

$$(T \times_n v)_{i_1,\ldots,i_{n-1},i_{n+1},\ldots,i_k} = \sum_{j=1}^{d} T_{i_1,\ldots,i_{n-1},j,i_{n+1},\ldots,i_k} v_j.$$

Note that $T \times_n v \in \mathbb{R}^{d^{k-1}}$, so the order of the resulting tensor is decreased by 1.

The above definition can also be generalized for tensor-matrix multiplication [24]. Finally we note that the Frobenius norm for matrices can be generalized to tensors in the following way.

Definition 2.10 (tensor Frobenius norm [24]). *The* Frobenius norm of a tensor $T \in \mathbb{R}^{d^k}$ *is the square root of the sum of the squares of all its elements,*

$$\|T\|_F = \sqrt{\sum_{i_1=1}^{d} \cdots \sum_{i_k=1}^{d} {T_{i_1,\ldots,i_k}}^2}.$$

Moments of a distribution have the special property in that they are symmetric in the following sense.

Definition 2.11 (symmetric tensor). *A tensor* $T \in \mathbb{R}^{d^k}$ *is* symmetric *if the tensor is invariant to permutations of the indices, i.e.,*

$$T_{i_1 \cdots i_k} = T_{p(i_1 \cdots i_k)}$$

for any permutation p.

Notice that if a tensor is symmetric, then the n-mode product is independent of the mode, i.e., if $T \in \mathbb{R}^{d^k}$ is symmetric, then

$$T \times_n v = T \times_m v$$

for any $1 \leq n, m \leq k$. The next lemma shows that the tensor Frobenius norm has a particularly simple formula for rank-1 tensors.

Lemma 2.12. *Let* $v \in \mathbb{R}^d$ *and* k *be a positive integer. Then the tensor Frobenius norm of the* k*th-order tensor product is the same as the Euclidean norm of* v *raised to the* k*, i.e.,*

$$\|v^{\otimes k}\|_F = \|v\|^k.$$

The proof of Lemma 2.12 is included in Appendix A.

2.4. Tensor eigenvectors and normalized power iteration. The key to our approximate CP decomposition is rank-1 approximation that is based on tensor eigenvectors. These tensor eigenvectors can be found with the HOPM which we review in this section.

Definition 2.13 (tensor eigenvectors and eigenvalues). *Let* $T \in \mathbb{R}^{d^k}$ *be a symmetric tensor. Then* $v \in \mathbb{R}^d$ *is an* eigenvector *and* $\lambda \in \mathbb{R}$ *is the corresponding* eigenvalue *of* T *if*

$$((T \times_1 v) \times_1 v) \cdots \times_1 v) = \lambda v.$$

Note that since T is symmetric, the choice of n-mode product does not affect the definition of a tensor eigenvector. The next lemma shows that an eigenvalue-eigenvector pair provides a rank-1 approximation of a tensor in the Frobenius norm.

Lemma 2.14. *Let* T *be a* k*-order symmetric tensor with dimension* d*, i.e., let* $T \in \mathbb{R}^{d^k}$ *and* $v \in \mathbb{R}^d$ *be a unit length eigenvector of* T *with eigenvalue* $\lambda \neq 0$*. Then*

$$\|T - \lambda v^{\otimes k}\|_F^2 = \|T\|_F^2 - \lambda^2$$

and $\|T\|_F \geq \lambda$.

The proof of Lemma 2.14 can be found in Appendix A. It immediately follows from Lemma 2.14 that the eigenvector with the largest eigenvalue will achieve the best rank-1 approximation among the eigenpairs. In fact, it has been shown that the eigenpair with the largest eigenvalue achieves the best possible rank-1 approximation of the tensor [22, 6]. This fact will form the basis for an effective algorithm for finding an approximate CP decomposition in the next section.

Finally, an effective algorithm for finding the eigenvector associated to the largest eigenvalue in absolute value is the HOPM, originally developed in [5] and further analyzed in [29, 6]. In the case of symmetric tensors the symmetric-HOPM has a simpler form that is very similar to normalized power iteration but is not guaranteed to converge [22]. The HOPM algorithm for a symmetric order-k tensor $T \in \mathbb{R}^{d^k}$ requires initialization with the left singular vector, u, corresponding to the largest singular value of the unfolding (reshaping) of the tensor into a $d \times d^{k-1}$ matrix. The HOPM then defines k sequences of vectors, $v_0^{(1)}, \dots, v_0^{(k)}$, by initializing them all to be equal to u, $v_0^{(1)} = \cdots = v_0^{(k)} = u$, and inductively updating

$$w = T \times_1 v_{j+1}^{(1)} \times_1 \cdots \times_1 v_{j+1}^{(i-1)} \times_1 v_j^{(i+1)} \times_1 \cdots \times_1 v_j^{(k)}, \tag{2.1}$$
$$v_{j+1}^{(i)} = \frac{w}{||w||}$$

for each $i = 1, \dots, k$ and then increments j. Note that in formula (2.1), the subscripts do not represent the indices of the vector; they refer to the iteration, whereas in Algorithm 2.1 subscripts indicate vector indices.

Notice that the product that updates $v_{j+1}^{(i)}$ is the tensor T multiplied by the $k-1$ other vectors, leaving out $v_j^{(i)}$. Also note that we use the already updated $(j+1)$-step vectors for the first $i-1$ products and the j-step vectors for the last $k-i$ products. The HOPM is guaranteed to converge to an eigenvector of T [29], and when T is symmetric all $v_j^{(1)}, \dots, v_j^{(k)}$ converge to the same eigenvector but may differ in sign for even order tensors. For completeness we summarize the HOPM algorithm of [5] in Algorithm 2.1.

Algorithm 2.1 Higher order power method [5].

Inputs: A k-tensor $T \in \mathbb{R}^{d^k}$
Outputs: Eigenvector $v \in \mathbb{R}^d$ and eigenvalue λ such that $T \times_1 v \times_1 \cdots \times_1 v = \lambda v$

Reshape T into a $d \times d^{k-1}$ matrix and compute the leading left singular vector, v_0
Initialize $v^{(1)} = v^{(2)} = \cdots = v^{(k)} = u$, $\lambda = \text{Inf}$, and $\lambda_{\text{prev}} = 0$
while $|\lambda - \lambda_{\text{prev}}| > \text{tol}$ **do**
  for $\ell = 1, \dots, k$ **do**
    Set $v_s^{(\ell)} = \sum_{i_1,\dots,i_{\ell-1},i_{\ell+1},\dots,i_k=1}^{d} T_{i_1,\dots,i_{\ell-1},s,i_{\ell+1},\dots,i_k} v_{i_1}^{(1)} \cdots v_{i_{\ell-1}}^{(\ell-1)} v_{i_{\ell+1}}^{(\ell+1)} \cdots v_{i_k}^{(k)}$
    Set $v_s = \frac{v_s}{\|v_s\|}$
  end for
  Set $\lambda_{\text{prev}} = \lambda$
  Set $\lambda = \sum_{i_1,\dots,i_k=1}^{d} T_{i_1,\dots,i_k} v_{i_1}^{(1)} \cdots v_{i_k}^{(1)}$
end while
Set $v = v^{(1)}$
Return v, λ

Unlike the case of matrices, for tensors of order greater than two the basins of attraction for multiple distinct eigenvalues can have nonzero measure. It has been observed [6, 29, 22] that initialization with the left singular vector, u, of the tensor unfolding typically leads to

convergence to the eigenvector with the largest eigenvalue. The next section will rely on the ability to find the eigenpair associated to the largest eigenvalue (in absolute value) so a guaranteed way to find an initial condition in the basin of the largest eigenvalue is still an important problem for future research.

3. Approximate CP decomposition. In this section we show how tensor eigenvectors can be used to form an approximate CP decomposition up to an arbitrary level of precision. Of course, this is not a method of finding the minimal CP decomposition, the computation of which is NP-complete [9, 14]. Moreover, we do not even see an exact CP decomposition. Instead, given an order-k tensor T, we seek a sequence of vectors v_ℓ and constants λ_ℓ such that $\sum_{\ell=1}^{p} \lambda_\ell v_\ell^{\otimes k}$ approximates T in the Frobenius norm up to an error that can be made arbitrarily small by increasing p. In the next section we will show that this approximate CP decomposition is a key component for generalizing the unscented ensemble to higher moments.

Our approach is motivated by a theorem of [22] which states that if v is the unit length eigenvector of an order-k tensor T associated to the largest eigenvalue λ (in absolute value), then $\lambda v^{\otimes k}$ is the best rank-1 approximation of T, namely

$$\|T - \lambda v^{\otimes k}\|$$

is minimized over all possible λ, $\|v\| = 1$. It is well known that subtracting the best rank-1 approximation does not produce an *exact* CP decomposition, and in fact may increase tensor rank [31, 23]. However, it was suggested in [22] that repeatedly subtracting the rank-1 approximations may result in an *approximate* CP decomposition. The following theorem shows that this process converges subject to a certain tensor eigenvalue inequality that will be shown in Lemma 3.2 below.

Theorem 3.1. *Let T be a k-order symmetric tensor with size d, i.e., $T \in \mathbb{R}^{d^k}$. Consider the process of finding an approximate CP decomposition of T by starting from $T_0 = T$ and setting $T_{\ell+1} = T_\ell - \lambda_\ell v_\ell^{\otimes k}$ where λ_ℓ is the largest eigenvalue in absolute value of T_ℓ and v_ℓ is the associated eigenvector. Assume also that there exists a universal constant $c \in (0, 1]$ such that $\lambda_\ell \geq c|(T_\ell)_{i_1 \ldots i_k}|$. Then $\|T_\ell\|_F \to 0$ and for $r = \sqrt{1 - \frac{c^2}{d^k}} \in (0, 1)$*

$$\frac{\|T_{\ell+1}\|_F}{\|T_\ell\|_F} \leq r \quad \text{and} \quad T = \sum_{\ell=1}^{p} \lambda_\ell v_\ell^{\otimes k} + \mathcal{O}(r^L)$$

for all $L \in \mathbb{N}$.

Proof. First let λ_{maxabs} be the largest eigenvalue in absolute value of a tensor T and assume $\lambda_{maxabs} \geq c|T_{i_1 \ldots i_k}|$ for all $i_1, \ldots, i_k$. We will show that there exists a constant $c_2 = \frac{c}{d^{k/2}} \in (0, 1]$ such that $\lambda_{maxabs} \geq c_2 \|T\|_F$. Since $\lambda_{maxabs} \geq c|T_{i_1 \ldots i_k}|$, we have

$$\lambda_{maxabs}^2 \geq c^2 T_{i_1 \ldots i_k}^2,$$

which implies that

$$d^k \lambda_{maxabs}^2 \geq c^2 \sum_{i_1, \ldots, i_k} T_{i_1 \ldots i_k}^2,$$

so we have $d^{k/2}\lambda_{maxabs} \geq c\sqrt{\sum_{i_1,\ldots,i_k} T_{i_1\ldots i_k}^2}$ and

$$\lambda_{maxabs} \geq \frac{c}{d^{k/2}}\|T\|_F, \tag{3.1}$$

where we take $c_2 = \frac{c}{d^{k/2}} \in (0,1)$, since $c \in (0,1)$ and $d \geq 1$. By Lemma 2.14 applied to T_ℓ, we have

$$\|T_{\ell+1}\|_F{}^2 = \|T_\ell - \lambda_\ell v_\ell^{\otimes k}\|_F{}^2 = \|T_\ell\|_F{}^2 - \lambda_\ell{}^2.$$

Since λ_ℓ is defined to be the largest eigenvalue of T_ℓ, (3.1) says that $\lambda_\ell \geq c_2\|T_\ell\|_F$ where $c_2 = \frac{c}{d^{k/2}}$ so

$$\begin{aligned}\|T_{\ell+1}\|_F{}^2 &\leq \|T_\ell\|_F{}^2 - c_2{}^2\|T_\ell\|_F{}^2 \\ &\leq (1-c_2{}^2)\|T_\ell\|_F{}^2.\end{aligned}$$

Thus, setting $r = \sqrt{1-c_2{}^2} \in (0,1)$ we have $\|T_{\ell+1}\|_F \leq r\|T_\ell\|_F$ and $\|T_{\ell+1}\|_F \leq r^2\|T_{\ell-1}\|_F$ and so forth and proceeding inductively we find,

$$\|T_{\ell+1}\|_F \leq r^{\ell+1}\|T_0\|_F = r^{\ell+1}\|T\|_F.$$

Since $0 < r < 1$, $\lim_{\ell\to\infty} r^{\ell+1} = 0$, so $0 \leq \|T_{\ell+1}\|_F \leq r^{\ell+1}\|T\|_F \to 0$ implies $\|T_{\ell+1}\| \to 0$ as $\ell \to \infty$. Since this limit is 0, an upper bound on the rate of convergence of $\|T_\ell\|_F$ is found by considering

$$\frac{\|T_{\ell+1}\|_F}{\|T_\ell\|_F} \leq r = \sqrt{1-\frac{c^2}{d^k}}.$$

■

Theorem 3.1 gives an effective algorithm for finding approximate CP decompositions of tensors; however, it requires an inequality of the form

$$\lambda_{maxabs} \geq c|T_{i_1,\ldots,i_k}|. \tag{3.2}$$

The inequality (3.2) holds for symmetric matrices with $c = 1$, since if $T \in \mathbb{R}^{d^2}$ is symmetric it has an orthogonal eigendecomposition $T = U^\top \Lambda U$ so by the Cauchy–Schwarz inequality,

$$|T_{ij}| = |\langle u_i, \lambda_j u_j\rangle| \leq ||u_i||\,||\lambda_j u_j|| = |\lambda_j| \leq \lambda_{maxabs}$$

and the identity matrix shows that $c = 1$ is the best possible constant for matrices. Of course, this method of proof cannot be generalized to arbitrary tensors due to the lack of a similar rank-1 eigendecomposition. Nevertheless, the next lemma shows that an inequality of the form (3.2) does hold for all symmetric tensors of orders 3 and 4.

Lemma 3.2. *If T is a symmetric 3-tensor, then*

$$\lambda_{maxabs} \geq \frac{2}{3+4\sqrt{2}+\sqrt{3}}|T_{ijk}|.$$

If T is a symmetric 4-tensor, then

$$\lambda_{maxabs} \geq \frac{6}{323}|T_{ijk\ell}|.$$

The proof of Lemma 3.2 is quite involved and can be found in the appendix. We conjecture that such an inequality holds for symmetric tensors of any order, and we note that the constants in Lemma 3.2 are not known to be sharp. For the purposes of this paper, we are focused on matching the skewness and kurtosis of a distribution so we only need the approximate CP decomposition for tensors up to order 4. In the next section we will show how to use the approximate CP decomposition to build an ensemble that simultaneously matches the mean, covariance, skewness, and kurtosis.

We summarize the approximate CP decomposition algorithm below.

Algorithm 3.1 Approximate CP decomposition.

Inputs: A k-tensor $T \in \mathbb{R}^{d^k}$ and a tolerance τ.

Outputs: Vectors, v_ℓ, and signs, $s_\ell \in \{-1, 1\}$ such that $\left|\left|\sum_{\ell=1}^{p} s_\ell v_\ell^{\otimes k} - T\right|\right|_F \leq \tau$.

Set $\ell = 1$
while $||T||_F > \tau$ **do**
  Apply the HOPM (Algorithm 2.1) to find an eigenpair (v, λ) of T.
  Set $s_\ell = \text{sign}(\lambda)$ (note that if k is odd we can always choose $s_\ell = 1$)
  Set $v_\ell = |\lambda|^{1/k} v$
  Set $T = T - s_\ell v_\ell^{\otimes k}$
  Set $\ell = \ell + 1$
end while
Return the set of all s_ℓ, v_ℓ

Finally, we demonstrate this algorithm on a random 3-tensor and 4-tensor with $d = 2$ and $d = 10$ in Figure 2. We note that in all cases the convergence is much faster than our theoretical upper bound; however, for $d = 10$ we see that the ratio of residual norms approaches much closer to our upper bound. Moreover, high dimensional tensors require a much larger number of vectors to achieve a given tolerance with the approximate CP decomposition. So while our approach provides an effective solution, it is likely that there is room for improvement, and the HOUT introduced in the next section can use any method of CP decomposition.

4. Higher order unscented transform.

The goal of the SUT is to generate a small ensemble that exactly matches the mean and covariance of a distribution, thus forming a quadrature rule that can be used to estimate the expected value of nonlinear functions. In this section we define the HOUT which matches the first four moments of a distribution, thus providing a quadrature rule with a higher degree of exactness. While we only describe the process explicitly for up to four moments, our method is based on the approximate tensor decomposition from the previous section and should allow generalization to an arbitrary number of moments.

Suppose we are given the following moments of the distribution of a random variable: the mean $\mu \in \mathbb{R}^d$, the covariance matrix $C \in \mathbb{R}^{d \times d}$, the skewness tensor $S \in \mathbb{R}^{d \times d \times d}$, and kurtosis tensor $K \in \mathbb{R}^{d \times d \times d \times d}$. Let τ be a parameter that specifies the tolerance of the approximate CP decompositions and let S and K have the approximate CP decompositions

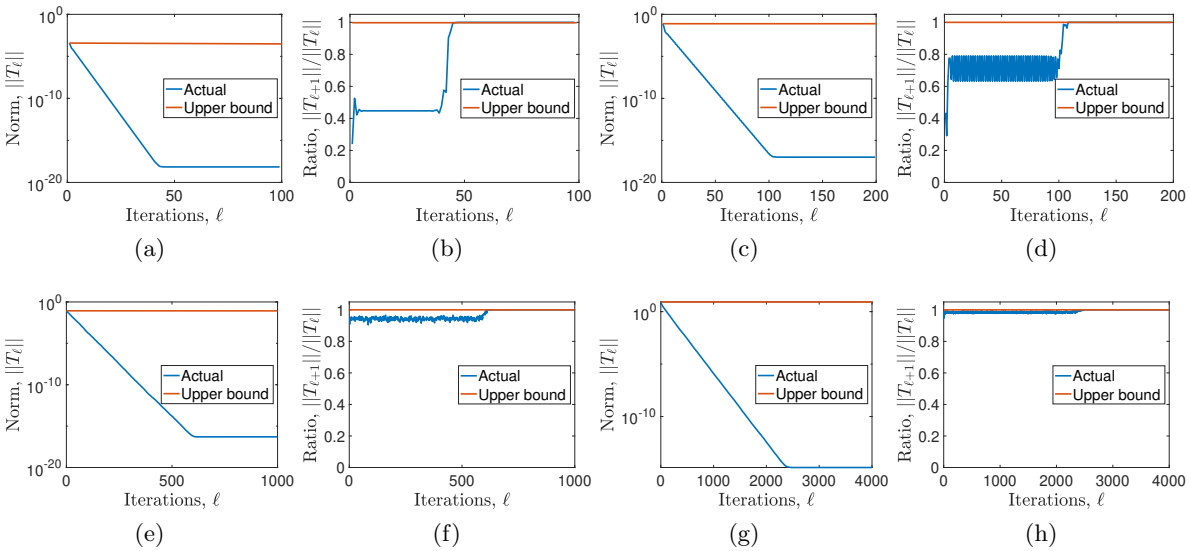


Figure 2. (a)–(d) *With $d = 2$ we demonstrate the approximate CP decomposition on a random* 3*-tensor* (a), (b) *and* 4*-tensor* (c), (d). *The norm of the residual* (a), (c) *(blue) decays to numerical zero faster than the theoretical upper bound, r^ℓ (red). The ratio of successive Frobenius norms* (b), (d) *(blue) is always less than the derived upper bound, r (red).* (e)–(h) *We repeat the experiment with $d = 10$.*

$$\left|\left|S - \sum_{i=1}^{J} \tilde{v}_i^{\otimes 3}\right|\right|_F \leq \tau/2, \qquad \left|\left|K - \sum_{i=1}^{L} s_i \tilde{u}_i^{\otimes 4}\right|\right|_F \leq \tau/2,$$

where $s_i \in \{-1, 1\}$ denote signs. Note that these approximate decompositions can be constructed by the algorithm described in Theorem 3.1 by moving the eigenvalues inside the tensor power by the rule $(cv)^{\otimes k} = c^k v^{\otimes k}$. Note that the signs s_i are required for the kurtosis since constants come out of even order tensor powers as absolute values.

The key to forming an ensemble that matches all four moments simultaneously is carefully balancing the interactions between the moments. For example, if we add new quadrature nodes of the form $\mu + \gamma\tilde{v}_i$ in order to try to match the skewness, these nodes will influence the mean of the ensemble. In order to balance these interactions we make the following definitions based on the approximate CP decompositions of the skewness and kurtosis:

$$\tilde{\mu} = \sum_{i=1}^{J} \tilde{v}_i, \qquad \hat{\mu} = -\gamma^{-2}\tilde{\mu}, \qquad \tilde{C} = \sum_{i=1}^{L} s_i \tilde{u}_i^{\otimes 2}, \qquad \hat{C} = C - \frac{1}{\delta^2}\tilde{C},$$

where $\hat{L} = \sum_{i=1}^{L} s_i$ and β, γ, δ are arbitrary positive constants that will define the 4-moment σ-points below. We note that C is assumed symmetric and positive definite since it is a covariance matrix and $\tilde{C}$ is symmetric by definition. In order to ensure that $\hat{C}$ is also positive definite, let $\lambda_{\max}^{\tilde{C}}$ be the largest eigenvalue of $\tilde{C}$ and let $\lambda_{\min}^{C}$ be the smallest eigenvalue of C; then we require that $\delta > \sqrt{\frac{\lambda_{\max}^{\tilde{C}}}{\lambda_{\min}^{C}}}$ which guarantees that $\hat{C}$ is positive definite. We note that this choice can be overly conservative especially when C is close to rank deficient. In these cases, it can be helpful to iteratively divide δ by 2 as long as $\hat{C}$ remains positive definite.

These choices balance out the interactions between the moments and are the key to proving Theorem 4.2 below. We are now ready to define the 4-moment σ-points.

Definition 4.1 (the 4-moment σ-points of the higher order unscented transform). *Let α, β, γ, δ be positive real numbers; we define the* 4-moment σ-points *by*

$$\sigma_i = \begin{cases} \mu & \text{if } i = -2, \\ \mu + \alpha\hat{\mu} & \text{if } i = -1, \\ \mu - \alpha\hat{\mu} & \text{if } i = 0, \\ \mu + \beta\sqrt{\hat{C}_i} & \text{if } i = 1, \dots, d, \\ \mu - \beta\sqrt{\hat{C}_{i-d}} & \text{if } i = d+1, \dots, 2d, \\ \mu + \gamma\tilde{v}_{i-2d} & \text{if } i = 2d+1, \dots, 2d+J, \\ \mu - \gamma\tilde{v}_{i-2d-J} & \text{if } i = 2d+J+1, \dots, 2d+2J, \\ \mu + \delta\tilde{u}_{i-2d-2J} & \text{if } i = 2d+2J+1, \dots, 2d+2J+L, \\ \mu - \delta\tilde{u}_{i-2d-2J-L} & \text{if } i = 2d+2J+L+1, \dots, N \end{cases}$$

and the corresponding weights *by*

$$w_i = \begin{cases} 1 - d\beta^{-2} - \hat{L}\delta^{-4} & \text{if } i = -2, \\ \frac{1}{2}\alpha^{-1} & \text{if } i = -1, \\ -\frac{1}{2}\alpha^{-1} & \text{if } i = 0, \\ \frac{1}{2}\beta^{-2} & \text{if } i = 1, \dots, 2d, \\ \frac{1}{2}\gamma^{-3} & \text{if } i = 2d+1, \dots, 2d+J, \\ -\frac{1}{2}\gamma^{-3} & \text{if } i = 2d+J+1, \dots, 2d+2J, \\ \frac{1}{2}\delta^{-4}s_{i-2d-2J} & \text{if } i = 2d+2J+1, \dots, 2d+2J+L, \\ \frac{1}{2}\delta^{-4}s_{i-2d-2J-L} & \text{if } i = 2d+2J+L+1, \dots, N. \end{cases}$$

For convenience, denote $N = 2(d + J + L)$.

The next theorem shows that the 4-moment σ-points match the first two moments exactly and match the skewness and kurtosis up to an error term that can be controlled below.

Theorem 4.2. *Given the 4-moment σ-points associated with μ, C, S, and K we have $\sum_{i=-2}^{N} w_i = 1$ and*

$$\sum_{i=-2}^{N} w_i\sigma_i = \mu,$$

$$\sum_{i=-2}^{N} w_i(\sigma_i - \mu)^{\otimes 2} = C,$$

$$\left|\left|\sum_{i=-2}^{N} w_i(\sigma_i - \mu)^{\otimes 3} - S\right|\right|_F \leq \tau/2 + \alpha^2 \left|\left|\hat{\mu}^{\otimes 3}\right|\right|_F,$$
$$\left|\left|\sum_{i=-2}^{N} w_i(\sigma_i - \mu)^{\otimes 4} - K\right|\right|_F \leq \tau/2 + \beta^2 \left|\left|\bar{C}\right|\right|_F,$$

where $\bar{C} = \sum_{i=1}^{d} \sqrt{\hat{C}_i}^{\otimes 4}$.

The proof can be found in the appendix. Notice that the third and fourth moment equations do not exactly match the skewness and kurtosis, respectively. Of course, we only used an approximate CP decomposition to begin with, which accounts for the τ term in the error. Thus, the real goal is to bound the other error term by the same tolerance, τ. The following corollary shows how to control the error terms on the skewness and kurtosis.

Corollary 4.3. *Let τ be a specified tolerance for the absolute error of the skewness and kurtosis and set $\bar{C} = \sum_{i=1}^{d} \sqrt{\hat{C}_i}^{\otimes 4}$ and $\hat{\mu}$ as in Theorem 4.2. If we choose parameters α, β such that*

$$\alpha < \sqrt{\frac{\tau}{2||\hat{\mu}^{\otimes 3}||_F}} \qquad \text{and} \qquad \beta < \sqrt{\frac{\tau}{2||\overline{C}||_F}},$$

then

$$\left|\left|\sum_{i=-2}^{N} w_i(\sigma_i - \mu)^{\otimes 3} - S\right|\right|_F < \tau \qquad \text{and} \qquad \left|\left|\sum_{i=-2}^{N} w_i(\sigma_i - \mu)^{\otimes 4} - K\right|\right|_F < \tau.$$

Proof. The inequality for β follows immediately from Theorem 4.2. Once β is chosen, then we can define

$$||\hat{\mu}^{\otimes 3}||_F = \left|\left|\left(\left(1 - d\beta^{-2} - \hat{L}\delta^{-4}\right)\mu - \gamma^{-2}\tilde{\mu}\right)^{\otimes 3}\right|\right|_F$$

and choosing $\alpha < \sqrt{\frac{\tau}{2||\hat{\mu}^{\otimes 3}||_F}}$ we have

$$\left|\left|\sum_{i=-2}^{N} w_i(\sigma_i - \mu)^{\otimes 3} - S\right|\right|_F \leq \tau/2 + \alpha^2 ||\hat{\mu}^{\otimes 3}||_F < \tau$$

as desired. ∎

Corollary 4.3 could easily be reformulated to control relative error if desired, and taken to the extreme we could make the quadrature rule exact up to numerical precision. As a practical matter, this is not an effective strategy since it would result in a larger condition number for the numerical quadrature as shown in the following remark.

Algorithm 4.1 Higher order unscented transform.

Inputs: A function f, tolerance τ, and the mean, μ, covariance, C, skewness, S, and kurtosis, K, of a random variable X.

Outputs: Estimate of $\mathbb{E}[f(X)]$ with degree of exactness 4.

Compute the approximate CP decomposition $\left|\left|S - \sum_{i=1}^{J} \tilde{v}_i^{\otimes 3}\right|\right|_F \leq \tau/2$

Compute the approximate CP decomposition $\left|\left|K - \sum_{i=1}^{L} s_i \tilde{u}_i^{\otimes 4}\right|\right|_F \leq \tau/2$

Set $\tilde{C} = \sum_{i=1}^{L} s_i \tilde{u}_i^{\otimes 2}$.

Compute the largest eigenvalue $\lambda_{\max}^{\tilde{C}}$ of $\tilde{C}$ and the smallest eigenvalue $\lambda_{\min}^{C}$ of C

Choose $\delta > \sqrt{\frac{\lambda_{\max}^{\tilde{C}}}{\lambda_{\min}^{C}}}$ (note that C is positive definite so $\lambda_{\min}^{C} > 0$)

(Optional) While $C - \delta^{-2}\tilde{C}$ is positive definite, set $\delta = \delta/2$

Set $\hat{C} = C - \delta^{-2}\tilde{C}$

Compute the symmetric square root of $\hat{C}$ with columns $\sqrt{\hat{C}_i}$

Set $\bar{C} = \sum_{i=1}^{d} \sqrt{\hat{C}_i}^{\otimes 4}$

Choose $\beta < \sqrt{\frac{\tau}{2||\bar{C}||_F}}$ and choose $\gamma > 0$ (default $\gamma = J^{-1/3}$)

Set $\hat{L} = \sum_{i=1}^{L} s_i$ and $\tilde{\mu} = \sum_{i=1}^{J} \tilde{v}_i$ and $\hat{\mu} = (1 - d\beta^{-2} - \hat{L}\delta^{-4})\mu - \gamma^{-2}\tilde{\mu}$

Choose $\alpha < \sqrt{\frac{\tau}{2||\hat{\mu}^{\otimes 3}||_F}}$

Define the 4-moment σ-points, σ_i, and weights, w_i, according to Definition 4.1

Output: $\sum_{i=-2}^{N} w_i f(\sigma_i)$

Remark 4.4. The absolute condition number of the HOUT is bounded above by $\sum_{i=0}^{N} |w_i|$. Using the bounds from Corollary 4.3 we find

$$\sum_{i=0}^{N} |w_i| = \frac{1}{\alpha} + \frac{d}{\beta^2} + \frac{J}{\gamma^3} + \frac{L}{\delta^4} > \sqrt{\frac{||\bar{\mu}^{\otimes 3}||_F}{\tau}} + \frac{d||\bar{C}||_F}{\tau} + \frac{J}{\gamma^3} + \frac{L}{\delta^4} = \mathcal{O}(\tau^{-1}),$$

which shows that the condition number has the potential to blow up as the tolerance is decreased.

We summarize the HOUT algorithm in Algorithm 4.1 and we now turn to some numerical experiments to demonstrate the HOUT.

5. Error analysis.

The standard approach to error estimates for the SUT is based on a Taylor's theorem approximation near the mean. These results can be immediately generalized to the HOUT as in the following theorem (see the proof in Appendix D).

Theorem 5.1 (Taylor-type HOUT error bound). *Let $f \in C^5(\mathbb{R}^d, \mathbb{R})$ and let $X \sim p$ be a random variable with distribution p that has compact support. Then the error in estimating $\mathbb{E}[f(X)]$ using the 4-moment σ-points of the HOUT and corresponding weights has the upper bound*

$$\left|\mathbb{E}[f(X)] - \sum_{i=1}^{m} w_i f(\sigma_i)\right| \leq ||D^5 f||_\infty \frac{d^5}{120}\left(||M_{5,abs}||_{\max} + ||\tilde{M}_{5,abs}||_{\max}\right),$$

where the $||D^5 f||_\infty$ *is taken on the support of the measure and* $M_{5,abs}, \tilde{M}_{5,abs}$ *are the absolute fifth moments of* p *and the quadrature, respectively (see Appendix* D *for details).*

While the assumption of compact support is not strictly necessary, one must make some assumption on the decay of the probability measure in order to control the error. Moreover, the Taylor's theorem approach does not allow less regular functions f or take advantage of additional regularity that may be present in f. Thus we take a more general approach based on the methods of polynomial approximation [3, 28].

The benefit of our ability to match four moments to arbitrary precision is that it allows us to apply the standard approach for quadrature error analysis based on polynomial approximation. In this section, we develop error bounds in the context of a quadrature that matches n moments. Since the HOUT matches four moments, the bounds developed in this section apply to the HOUT with $n = 4$. Of course, this immediately requires an assumption on the probability measure dp that the first n-moments exist. However, we will not require the existence of a density or any regularity assumptions on the measure.

The polynomials $1, x, (x-\mu)^{\otimes 2}, \ldots, (x-\mu)^{\otimes n}$ form a basis for the space of degree n polynomials in the components of $x \in \mathbb{R}^d$, denoted Π_n^d. Since expectations are linear, a quadrature which is exact on these basis polynomials will be exact for all polynomials of degree less than or equal to n, namely, $\mathbb{E}[q] = \sum_{i=1}^m w_i q(\sigma_i)$ for any $q \in \Pi_n^d$. Of course, the quadrature may only be accurate up to a threshold and in finite precision arithmetic it cannot be exact. Moreover, the moments that the quadrature is matching may only be estimates of the true moments. To understand the propagation of such errors, we write the polynomial $q(x) = \sum_{s=0}^n \sum_{j_1,\ldots,j_s=1}^d a_{j_1\cdots j_s}(x-\mu)^{\otimes s}_{j_1\cdots j_s}$ in the basis of moments. Note that

$$\begin{aligned} E_{\text{moments}} &\equiv \left|\mathbb{E}[q] - \sum_{i=1}^m w_i q(\sigma_i)\right| \\ &= \left|\sum_{s=0}^n \sum_{j_1,\ldots,j_s=1}^d a_{j_1\cdots j_s}\left(\mathbb{E}[(x-\mu)^{\otimes s}_{j_1\cdots j_s}] - \sum_{i=1}^m w_i(\sigma_i-\mu)^{\otimes s}_{j_1\cdots j_s}\right)\right| \\ &\leq c(q)\sum_{s=0}^n ||M_s - \tilde{M}_s||_{\max}, \end{aligned}$$

where $c(q)$ is a constant depending only on the polynomial q and $M_s = \mathbb{E}[(x-\mu)^{\otimes s}]$ are the true moments and $\tilde{M}_s = \sum_{i=1}^m w_i(\sigma_i - \mu)^{\otimes s}$ are the moments matched by the algorithm.

Whenever we approximate a function f by a polynomial $q \in \Pi_n^d$, we should expect unbounded errors as the inputs approach infinity. Thus, in order to control the error on $\mathbb{E}[f]$ by polynomial approximation, we need to split the domain into the interior and exterior of a ball $\mathbb{B}_r(\mu)$ of radius r centered on μ. Outside the ball we define the error by

$$E_{\text{outside}} \equiv \int_{\mathbb{R}^d \cap \mathbb{B}_r(\mu)^c} |f - q|\, dp,$$

and bounding this error requires assuming that the probability measure decays sufficiently fast to control the error between f and q. Inside the ball we define the polynomial approximation error by

$$E_{\text{inside}} \equiv ||f - q||_\infty = \sup_{x \in \mathbb{B}_r(\mu)} |f(x) - q(x)|$$

and bounding this error will require an appropriate regularity assumption on f.

By combining these error terms, we can control the error of a quadrature formula on any function f by any polynomial q of degree n, namely,

$$\begin{aligned}
E_{\text{total}} &\equiv \left| \mathbb{E}[f] - \sum_{i=1}^m w_i f(\sigma_i) \right| \\
&\le |\mathbb{E}[f] - \mathbb{E}[q]| + E_{\text{moments}} + \left| \sum_{i=1}^m w_i q(\sigma_i) - \sum_{i=1}^m w_i f(\sigma_i) \right| \\
&\le E_{\text{moments}} + \int_{\mathbb{R}^d} |f - q| \, dp + \sum_{i=1}^m w_i |f(\sigma_i) - q(\sigma_i)| \\
&\le E_{\text{moments}} + \int_{\mathbb{R}^d \cap \mathbb{B}_r(\mu)^c} |f - q| \, dp + \int_{\mathbb{B}_r(\mu)} ||f - q||_\infty \, dp + \sum_{i=1}^m w_i ||f - q||_\infty \\
&\le E_{\text{moments}} + E_{\text{outside}} + 2E_{\text{inside}},
\end{aligned}$$

where we assume that r is sufficiently large that $\sigma_i \in \mathbb{B}_r(\mu)$ for all $i = 1, \dots, m$. Notice that the three error terms all depend on the choice of the polynomial q, and since the inequality holds for all $q \in \Pi_n^d$ we can write

$$\left| \mathbb{E}[f] - \sum_{i=1}^m w_i f(\sigma_i) \right| \le \inf_{q \in \Pi_n^d} \{E_{\text{moments}} + E_{\text{outside}} + 2E_{\text{inside}}\}.$$

From this general framework, many potential results can be derived depending on the localization of the probability measure and the regularity of f. If we assume that the moments are exactly approximated, then one such result would be the following theorem.

Theorem 5.2 (general HOUT error bound). *Let $f \in C^k(\mathbb{R}^d, \mathbb{R})$ be bounded in absolute value by a polynomial, $|f(x)| \le a + b||x||^t$. Let x be a random variable with probability density $p(x) < ce^{-\alpha||x-\mu||^\beta}$ for some $\alpha, \beta > 0$ and all $||x - \mu|| > r_0$. Let $Q(f) \equiv \sum_{i=1}^m w_i f(\sigma_i)$ be exact on the first n moments of p. For any radius $r \ge r_0$ such that $\sigma_i \in \mathbb{B}_r(\mu)$ we have*

$$|\mathbb{E}[f] - Q(f)| \le c_1 \left(\frac{r}{n}\right)^k \left(\frac{||D^k f||_\infty}{n} + \sum_{|\gamma|=k} \sup_{|x-y|<\frac{1}{n}} |D_\gamma^k f(x) - D_\gamma^k f(y)| \right) + c_2 \, n \, r^{t+n+d-\beta} e^{-\alpha r^\beta},$$

where c_1 depends on k, d and c_2 depends on a, b, α, β.

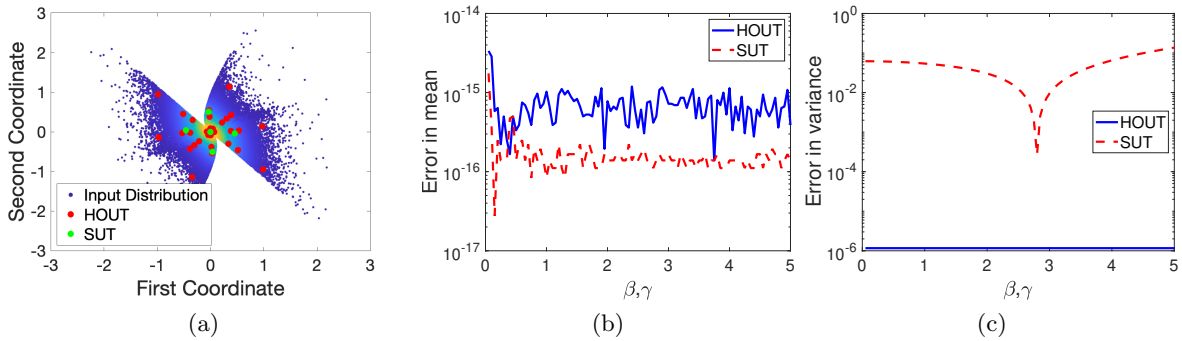
The proof of Theorem 5.2 is included in Appendix D and follows from upper bounds on the error of the multivariate polynomial of best approximation found in [3, 28] together with bounds on the integrals of polynomials multiplied by an exponential.

Of course, the HOUT currently has only been derived for $n = 4$; however, we chose to derive the general error bounds to show how matching more moments can potentially improve the estimation in the future.

6. Numerical experiments. We first compare the HOUT and SUT on various polynomials applied to a two-dimensional input distribution. In order to generate a non-Gaussian input distribution, we start by generating an ensemble of 10^5 standard Gaussian random variables, $Z \in \mathbb{R}^2$, and then transforming them by a map $X = AZ + B(Z \odot Z \odot \text{sign}(Z))$ where A, B are random 2×2 matrices with entries chosen from a Gaussian distribution with mean 0 and standard deviation $1/10$ and $\odot$ is componentwise multiplication. The resulting ensemble is shown in Figure 3(a) along with the HOUT (red dots) and SUT (green dots) ensembles.

The SUT has the free parameter β but the HOUT requires a certain inequality for β and instead the HOUT has γ as a free parameter. In order to explore the effect of these parameters on the SUT and HOUT, we considered a random quadratic polynomial $f : \mathbb{R}^2 \to \mathbb{R}$. In Figure 3 we show the error of the HOUT and SUT estimates of the mean $\mathbb{E}[f(X)]$ and variance $\mathbb{E}[(f(X) - \mathbb{E}[f(X)])^2]$ as a function of β for the SUT and γ for the HOUT. Notice that since f is a quadratic polynomial, the mean is also a quadratic polynomial, whereas the variance is a quartic polynomial. Since the SUT has degree of exactness two, it is exact on the mean but not on the variance. The HOUT has degree of exactness four and is exact on both up to the specified tolerance (10^{-5} in these experiments). Reducing the tolerance below this point led to increased error, most likely due to the conditioning of the HOUT quadrature rule.

Using the same two-dimensional distribution, X, we passed it through several polynomial functions of the form $f(x) = ax + bcx^n$ for $n = 2, 3, 4, 5$ where a and b are made random 1×2 vectors. To show the influence of the strength of the nonlinearity, we sweep through different values of c. In Figure 4 we compare the HOUT and SUT for estimating the mean and variance of the output of each of these polynomials. As expected, the HOUT is exact for the means up to $n = 4$ and for the variances up to $n = 2$ due to having degree of exactness four. For higher degree polynomials, the HOUT has comparable or better performance. Whenever the



(a) (b) (c)

Figure 3. (a) *Comparison between the HOUT (red dots) and the SUT ensemble (green dots) on a non-Gaussian distribution. Note that the SUT uses* 5 σ*-points while the HOUT uses* 69 σ*-points.* (b), (c) *Estimating the output mean and covariance for various values of* β *in the SUT and various values of* γ *in the HOUT.*

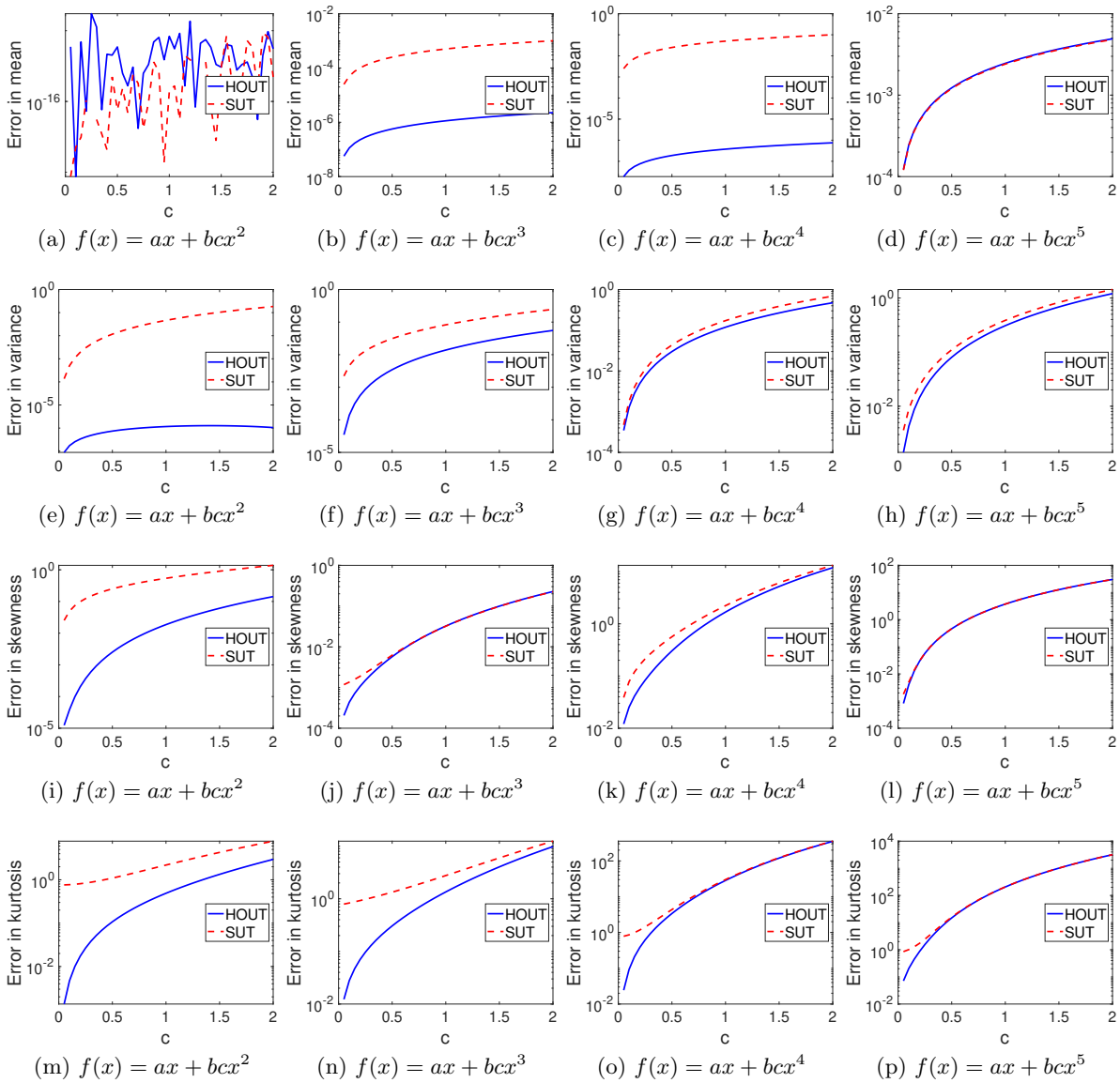


Figure 4. *Comparison between the HOUT and the SUT when estimating the mean (top row), variance (second row), skewness (third row), and kurtosis (bottom row) with different polynomials. Notice that the SUT has degree of exactness two, while the HOUT has degree of exactness four.*

nonlinearity is not too strong, such as when c is small and/or the power n is small, the HOUT has a big advantage. However, for some strong nonlinearities when c and the power n is large then the HOUT and SUT may have similar performance.

Of course, the HOUT and SUT are intended for use beyond polynomial functions. In fact, the most common application is for forecasting dynamical systems. Next, we consider the problem of forecasting the chaotic Lorenz-63 dynamical system [26]. We integrate the Lorenz-63 system with a Runge–Kutta order 4 method and a time step $\tau = 0.1$. In order to

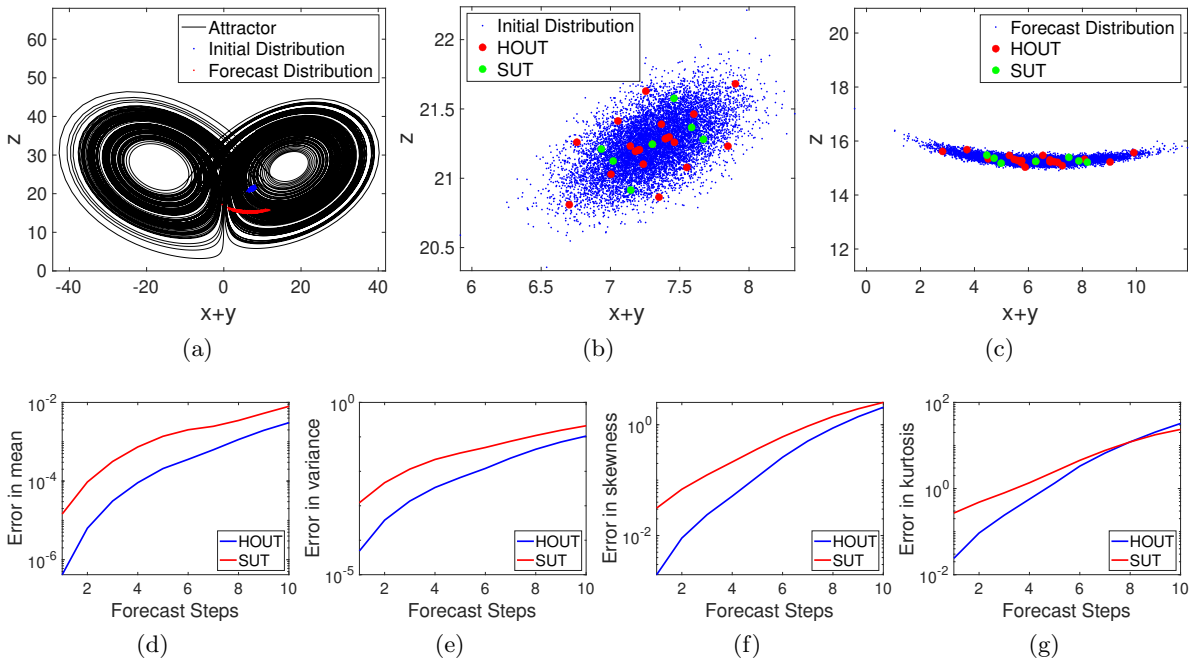


Figure 5. *Comparison between the HOUT and the SUT when estimating the mean* $\mathbb{E}[f(X)]$ *(top row) and higher moments of the Lorenz-63 model at various forecast horizons. In* (a) *we show the Lorenz-63 attractor (black) along with an example initial ensemble (blue) and forecast ensemble (red) used to compute the true statistics. In* (b), (c) *we show the initial and forecast ensembles (blue) together with the HOUT (red) and SUT (green) ensembles. Results in* (d)–(g) *show the forecast accuracy versus the forecast steps and are geometrically averaged over* 500 *different initial conditions on the attractor.*

generate a non-Gaussian initial state, we start by choosing a random point on the attractor and adding a small amount of Gaussian noise. We then run the ensemble forward $N_1 = 5$ steps and we consider this the initial state; see Figures 5(a) (blue) and 5(b) (blue). We compute the statistics of the initial state using the ensemble shown and use these statistics to generate the HOUT and SUT as shown in Figure 5(b). All three ensembles are then integrated forward in time N_2 additional steps and the true forecast statistics from the large ensemble are compared to the HOUT and SUT estimates. An example is shown in Figure 5(c) with $N_2 = 15$.

We then repeat this experiment 500 times with different randomly selected initial states on the attractor and we compute the geometric average of the error between the HOUT estimate and the true statistics at each forecast time, shown in Figures 5(d)–(g) (blue). Similarly, we compute the geometric average of the error between the SUT estimate and the true statistics (red) at each forecast time, shown in Figures 5(d)–(g) (red). We note that the HOUT provides improved estimates of the first four moments up to at least four forecast steps, which is 0.4 model time units. In particular, the mean forecast is improved by an order of magnitude in this forecast range.

7. Conclusions and future directions.

The SUT is a highly efficient and successful strategy for uncertainty quantification with many applications. As computational resources expand, there is a growing demand for larger ensembles that are similarly well designed. At the

same time, complex systems demand better uncertainty quantification such as skewness and kurtosis to capture fat-tails. The HOUT generalizes the SUT to efficiently leverage additional computation resources to meet the growing uncertainty quantification demand. There are several promising directions of future research that we expect to result from this work.

First, there are many applications, such as Kalman filtering and smoothing for nonlinear systems, that use the SUT to track the mean and covariance of hidden variables based on noisy observations. If these filters and smoothers can be generalized to track four moments, they could be integrated with the HOUT to achieve better stability and accuracy along with additional uncertainty quantification. Moreover, these methods are often difficult to analyze theoretically due to the lack of a natural limit. This compares to the relative ease of theoretical analysis of particle filters where one may consider the infinite particle (Monte Carlo) limit. The HOUT opens up the possibility that generalized Kalman-based approaches may be analyzed in the limit of infinitely many moments; in a sense this is a kind of spectral solver convergence. Of course, as the number of moments increases, the size of the ensemble required and the computational complexity will also increase. While we only explicitly derive the 4-moment version of the HOUT, the methods used should allow generalization to match an arbitrary number of moments. Such a convergence result would finally allow moment-based methods to be compared to particle filters, where one can show convergence as the number of particles goes to infinity.

A second promising direction for future research concerns deriving error bounds for the SUT and HOUT. Current error bounds for the SUT are based on Taylor expansion [18, 21, 19], and a similar analysis could be carried out for the HOUT. However, this analysis requires decay of the moments and a highly localized input density; moreover it is not the natural method of analyzing quadrature error. A more natural approach would be based on multivariate polynomial approximation error bounds, which would be analogous to the univariate quadrature error bound analysis.

Finally, a more efficient CP decomposition can immediately improve the efficiency of the HOUT. Similarly, improved/sharp bounds on the relationship between tensor eigenvalues and their entries could improve understanding of the convergence rate as a function of dimension and tensor order.

Appendix A. Tensor multiplication and proofs of Lemmas 2.12 and 2.14. To discuss how tensor multiplication works, let us first look at the simplest case where we multiply a 2-tensor with a 1-tensor. Recall that for a matrix $A \in \mathbb{R}^{d\times d}$ and $v \in \mathbb{R}^d$, the matrix-vector multiplication Av is given by $(Av)_i = \sum_{j=1}^d A_{ij}v_j$ so we define two natural tensor-vector products

$$(A \times_1 v)_i = \sum_{j=1}^d A_{ji}v_j = (A^\top v)_i \qquad \text{and} \qquad (A \times_2 v)_i = \sum_{j=1}^d A_{ij}v_j = (Av)_i.$$

Analogously, for a 3-tensor $S \in \mathbb{R}^{d\times d\times d}$ and a vector $v \in \mathbb{R}^d$, the tensor-vector multiplication is carried out as follows, each case resulting in a $d \times d$ matrix:

$$(S \times_1 v)_{ik} = \sum_{j=1}^d S_{jik}v_j, \qquad (S \times_2 v)_{ik} = \sum_{j=1}^d S_{ijk}v_j, \qquad (S \times_3 v)_{ik} = \sum_{j=1}^d S_{ikj}v_j.$$

For example, if $S \in \mathbb{R}^{3\times3\times3}$ and $v \in \mathbb{R}^3$ such that

$$S = \left[\begin{array}{ccc} S_{111} & S_{121} & S_{131} \\ S_{211} & S_{221} & S_{231} \\ S_{311} & S_{321} & S_{331} \end{array}\right] \left[\begin{array}{ccc} S_{112} & S_{122} & S_{132} \\ S_{212} & S_{222} & S_{232} \\ S_{312} & S_{322} & S_{332} \end{array}\right] \left[\begin{array}{ccc} S_{113} & S_{123} & S_{133} \\ S_{213} & S_{223} & S_{233} \\ S_{313} & S_{323} & S_{333} \end{array}\right],$$

then

$$S \times_1 v = \begin{bmatrix} S_{111}v_1 + S_{211}v_2 + S_{311}v_3 & S_{112}v_1 + S_{212}v_2 + S_{312}v_3 & S_{113}v_1 + S_{213}v_2 + S_{313}v_3 \\ S_{121}v_1 + S_{221}v_2 + S_{321}v_3 & S_{122}v_1 + S_{222}v_2 + S_{322}v_3 & S_{123}v_1 + S_{223}v_2 + S_{323}v_3 \\ S_{131}v_1 + S_{231}v_2 + S_{331}v_3 & S_{132}v_1 + S_{232}v_2 + S_{332}v_3 & S_{133}v_1 + S_{233}v_2 + S_{333}v_3 \end{bmatrix}.$$

Generalizing this to arbitrary order tensors yields Definition 2.9.

We now turn to the proof of Lemma 2.12.

Proof. By the definition of the tensor Frobenius norm,

$$\|v^{\otimes k}\|_F^2 = \sum_{i_1=1}^{d} \cdots \sum_{i_k=1}^{d} [(v^{\otimes k})_{i_1,\ldots,i_k}]^2$$

and since $(v^{\otimes k})_{i_1\ldots i_k} = v_{i_1}v_{i_2}\cdots v_{i_k}$, we have $\|v^{\otimes k}\|_F^2 = \sum_{i_1=1}^{d}\cdots\sum_{i_k=1}^{d} v_{i_1}^2\cdots v_{i_k}^2$, so

$$\|v^{\otimes k}\|_F^2 = \sum_{i_1=1}^{d} v_{i_1}^2 \sum_{i_2=1}^{d} v_{i_2}^2 \cdots \sum_{i_k=1}^{d} v_{i_k}^2 = \underbrace{\|v\|^2\|v\|^2\cdots\|v\|^2}_{k \text{ times}}$$

by definition of $\|v\|$ so

$$\|v^{\otimes k}\|_F = \|v\|^k.$$

Finally, we include the proof of Lemma 2.14.

Proof. We first wish to show that $\|T - \lambda v^{\otimes k}\|_F^2 = \|T\|_F^2 - \lambda^2$.

$$\begin{aligned} \|T - \lambda v^{\otimes k}\|_F^2 &= \sum_{i_1=1}^{d}\cdots\sum_{i_k=1}^{d}[(T - \lambda v^{\otimes k})_{i_1,\ldots,i_k}]^2 = \sum_{i_1=1}^{d}\cdots\sum_{i_k=1}^{d}[T_{i_1,\ldots,i_k} - \lambda(v^{\otimes k})_{i_1,\ldots,i_k}]^2 \\ &= \sum_{i_1=1}^{d}\cdots\sum_{i_k=1}^{d}(T_{i_1,\ldots,i_k}^2 - 2\lambda T_{i_1,\ldots,i_k}v_{i_1}v_{i_2}\cdots v_{i_k} + \lambda^2 v_{i_1}^2 v_{i_2}^2\cdots v_{i_k}^2) \\ &= \|T\|_F^2 - 2\lambda\sum_{i=1}^{d} v_i(T\times_2 v\times_3 v\times_4\cdots\times_k v)_i + \lambda^2\|v^{\otimes k}\|_F. \end{aligned}$$

Since $\|v\| = 1$ and by Lemma 2.12, $\|v^{\otimes k}\|_F = 1$, hence

$$\|T - \lambda v^{\otimes k}\|_F^2 = \|T\|_F^2 - 2\lambda\langle v, \lambda v\rangle + \lambda^2 = \|T\|_F^2 - 2\lambda^2\|v\|_2^2 + \lambda^2 = \|T\|_F^2 - \lambda^2.$$

Since $\|T - \lambda v^{\otimes k}\|_F \geq 0$, $\|T\|_F^2 - \lambda^2 \geq 0$ so $\|T\|_F^2 \geq \lambda^2$ and taking square roots, $\|T\|_F \geq |\lambda|$. ∎

Appendix B. Proof of Lemma 3.2.

Proof. First note that for 3-tensors, if λ is an eigenvalue then $(T \times_1 v) \times_1 v = \lambda v$ so $(T \times_1 (-v)) \times_1 (-v) = -\lambda(-v)$ so $-\lambda$ is also an eigenvalue. Therefore for 3-tensors, $\lambda_{max} = \lambda_{maxabs}$, and in fact this is true for any odd order tensor.

Next, by the symmetry of the matrix 3-tensor T

$$\sum_{i,j,k} T_{ijk}v_iv_jv_k = \sum_{i=1}^n T_{iii}v_i^3 + 3\sum_{i=1}^n \sum_{k\neq i} T_{iik}v_i^2v_k + \sum_{i=1}^n \sum_{j\neq i} \sum_{k\neq i,j} T_{ijk}v_iv_jv_k.$$

Let us fix $s = 1, \ldots, n$ and let $(v_s)_i = \text{sign}(T_{sss})\,\delta_{is}$. Then $\|v_s\| = 1$ so

$$\begin{aligned}\lambda_{max} &= \max_{\|v\|=1} \sum_{i,j,k} T_{ijk}v_iv_jv_k \geq \sum_{i,j,k} T_{ijk}(v_s)_i(v_s)_j(v_s)_k \\ &= (\text{sign}(T_{sss}))^3 \sum_{i,j,k} T_{ijk}\delta_{is}\delta_{js}\delta_{ks} = \text{sign}(T_{sss}) \sum_{i,j,k} T_{ijk}\delta_{is}\delta_{js}\delta_{ks} \\ &= \text{sign}(T_{sss})T_{sss} = |T_{sss}|.\end{aligned}$$

Thus, in this case we have $\lambda_{max} \geq |T_{sss}|$ for all $s = 1, \ldots, n$. Next, fix $s, t \in \{1, \ldots, n\}$ and let

$$(w_{s,t})_i = \text{sign}(T_{stt})\frac{\delta_{is} + \delta_{it}}{\sqrt{2}}.$$

Then $\|w_{s,t}\| = 1$ and so

$$\lambda_{max} \geq \sum_{i,j,k} T_{ijk}(w_{s,t})_i(w_{s,t})_j(w_{s,t})_k = \left(\frac{\text{sign}(T_{stt})}{\sqrt{2}}\right)^3 (T_{sss} + T_{ttt} + 3T_{sst} + 3T_{stt})$$

and therefore

$$2^{3/2}\lambda_{max} \geq \text{sign}(T_{stt})(T_{sss} + T_{ttt} + 3T_{sst} + 3T_{stt}). \tag{B.1}$$

Now let

$$(\tilde{w}_{s,t})_i = \text{sign}(T_{stt})\frac{\delta_{is} - \delta_{it}}{\sqrt{2}}$$

so we have

$$\lambda_{max} \geq \sum_{i,j,k} T_{ijk}(\tilde{w}_{s,t})_i(\tilde{w}_{s,t})_j(\tilde{w}_{s,t})_k = \left(\frac{\text{sign}(T_{stt})}{\sqrt{2}}\right)^3 (T_{sss} - T_{ttt} - 3T_{sst} + 3T_{stt})$$

and

$$2^{3/2}\lambda_{max} \geq \text{sign}(T_{stt})(T_{sss} - T_{ttt} - 3T_{sst} + 3T_{stt}). \tag{B.2}$$

Adding (B.1) and (B.2), we get

$$2(2^{3/2}\lambda_{max}) \geq \text{sign}(T_{stt})(2T_{sss} + 6T_{stt}),$$
$$2^{3/2}\lambda_{max} \geq \text{sign}(T_{stt})(T_{sss} + 3T_{stt}).$$

Recall that $\lambda_{max} \geq |T_{sss}| \geq -\text{sign}(T_{stt})T_{sss}$, so

$$2^{3/2}\lambda_{max} + \lambda_{max} \geq \text{sign}(T_{stt})(T_{sss} + 3T_{stt} - T_{sss}),$$
$$(2^{3/2} + 1)\lambda_{max} \geq 3\,\text{sign}(T_{stt})T_{stt}.$$

Therefore

$$\lambda_{max} \geq \frac{3}{2^{3/2} + 1}|T_{stt}|.$$

Last, fix $s, t, u \in \{1, \ldots, n\}$ and let

$$(w_{s,t,u})_i = \text{sign}(T_{stu})\frac{\delta_{is} + \delta_{it} + \delta_{iu}}{\sqrt{3}}.$$

Then $\|w_{s,t,u}\| = 1$ and so

$$\lambda_{max} \geq \sum_{i,j,k} T_{ijk}(w_{s,t,u})_i(w_{s,t,u})_j(w_{s,t,u})_k = \left(\frac{\text{sign}(T_{stu})}{\sqrt{3}}\right)^3 (T_{sss} + T_{ttt} + T_{uuu} + 3T_{sst}$$
$$+3T_{stt} + 3T_{ssu} + 3T_{ttu} + 3T_{suu} + 3T_{tuu} + 6T_{stu}).$$

Note that since λ_{max} is greater than or equal to $-\text{sign}(T_{stu})T_{sss}$, $-\text{sign}(T_{stu})T_{ttt}$, and $-\text{sign}(T_{stu})T_{uuu}$ we can factor out $\text{sign}(T_{stu})$ so that

$$\left(3^{3/2} + 3\right)\lambda_{max} \geq 3\,\text{sign}(T_{stu})(T_{sst} + T_{stt} + T_{ssu} + T_{ttu} + T_{suu} + T_{tuu} + 2T_{stu}).$$

Now note that $(2^{3/2} + 1)\lambda_{max} \geq 3|T_{stt}| \geq -3\,\text{sign}(T_{stu})T_{stt}$, which implies

$$(2^{3/2} + 1)\lambda_{max} \geq -3\,\text{sign}(T_{stu})T_{ssu},$$
$$(2^{3/2} + 1)\lambda_{max} \geq -3\,\text{sign}(T_{stu})T_{ttu},$$
$$(2^{3/2} + 1)\lambda_{max} \geq -3\,\text{sign}(T_{stu})T_{suu},$$
$$(2^{3/2} + 1)\lambda_{max} \geq -3\,\text{sign}(T_{stu})T_{tuu},$$

and together these imply

$$\left(3^{3/2} + 3 + 6(2^{3/2} + 1)\right)\lambda_{max} \geq 6\,\text{sign}(T_{stu})T_{stu},$$
$$\left(3 + 4\sqrt{2} + \sqrt{3}\right)\lambda_{max} \geq 2|T_{stu}|,$$
$$\lambda_{max} \geq \frac{2}{3 + 4\sqrt{2} + \sqrt{3}}|T_{stu}|.$$

Comparing the lower bounds found in the above three cases, we see that the conclusion holds if we set

$$c = \frac{2}{3 + 4\sqrt{2} + \sqrt{3}}.$$

This completes the proof for 3-tensors. Next we follow a similar strategy for 4-tensors.

By the symmetry of the 4-tensor T, we have

$$\sum_{i,j,k,\ell} T_{ijk\ell} v_i v_j v_k v_\ell = \sum_{i=1}^{n} T_{iiii} v_i^4 + 6 \sum_{i=1}^{n} \sum_{j \neq i} T_{iijj} v_i^2 v_j^2 + 4 \sum_{i=1}^{n} \sum_{j \neq i} T_{iiij} v_i^3 v_j$$
$$+12 \sum_{i=1}^{n} \sum_{j \neq i} \sum_{k \neq i,j} T_{iijk} v_i^2 v_j v_k + 24 \sum_{i=1}^{n} \sum_{j \neq i} \sum_{k \neq i,j} \sum_{\ell \neq i,j,k} T_{ijk\ell} v_i v_j v_k v_\ell.$$

We wish to show that for some constant $c \in (0, 1]$,

$$\lambda_{maxabs} \geq c |T_{ijk\ell}| \quad \text{for all} \quad i, j, k, \ell \in \{1, 2, \ldots, n\}.$$

We are going to carry out the proof in five steps by looking at the following cases:

1. $i = j = k = \ell$.
2. $i = j \neq k = \ell$.
3. $i = j = k \neq \ell$.
4. $i = j$ distinct from k, ℓ and $k \neq \ell$.
5. i, j, k, ℓ all distinct.

1. Let us fix $s = 1, \ldots, n$ and let $(v_s)_i = \delta_{is}$. Then $\|v_s\| = 1$ so

$$\lambda_{max} = \max_{\|v\|=1} \sum_{i,j,k,\ell} T_{ijk\ell} v_i v_j v_k v_\ell \geq \sum_{i,j,k,\ell} T_{ijk\ell} (v_s)_i (v_s)_j (v_s)_k (v_s)_\ell = T_{ssss},$$
$$\lambda_{min} = \min_{\|v\|=1} \sum_{i,j,k,\ell} T_{ijk\ell} v_i v_j v_k v_\ell \leq \sum_{i,j,k,\ell} T_{ijk\ell} (v_s)_i (v_s)_j (v_s)_k (v_s)_\ell = T_{ssss}.$$

Thus $-\lambda_{min} \geq -T_{ssss}$. Therefore, for all $s = 1, \ldots, n$, we have

$$\lambda_{maxabs} = \max\{|\lambda_{max}|, |\lambda_{min}|\} \geq \lambda_{max} \geq T_{ssss},$$
$$\lambda_{maxabs} = \max\{|\lambda_{max}|, |\lambda_{min}|\} \geq -\lambda_{min} \geq -T_{ssss}.$$

Thus

$$\lambda_{maxabs} \geq |T_{ssss}| \quad \text{for each} \quad s = 1, \ldots, n.$$

2. Next, fix $s, t \in \{1, \ldots, n\}$ and let

$$(w_{s,t})_i = \frac{\delta_{is} + \delta_{it}}{\sqrt{2}}.$$

Then $\|w_{s,t}\| = 1$ and so

$$\lambda_{max} \geq \sum_{i,j,k,\ell} T_{ijk\ell} (w_{s,t})_i (w_{s,t})_j (w_{s,t})_k (w_{s,t})_\ell$$
$$= \left(\frac{1}{\sqrt{2}}\right)^4 \left(T_{ssss} + T_{tttt} + \binom{4}{2} T_{sstt} + \binom{4}{1} T_{ssst} + \binom{4}{1} T_{stttt} \right).$$

Hence

$$4\lambda_{max} \geq T_{ssss} + T_{tttt} + 6T_{sstt} + 4T_{ssst} + 4T_{sttt}. \tag{B.3}$$

Now let

$$(\tilde{w}_{s,t})_i = \frac{\delta_{is} - \delta_{it}}{\sqrt{2}}.$$

Then

$$\begin{aligned}\lambda_{max} &\geq \sum_{i,j,k,\ell} T_{ijk\ell}(\tilde{w}_{s,t})_i(\tilde{w}_{s,t})_j(\tilde{w}_{s,t})_k(\tilde{w}_{s,t})_\ell \\ &= \left(\frac{1}{\sqrt{2}}\right)^4 \left(T_{ssss} + T_{tttt} + \binom{4}{2} T_{sstt} - \binom{4}{1} T_{ssst} - \binom{4}{1} T_{sttt}\right).\end{aligned}$$

Thus

$$4\lambda_{max} \geq T_{ssss} + T_{tttt} + 6T_{sstt} - 4T_{ssst} - 4T_{sttt}. \tag{B.4}$$

Adding inequalities (B.3) and (B.4), and dividing by 2, we obtain

$$4\lambda_{max} \geq T_{ssss} + T_{tttt} + 6T_{sstt}. \tag{B.5}$$

Since $-\lambda_{min}$ is no smaller than both $-T_{ssss}$ and $-T_{tttt}$, subtracting $2\lambda_{min}$ from (B.5) gives us

$$6\lambda_{maxabs} \geq 4\lambda_{max} - 2\lambda_{min} \geq 6T_{sstt}. \tag{B.6}$$

Therefore

$$\lambda_{maxabs} \geq T_{sstt}. \tag{B.7}$$

Arguing similarly, we can show that

$$4\lambda_{min} \leq T_{ssss} + T_{tttt} + 6T_{sstt}. \tag{B.8}$$

Since $-\lambda_{max}$ is no larger than both $-T_{ssss}$ and $-T_{tttt}$, subtracting $2\lambda_{max}$ from (B.8) yields

$$4\lambda_{min} - 2\lambda_{max} \leq 6T_{sstt}. \tag{B.9}$$

Thus

$$6\lambda_{maxabs} \geq 2\lambda_{max} - 4\lambda_{min} \geq -6T_{sstt}.$$

Hence $\lambda_{maxabs} \geq -T_{sstt}$. Using this and (B.7), we obtain

$$\lambda_{maxabs} \geq |T_{sstt}|.$$

3. Next, fix distinct $s, t \in \{1, \dots, n\}$ and let

$$(w_{s,t})_i = \frac{\delta_{is} - 2\delta_{it}}{\sqrt{3}}.$$

Then $\|w_{s,t}\| = 1$ and so

$$\begin{aligned}\lambda_{max} &\geq \sum_{i,j,k,\ell} T_{ijk\ell}(w_{s,t})_i(w_{s,t})_j(w_{s,t})_k(w_{s,t})_\ell \\ &= \left(\frac{1}{\sqrt{3}}\right)^4 (T_{ssss} + 16T_{tttt} + 24T_{sstt} - 8T_{ssst} - 32T_{sttt}).\end{aligned}$$

Hence

$$9\lambda_{max} \geq T_{ssss} + 16T_{tttt} + 24T_{sstt} - 8T_{ssst} - 32T_{sttt}. \tag{B.10}$$

Adding inequalities (B.3) multiplied by 8 and (B.10), we get

$$41\lambda_{max} \geq 9T_{ssss} + 24T_{tttt} + 72T_{sstt} + 24T_{ssst}. \tag{B.11}$$

Since $-\lambda_{min}$ is no smaller than both $-T_{ssss}$ and $-T_{tttt}$,

$$-33\lambda_{min} \geq -9T_{ssss} - 24T_{tttt}. \tag{B.12}$$

Moreover, by (B.9),

$$\lambda_{max} - 2\lambda_{min} \geq -3T_{sstt}. \tag{B.13}$$

Thus, by (B.13) and using (B.11) and (B.12), we have

$$\begin{aligned}65\lambda_{max} - 81\lambda_{min} &= 41\lambda_{max} - 33\lambda_{min} + 24(\lambda_{max} - 2\lambda_{min}) \\ &\geq 9T_{ssss} + 24T_{tttt} + 72T_{sstt} + 24T_{ssst} \\ &\quad - 9T_{ssss} - 24T_{tttt} - 72T_{sstt} \\ &= 24T_{ssst}.\end{aligned}$$

Thus

$$146\lambda_{maxabs} \geq 65\lambda_{max} - 81\lambda_{min} \geq 24T_{ssst}. \tag{B.14}$$

Similarly, we can show that

$$65\lambda_{min} - 81\lambda_{max} \leq 24T_{ssst}.$$

Thus

$$146\lambda_{maxabs} \geq 81\lambda_{max} - 65\lambda_{min} \geq -24T_{ssst}. \tag{B.15}$$

Therefore, by (B.14) and (B.15), we obtain

$$\lambda_{maxabs} \geq \frac{12}{73}|T_{ssst}|.$$

4. Next, fix distinct $s,t,u \in \{1,\dots,n\}$ and let

$$(w_{s,t,u})_i = \frac{\delta_{is}+\delta_{it}+\delta_{iu}}{\sqrt{3}}.$$

Then $\|w_{s,t,u}\| = 1$ and so

$$\begin{aligned}\lambda_{max} &\geq \sum_{i,j,k,\ell} T_{ijk\ell}(w_{s,t,u})_i(w_{s,t,u})_j(w_{s,t,u})_k(w_{s,t,u})_\ell \\ &= \left(\frac{1}{\sqrt{3}}\right)^4 (T_{ssss}+T_{tttt}+T_{uuuu}+12T_{sstu}+12T_{sttu}+12T_{stuu}+6T_{sstt} \\ &\quad +6T_{ssuu}+6T_{ttuu}+4T_{ssst}+4T_{sssu}+4T_{sttt}+4T_{suuu}+4T_{tuuu}+4T_{uttt}).\end{aligned}$$

Thus,

$$\begin{aligned}9\lambda_{max} \geq\ & T_{ssss}+T_{tttt}+T_{uuuu}+12T_{sstu}+12T_{sttu}+12T_{stuu} \\ & +6T_{sstt}+6T_{ssuu}+6T_{ttuu}+4T_{ssst}+4T_{sssu}+4T_{sttt} \\ & +4T_{suuu}+4T_{tuuu}+4T_{uttt}.\end{aligned} \tag{B.16}$$

Now let

$$(\tilde{w}_{s,t,u})_i = \frac{\delta_{is}-\delta_{it}-\delta_{iu}}{\sqrt{3}}.$$

Then

$$\begin{aligned}\lambda_{max} &\geq \sum_{i,j,k,\ell} T_{ijk\ell}(\tilde{w}_{s,t,u})_i(\tilde{w}_{s,t,u})_j(\tilde{w}_{s,t,u})_k(\tilde{w}_{s,t,u})_\ell \\ &= \left(\frac{1}{\sqrt{3}}\right)^4 (T_{ssss}+T_{tttt}+T_{uuuu}+12T_{sstu}-12T_{sttu}-12T_{stuu}+6T_{sstt} \\ &\quad +6T_{ssuu}+6T_{ttuu}-4T_{ssst}-4T_{sssu}-4T_{sttt}-4T_{suuu}+4T_{tuuu}+4T_{uttt}).\end{aligned}$$

Thus,

$$\begin{aligned}9\lambda_{max} \geq\ & T_{ssss}+T_{tttt}+T_{uuuu}+12T_{sstu}-12T_{sttu}-12T_{stuu} \\ & +6T_{sstt}+6T_{ssuu}+6T_{ttuu}-4T_{ssst}-4T_{sssu}-4T_{sttt} \\ & -4T_{suuu}+4T_{tuuu}+4T_{uttt}.\end{aligned} \tag{B.17}$$

Adding inequalities (B.16) and (B.17), we get

$$\begin{aligned}18\lambda_{max} \geq\ & 2(T_{ssss}+T_{tttt}+T_{uuuu})+24T_{sstu} \\ & +12(T_{sstt}+T_{ssuu}+T_{ttuu})+8(T_{tuuu}+T_{uttt}).\end{aligned} \tag{B.18}$$

Recall from (B.15) that

$$81\lambda_{max}-65\lambda_{min} \geq -24T_{ssst}. \tag{B.19}$$

Applying twice this estimate to the indices t and u after dividing both sides by 3, we obtain

$$54\lambda_{max} - \frac{130}{3}\lambda_{min} \geq -8(T_{tuuu} + T_{uttt}). \tag{B.20}$$

Since $-\lambda_{min}$ is no smaller than $-T_{ssss}$, $-T_{tttt}$, and $-T_{uuuu}$, using (B.13) and (B.20), inequality (B.18) yields

$$\begin{aligned} 18\lambda_{max} - 6\lambda_{min} + 12\lambda_{max} - 24\lambda_{min} + 54\lambda_{max} - \frac{130}{3}\lambda_{min} \\ \geq 2(T_{ssss} + T_{tttt} + T_{uuuu}) + 24T_{sstu} + 12(T_{sstt} + T_{ssuu} + T_{ttuu}) \\ +8(T_{tuuu} + T_{uttt}) - 2(T_{ssss} + T_{tttt} + T_{uuuu}) \\ -12(T_{sstt} + T_{ssuu} + T_{ttuu}) - 8(T_{tuuu} + T_{uttt}). \end{aligned}$$

Hence

$$84\lambda_{max} - \frac{220}{3}\lambda_{min} \geq 24T_{sstu}.$$

Multiplying by 3, we obtain

$$472\lambda_{maxabs} \geq 252\lambda_{max} - 220\lambda_{min} \geq 72T_{sstu}.$$

Following the same argument, we can show that

$$252\lambda_{min} - 220\lambda_{max} \leq 72T_{sstu}.$$

Thus

$$472\lambda_{maxabs} \geq 220\lambda_{max} - 252\lambda_{min} \geq -72T_{sstu}. \tag{B.21}$$

Therefore

$$\lambda_{maxabs} \geq \frac{9}{59}|T_{sstu}|.$$

5. Next, fix distinct $s, t, u, v \in \{1, \ldots, n\}$ and let

$$(w_{s,t,u,v})_i = \frac{\delta_{is} + \delta_{it} + \delta_{iu} + \delta_{iv}}{2}.$$

Then $\|w_{s,t,u,v}\| = 1$ and so

$$\begin{aligned} \lambda_{max} &\geq \sum_{i,j,k,\ell} T_{ijk\ell}(w_{s,t,u,v})_i(w_{s,t,u,v})_j(w_{s,t,u,v})_k(w_{s,t,u,v})_\ell \\ &= \left(\frac{1}{2}\right)^4 (T_{ssss} + T_{tttt} + T_{uuuu} + T_{vvvv} \\ &\quad + 6(T_{sstt} + T_{ssuu} + T_{ssvv} + T_{ttuu} + T_{ttvv} + T_{uuvv}) \\ &\quad + 12(T_{sstu} + T_{sstv} + T_{ssuv} + T_{ttsu} + T_{ttsv} + T_{ttuv} \\ &\quad + T_{uust} + T_{uusv} + T_{uutv} + T_{vvst} + T_{vvsu} + T_{vvtu}) \\ &\quad + 4(T_{ssst} + T_{sssu} + T_{sssv} + T_{ttts} + T_{tttu} + T_{tttv} \\ &\quad + T_{uuus} + T_{uuut} + T_{uuuv} + T_{vvvs} + T_{vvvt} + T_{vvvu}) \\ &\quad + 24T_{stuv}. \end{aligned}$$

Thus,

$$\begin{aligned}16\lambda_{max} \geq\ & T_{ssss} + T_{tttt} + T_{uuuu} + T_{vvvv} \\ & + 6(T_{sstt} + T_{ssuu} + T_{ssvv} + T_{ttuu} + T_{ttvv} + T_{uuvv}) \\ & + 12(T_{sstu} + T_{sstv} + T_{ssuv} + T_{ttsu} + T_{ttsv} + T_{ttuv} \\ & + T_{uust} + T_{uusv} + T_{uutv} + T_{vvst} + T_{vvsu} + T_{vvtu}) \\ & + 4(T_{ssst} + T_{sssu} + T_{sssv} + T_{ttts} + T_{tttu} + T_{tttv} \\ & + T_{uuus} + T_{uuut} + T_{uuuv} + T_{vvvs} + T_{vvvt} + T_{vvvu}) \\ & + 24T_{stuv}.\end{aligned}$$

Since $-\lambda_{min}$ is no smaller than the quantities $-T_{ssss}$, $-T_{tttt}$, $-T_{uuuu}$, and $-T_{vvvv}$, by (B.13) applied to the pairs of indices $\{s,t\}$, $\{s,u\}$, $\{s,v\}$, $\{t,u\}$, $\{t,v\}$, and $\{u,v\}$, and, in addition, using (B.19) and (B.21), we have

$$16\lambda_{max} - 4\lambda_{min} + 12\lambda_{max} - 24\lambda_{min} + 440\lambda_{max} - 504\lambda_{min} + 162\lambda_{max} - 130\lambda_{min} \geq 24T_{stuv},$$

which reduces to

$$630\lambda_{max} - 662\lambda_{min} \geq 24T_{stuv}.$$

Hence

$$1292\lambda_{maxabs} \geq 630\lambda_{max} - 662\lambda_{min} \geq 24T_{stuv}.$$

Similarly, we can show that

$$630\lambda_{min} - 662\lambda_{max} \leq 24T_{stuv}.$$

Thus

$$1292\lambda_{maxabs} \geq 662\lambda_{max} - 630\lambda_{min} \geq -24T_{stuv}.$$

Therefore, simplifying, we obtain

$$\lambda_{maxabs} \geq \frac{6}{323}|T_{stuv}|.$$

Comparing the lower bounds found in the above three cases, we see that the conclusion holds if we set

$$c = \frac{6}{323}.$$

■

Appendix C. Proof of Theorem 4.2.

Proof. We first we wish to show that the first moment equation matches our mean. We begin by splitting the sum

$$\sum_{i=-2}^{N} w_i\sigma_i = \sum_{i=-2}^{0} w_i\sigma_i + \sum_{i=1}^{2d} w_i\sigma_i + \sum_{i=2d+1}^{2d+2J} w_i\sigma_i + \sum_{i=2d+2J+1}^{N} w_i\sigma_i.$$

Using the expressions defining the 4-moment σ-points σ_i and the corresponding weights w_i, we have

$$\begin{aligned}\sum_{i=-2}^{N} w_i\sigma_i &= (1-d\beta^{-2}-\hat{L}\delta^{-4})\mu + \frac{1}{2\alpha}(\mu+\alpha\hat{\mu}) - \frac{1}{2\alpha}(\mu-\alpha\hat{\mu})\\ &\quad + \sum_{i=1}^{d}\frac{1}{2\beta^2}\big(\mu+\beta\sqrt{\hat{C}}_i\big) + \sum_{j=d+1}^{2d}\frac{1}{2\beta^2}\big(\mu-\beta\sqrt{\hat{C}}_{i-d}\big)\\ &\quad + \sum_{i=2d+1}^{2d+J}\frac{1}{2\gamma^3}\big(\mu+\gamma\tilde{v}_{i-2d}\big) + \sum_{j=2d+J+1}^{2d+2J}\frac{-1}{2\gamma^3}\big(\mu-\gamma\tilde{v}_{i-2d-J}\big)\\ &\quad + \sum_{i=2d+2J+1}^{2d+2J+L}\frac{1}{2\delta^4}s_{i-2d-2J}\big(\mu+\delta\tilde{u}_{i-2d-2J}\big)\\ &\quad + \sum_{j=2d+2J+L+1}^{N}\frac{1}{2\delta^4}s_{i-2d-2J-L}\big(\mu-\delta\tilde{u}_{i-2d-2J-L}\big),\end{aligned}$$

and regrouping like terms, we obtain

$$\begin{aligned}\sum_{i=-2}^{N} w_i\sigma_i &= (1-d\beta^{-2}-\hat{L}\delta^{-4})\mu+\hat{\mu}+\sum_{i=1}^{d}\frac{1}{2\beta^2}\big(2\mu+\beta\sqrt{\hat{C}}_i-\beta\sqrt{\hat{C}}_i\big)\\ &\quad + \sum_{i=1}^{J}\left(\frac{1}{2\gamma^3}\big(\mu+\gamma\tilde{v}_i\big)-\frac{1}{2\gamma^3}\big(\mu-\gamma\tilde{v}_i\big)\right)+\sum_{i=1}^{L}\frac{1}{2\delta^4}s_i\big(2\mu+\delta\tilde{u}_i-\delta\tilde{u}_i\big)\\ &= (1-d\beta^{-2}-\hat{L}\delta^{-4})\mu+\hat{\mu}+\sum_{i=1}^{d}\frac{\mu}{\beta^2}+\sum_{i=1}^{J}\frac{\tilde{v}_i}{\gamma^2}+\sum_{i=1}^{L}\frac{s_i\mu}{\delta^4}\\ &= (1-d\beta^{-2}-\hat{L}\delta^{-4})\mu+\hat{\mu}+d\beta^{-2}\mu+\gamma^{-2}\sum_{i=1}^{J}\tilde{v}_i+\delta^{-4}\mu\sum_{i=1}^{L}s_i\\ &= \mu+\hat{\mu}+\gamma^{-2}\tilde{\mu}\\ &= \mu\end{aligned}$$

using the definition $\hat{\mu}=-\gamma^{-2}\tilde{\mu}$ for the last equality.

To look at the other moment equations, let's first observe that for $n=2,3,4$,

$$\begin{aligned}\sum_{i=-2}^{N} w_i(\sigma_i-\mu)^{\otimes n} &= \sum_{i=-1}^{0} w_i(\sigma_i-\mu)^{\otimes n}+\sum_{i=1}^{2d} w_i(\sigma_i-\mu)^{\otimes n}\\ &\quad + \sum_{i=2d+1}^{2d+2J} w_i(\sigma_i-\mu)^{\otimes n}+\sum_{i=2d+2J+1}^{N} w_i(\sigma_i-\mu)^{\otimes n}.\end{aligned}$$

Notice that since the first σ-point σ_{-2} is μ, the term $w_{-2}(\sigma_{-2}-\mu)^{\otimes n}=0$. By the definition of σ-points and corresponding weights,

$$\begin{aligned}\sum_{i=-2}^{N} w_i(\sigma_i-\mu)^{\otimes n} &= \frac{\alpha^{n-1}}{2}\left(\hat{\mu}^{\otimes n}-(-\hat{\mu})^{\otimes n}\right)+\frac{\beta^{n-2}}{2}\left(\sum_{i=1}^{d}\left(\sqrt{\hat{C}_i}\right)^{\otimes n}+\sum_{j=d+1}^{2d}\left(-\sqrt{\hat{C}}_{i-d}\right)^{\otimes n}\right)\\ &+\frac{\gamma^{n-3}}{2}\left(\sum_{i=2d+1}^{2d+J}\left(\tilde{v}_{i-2d}\right)^{\otimes n}-\sum_{j=2d+J+1}^{2d+2J}\left(-\tilde{v}_{i-2d-J}\right)^{\otimes n}\right)\\ &+\frac{\delta^{n-4}}{2}\sum_{i=2d+2J+1}^{2d+2J+L} s_{i-2d-2J}\left(\tilde{u}_{i-2d-2J}\right)^{\otimes n}\\ &+\frac{\delta^{n-4}}{2}\sum_{j=2d+2J+L+1}^{N} s_{i-2d-2J-L}\left(-\tilde{u}_{i-2d-2J-L}\right)^{\otimes n},\end{aligned}$$

where we used the property $(av)^{\otimes n}=a^n v^{\otimes n}$, where a is any real number and v is a vector. When n is even, we have

$$\sum_{i=-2}^{N} w_i(\sigma_i-\mu)^{\otimes n}=\beta^{n-2}\sum_{i=1}^{d}\left(\sqrt{\hat{C}_i}\right)^{\otimes n}+\delta^{n-4}\sum_{i=1}^{L} s_i\left(\tilde{u}_i\right)^{\otimes n}, \tag{C.1}$$

and when n is odd, we obtain

$$\sum_{i=-2}^{N} w_i(\sigma_i-\mu)^{\otimes n}=\alpha^{n-1}\hat{\mu}^{\otimes n}+\gamma^{n-3}\sum_{i=1}^{J}\left(\tilde{v}_i\right)^{\otimes n}. \tag{C.2}$$

Now we wish to show that the second moment equation matches our covariance. By (C.1), setting $n=2$ we have

$$\sum_{i=-2}^{N} w_i(\sigma_i-\mu)^{\otimes 2}=\sum_{i=1}^{d}\sqrt{\hat{C}_i}^{\otimes 2}+\delta^{-2}\sum_{i=1}^{L} s_i\tilde{u}_i^{\otimes 2}=\hat{C}+\delta^{-2}\tilde{C},$$

and applying the definition of $\hat{C}=C-\delta^{-2}\tilde{C}$ we have

$$\sum_{i=-2}^{N} w_i(\sigma_i-\mu)^{\otimes 2}=C,$$

as desired. Next, observe that by (C.2) and the definition of S,

$$\sum_{j=0}^{N} w_i(\sigma_i-\mu)^{\otimes 3}=\alpha^2\hat{\mu}^{\otimes 3}+\sum_{i=1}^{J}\tilde{v}_i^{\otimes 3},$$

and since we assume that $||\sum_{i=1}^{J}\tilde{v}_i^{\otimes 3}-S||_F\leq\frac{\tau}{2}$, by the triangle inequality we have

$$\left|\left|\sum_{j=0}^{N} w_i(\sigma_i - \mu)^{\otimes 3} - S\right|\right|_F \leq \frac{\tau}{2} + \alpha^2 ||\hat{\mu}^{\otimes 3}||_F$$

as desired. Last, we wish to show that the fourth moment equation matches our kurtosis. By (C.1) and the definition of K, we have

$$\begin{aligned}\sum_{i=-2}^{N} w_i(\sigma_i - \mu)^{\otimes 4} &= \beta^2 \sum_{i=1}^{d} \sqrt{\hat{C}_i}^{\otimes 4} + \sum_{i=1}^{L} s_i \tilde{u}_i^{\otimes 4} \\ &= \beta^2 \bar{C} + \sum_{i=1}^{L} s_i \tilde{u}_i^{\otimes 4},\end{aligned}$$

and since we assume that $||\sum_{i=1}^{L} s_i \tilde{u}_i^{\otimes 4} - K||_F \leq \frac{\tau}{2}$, by the triangle inequality we have

$$\left|\left|\sum_{j=0}^{N} w_i(\sigma_i - \mu)^{\otimes 4} - K\right|\right|_F \leq \frac{\tau}{2} + \beta^2 ||\bar{C}||_F,$$

which completes the proof. ∎

Appendix D. Proofs of Theorems 5.1 and 5.2.

We first prove Theorem 5.1.

Proof. Suppose $f \in C^5(\mathbb{R}^d, \mathbb{R})$. Now we wish to find the error bound where $\mathbb{E}[f(x)]$ with $x \sim p$ is the truth and $\sum_{i=1}^{m} w_i f(\sigma_i)$ is our estimate where m is the number of σ-points (nodes) in the quadrature. By Taylor's theorem with remainder we can expand f centered at μ as

$$\begin{aligned}f(x) = f(\mu) + \nabla f(\mu)(x-\mu) + \frac{1}{2}\sum_{j,k=1}^{d} Hf(\mu)_{jk}(x-\mu)_j(x-\mu)_k \\ + \frac{1}{6}\sum_{j,k,l=1}^{d} D^3 f(\mu)_{jkl}(x-\mu)_j(x-\mu)_k(x-\mu)_l \\ + \frac{1}{24}\sum_{j,k,l,r=1}^{d} D^4 f(\mu)_{jklr}(x-\mu)_j(x-\mu)_k(x-\mu)_l(x-\mu)_r \\ + \frac{1}{120}\sum_{j,k,l,r,s=1}^{d} D^5 f(\mu^*)_{jklrs}(x-\mu)_j(x-\mu)_k(x-\mu)_l(x-\mu)_r(x-\mu)_s,\end{aligned}$$

where $\mu^* \in B_{\|x-\mu\|}(\mu)$ (i.e., $\|\mu^* - \mu\| < \|x - \mu\|$), and $D^k f(x)_{j_1 \cdots j_k} \equiv \frac{\partial^k f}{\partial x_{j_k} \cdots \partial x_{j_1}}(x)$. Thus

$$
\begin{aligned}
\mathbb{E}[f(x)] &= \int_{\mathbb{R}^d} f(x)p(x)\,dx \\
&= f(\mu)\int_{\mathbb{R}^d} p(x)\,dx + \nabla f(\mu)\int_{\mathbb{R}^d}(x-\mu)p(x)\,dx + \frac{1}{2}\sum_{j,k=1}^{d} Hf(\mu)_{jk}\int_{\mathbb{R}^d}(x-\mu)_j(x-\mu)_k p(x)\,dx \\
&\quad + \frac{1}{6}\sum_{j,k,l=1}^{d} D^3 f(\mu)_{jkl}\int_{\mathbb{R}^d}(x-\mu)_j(x-\mu)_k(x-\mu)_l p(x)\,dx \\
&\quad + \frac{1}{24}\sum_{j,k,l,r=1}^{d} D^4 f(\mu)_{jklr}\int_{\mathbb{R}^d}(x-\mu)_j(x-\mu)_k(x-\mu)_l(x-\mu)_r p(x)\,dx \\
&\quad + \frac{1}{120}\sum_{j,k,l,r,s=1}^{d}\int_{\mathbb{R}^d} D^5 f(\mu_x^*)_{jklrs}(x-\mu)_j(x-\mu)_k(x-\mu)_l(x-\mu)_r(x-\mu)_s p(x)\,dx \\
&= f(\mu) + \frac{1}{2}\sum_{j,k=1}^{d} Hf(\mu)_{jk}C_{jk} + \frac{1}{6}\sum_{j,k,l=1}^{d} D^3 f(\mu)_{jkl}S_{jkl} + \frac{1}{24}\sum_{j,k,l,r=1}^{d} D^4 f(\mu)_{jklr}K_{jklr} \\
&\quad + \frac{1}{120}\sum_{j,k,l,r,s=1}^{d}\int_{\mathbb{R}^d} D^5 f(\mu_x^*)_{jklrs}(x-\mu)_j(x-\mu)_k(x-\mu)_l(x-\mu)_r(x-\mu)_s p(x)\,dx,
\end{aligned}
$$

where the subscript on μ_x^* denotes the implicit dependence on x of the remainder in Taylor's theorem. Since the quadrature exactly matches the first four moments we have

$$
\begin{aligned}
1 &= \sum_{i=1}^{m} w_i, \\
\mu &= \sum_{i=1}^{m} w_i\sigma_i, \\
C_{jk} &= \sum_{i=1}^{m} w_i(\sigma_i-\mu)_j(\sigma_i-\mu)_k, \\
S_{jkl} &= \sum_{i=1}^{m} w_i(\sigma_i-\mu)_j(\sigma_i-\mu)_k(\sigma_i-\mu)_l, \\
K_{jklr} &= \sum_{i=1}^{m} w_i(\sigma_i-\mu)_j(\sigma_i-\mu)_k(\sigma_i-\mu)_l(\sigma_i-\mu)_r.
\end{aligned}
$$

So applying Taylor's theorem inside the quadrature formula yields

$$
\begin{aligned}
\sum_{i=1}^{m} w_i f(\sigma_i) &= \sum_{i=1}^{m} w_i\Bigg(f(\mu) + \nabla f(\mu)(\sigma_i-\mu) + \frac{1}{2}\sum_{j,k=1}^{d} Hf(\mu)_{jk}(\sigma_i-\mu)_j(\sigma_i-\mu)_k \\
&\quad + \frac{1}{6}\sum_{j,k,l=1}^{d} D^3 f(\mu)_{jkl}(\sigma_i-\mu)_j(\sigma_i-\mu)_k(\sigma_i-\mu)_l \\
&\quad + \frac{1}{24}\sum_{j,k,l,r=1}^{d} D^4 f(\mu)_{jklr}(\sigma_i-\mu)_j(\sigma_i-\mu)_k(\sigma_i-\mu)_l(\sigma_i-\mu)_r
\end{aligned}
$$

$$+\frac{1}{120}\sum_{j,k,l,r,s=1}^{d} D^5 f(\mu^*)_{jklrs}(x-\mu)_j(x-\mu)_k(x-\mu)_l(x-\mu)_r(x-\mu)_s\Bigg)$$

$$= f(\mu)+\frac{1}{2}\sum_{j,k=1}^{d} Hf(\mu)_{jk}C_{jk}+\frac{1}{6}\sum_{j,k,l=1}^{d} D^3 f(\mu)_{jkl}S_{jkl}+\frac{1}{24}\sum_{j,k,l,r=1}^{d} D^4 f(\mu)_{jklr}K_{jklr}$$

$$+\frac{1}{120}\sum_{j,k,l,r,s=1}^{d}\left(\sum_{i=1}^{m} D^5 f(\mu^*_{\sigma_i})_{jklrs}w_i(\sigma_i-\mu)_j(\sigma_i-\mu)_k(\sigma_i-\mu)_l(\sigma_i-\mu)_r(\sigma_i-\mu)_s\right).$$

Notice that the first four terms of the true expectation and the quadrature formula agree. Since the first four terms cancel, the error becomes

$$\begin{aligned}
&\left|\mathbb{E}[f(x)]-\sum_{i=1}^{m} w_i f(\sigma_i)\right| \\
&\quad=\frac{1}{120}\left|\sum_{j,k,l,r,s=1}^{d}\int_{\mathbb{R}^d} D^5 f(\mu^*_x)_{jklrs}(x-\mu)^{\otimes 5}_{jklrs}\,dp-\sum_{i=1}^{m} D^5 f(\mu^*_{\sigma_i})_{jklrs}w_i(\sigma_i-\mu)^{\otimes 5}_{jklrs}\right| \\
&\quad\le\frac{1}{120}\sum_{j,k,l,r,s=1}^{d}\int_{\mathbb{R}^d}\left|D^5 f(\mu^*_x)_{jklrs}(x-\mu)^{\otimes 5}_{jklrs}\right|\,dp+\sum_{i=1}^{m}\left|D^5 f(\mu^*_{\sigma_i})_{jklrs}w_i(\sigma_i-\mu)^{\otimes 5}_{jklrs}\right| \\
&\quad\le\frac{||D^5 f||_\infty}{120}\sum_{j,k,l,r,s=1}^{d}\int_{\mathbb{R}^d}\left|(x-\mu)^{\otimes 5}_{jklrs}\right|\,dp+\sum_{i=1}^{m}\left|w_i(\sigma_i-\mu)^{\otimes 5}_{jklrs}\right| \\
&\quad\le||D^5 f||_\infty\frac{d^5}{120}\left(||M_{5,abs}||_{\max}+||\tilde{M}_{5,abs}||_{\max}\right).
\end{aligned}$$

We next turn to the proof of Theorem 5.2.

Proof. Recall that the total quadrature error is bounded above by the sum of the error due to the moments, E_{moments}, the error inside the ball, E_{inside}, and the error outside, E_{outside}. Since we assume that the quadrature exactly matches the first n moments, we have $E_{\text{moments}} = 0$. Next, combining the bound on f and the exponential decay bound on the density we have

$$\begin{aligned}
E_{\text{outside}} &= \int_{\mathbb{B}_r(\mu)^c}|f-q|\,dp\le\int_{\mathbb{B}_r(\mu)^c}(a+b||x-\mu||^t+b_2||x-\mu||^n)ce^{-\alpha||x-\mu||^\beta}\,dx \\
&=\omega_d\int_r^\infty(as^{d-1}+bs^{t+d-1}+b_2s^{n+d-1})ce^{-\alpha s^\beta}\,ds \\
&\le c\omega_d\int_r^\infty as^{d-\beta}s^{\beta-1}e^{-\alpha s^\beta}+bs^{t+d-\beta}s^{\beta-1}e^{-\alpha s^\beta}+b_2s^{n+d-\beta}s^{\beta-1}e^{-\alpha s^\beta}\,ds \\
&=(c_3r^{d-\beta}+c_4r^{t+d-\beta}+c_5r^{n+d-\beta})e^{-\alpha r^\beta} \\
&\quad+c\omega_d\int_r^\infty a_1s^{d-\beta-1}e^{-\alpha s^\beta}+b_3s^{t+d-\beta-1}e^{-\alpha s^\beta}+b_4s^{n+d-\beta-1}e^{-\alpha s^\beta}\,ds.
\end{aligned}$$

The above integration by parts can be repeated until $d-\beta, t+d-\beta, n+d-\beta$ are all less than $\beta-1$; then the integrands are bounded above by $s^{\beta-1}e^{-\alpha s^\beta}$ which is integrable exactly.

These integrations by parts pick up polynomial terms multiplied by $e^{-\alpha r^\beta}$, all of which are bounded by $r^{t+n+d-\beta}e^{-\alpha r^\beta}$. Since there are fewer than n such terms, we have

$$E_{\text{outside}} \leq c_2\, n\, r^{t+n+d-\beta}e^{-\alpha r^\beta}.$$

Finally, we turn to the error of polynomial approximation inside $\mathbb{B}_r(\mu)$. Defining $\tilde{f}(x) = f(rx+\mu)$ on the unit ball, we can apply Theorem 3.4 of [28], which says there exists a polynomial $\tilde{q}$ such that

$$||\tilde{f}-q||_{\infty,\mathbb{B}_1(0)} \leq \frac{c_1}{n^k}\left(\frac{||D^k\tilde{f}||_\infty}{n} + \sum_{|\gamma|=k}\sup_{|x-y|<1/n}|D^k_\gamma\tilde{f}(x) - D^k_\gamma\tilde{f}(y)|\right).$$

By the chain rule we have $|D^k\tilde{f}| = r^k|D^kf|$ so that

$$||f-q||_\infty \leq c_1\left(\frac{r}{n}\right)^k\left(\frac{||D^kf||_\infty}{n} + \sum_{|\gamma|=k}\sup_{|x-y|<1/n}|D^k_\gamma f(x) - D^k_\gamma f(y)|\right),$$

where $q(x) = \tilde{q}((x-\mu)/r)$. ∎

REFERENCES

[1] J. L. Anderson, *An adaptive covariance inflation error correction algorithm for ensemble filters*, Tellus A Dynamic Meteorology and Oceanography, 59 (2007), pp. 210–224.

[2] I. Arasaratnam and S. Haykin, *Cubature Kalman filters*, IEEE Trans. Automat. Control, 54 (2009), pp. 1254–1269.

[3] K. Atkinson and W. Han, *Theoretical Numerical Analysis*, Texts Appl. 39, Springer, New York, 2005.

[4] T. Berry and T. Sauer, *Adaptive ensemble Kalman filtering of non-linear systems*, Tellus A Dynamic Meteorology and Oceanography, 65 (2013), 20331.

[5] L. De Lathauwer, P. Comon, B. De Moor, and J. Vandewalle, *Higher-order power method*, in Proceedings of Nonlinear Theory and Its Applications, Vol. 1, 1995, pp. 91–96.

[6] L. De Lathauwer, B. De Moor, and J. Vandewalle, *On the best rank-*1 *and rank-*$(r_1, r_2, \dots r_n)$ *approximation of higher-order tensors*, SIAM J. Matrix Anal. Appl., 21 (2000), pp. 1324–1342.

[7] A. Falco and A. Nouy, *A proper generalized decomposition for the solution of elliptic problems in abstract form by using a functional Eckart–Young approach*, J. Math. Anal. Appl., 376 (2011), pp. 469–480.

[8] L. Grasedyck, D. Kressner, and C. Tobler, *A literature survey of low-rank tensor approximation techniques*, GAMM-Mitt., 36 (2013), pp. 53–78.

[9] J. Haastad, *Tensor rank is NP-complete*, in International Colloquium on Automata, Languages, and Programming, Springer, New York, 1989, pp. 451–460.

[10] W. Hackbusch, *Tensor Spaces and Numerical Tensor Calculus*, Springer Ser. Comput. Math. 42, Springer, New York, 2012.

[11] F. Hamilton, T. Berry, N. Peixoto, and T. Sauer, *Real-time tracking of neuronal network structure using data assimilation*, Phys. Rev. E, 88 (2013), 052715.

[12] F. Hamilton, T. Berry, and T. Sauer, *Ensemble Kalman filtering without a model*, Phys. Rev. X, 6 (2016), 011021.

[13] J. Harlim and A. J. Majda, *Catastrophic filter divergence in filtering nonlinear dissipative systems*, Commun. Math. Sci., 8 (2010), pp. 27–43.

[14] C. J. Hillar and L.-H. Lim, *Most tensor problems are NP-hard*, J. ACM, 60 (2013), pp. 1–39.

[15] B. Jia, M. Xin, and Y. Cheng, *Sparse-grid quadrature nonlinear filtering*, Automatica, 48 (2012), pp. 327–341.
[16] S. J. Julier, *Skewed approach to filtering*, in Signal and Data Processing of Small Targets 1998, SPIE 3373, International Society for Optics and Photonics, 1998, pp. 271–282.
[17] S. J. Julier, *The scaled unscented transformation*, in Proceedings of the 2002 American Control Conference, Vol. 6, IEEE, 2002, pp. 4555–4559.
[18] S. J. Julier and J. K. Uhlmann, *A General Method for Approximating Nonlinear Transformations of Probability Distributions*, Technical report, Robotics Research Group, Department of Engineering Science, University of Oxford, Oxford, 1996.
[19] S. J. Julier and J. K. Uhlmann, *New extension of the Kalman filter to nonlinear systems*, in Signal Processing, Sensor Fusion, and Target Recognition VI, SPIE. 3068, International Society for Optics and Photonics, 1997, pp. 182–193.
[20] S. J. Julier and J. K. Uhlmann, *Unscented filtering and nonlinear estimation*, Proc. IEEE, 92 (2004), pp. 401–422.
[21] S. J. Julier, J. K. Uhlmann, and H. F. Durrant-Whyte, *A new approach for filtering nonlinear systems*, in Proceedings of 1995 American Control Conference, Vol. 3, IEEE, 1995, pp. 1628–1632.
[22] E. Kofidis and P. A. Regalia, *On the best rank-1 approximation of higher-order supersymmetric tensors*, SIAM J. Matrix Anal. Appl., 23 (2002), pp. 863–884.
[23] T. G. Kolda, *A counterexample to the possibility of an extension of the Eckart–Young low-rank approximation theorem for the orthogonal rank tensor decomposition*, SIAM J. Matrix Anal. Appl., 24 (2003), pp. 762–767.
[24] T. G. Kolda and B. W. Bader, *Tensor decompositions and applications*, SIAM Rev., 51 (2009), pp. 455–500.
[25] O. Le Maître and O. M. Knio, *Spectral Methods for Uncertainty Quantification: With Applications to Computational Fluid Dynamics*, Springer, New York, 2010.
[26] E. N. Lorenz, *Deterministic nonperiodic flow*, J. Atmospheric Sci., 20 (1963), pp. 130–141.
[27] F. Nobile, R. Tempone, and C. G. Webster, *A sparse grid stochastic collocation method for partial differential equations with random input data*, SIAM J. Numer. Anal., 46 (2008), pp. 2309–2345.
[28] D. L. Ragozin, *Constructive polynomial approximation on spheres and projective spaces*, Trans. Amer. Math. Soc., 162 (1971), pp. 157–170.
[29] P. A. Regalia and E. Kofidis, *The higher-order power method revisited: Convergence proofs and effective initialization*, in Proceedings of the 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 5, IEEE, 2000, pp. 2709–2712.
[30] S. Särkkä, *Unscented Rauch–Tung–Striebel smoother*, IEEE Trans. Automat. Control, 53 (2008), pp. 845–849.
[31] A. Stegeman and P. Comon, *Subtracting a best rank-1 approximation may increase tensor rank*, Linear Algebra Appl., 433 (2010), pp. 1276–1300.
[32] R. Van Der Merwe, A. Doucet, N. De Freitas, and E. A. Wan, *The unscented particle filter*, in Advances in Neural Information Processing Systems, 2001, pp. 584–590.
[33] E. A. Wan and R. Van Der Merwe, *The unscented Kalman filter for nonlinear estimation*, in Proceedings of the IEEE 2000 Adaptive Systems for Signal Processing, Communications, and Control Symposium, 2000, pp. 153–158.
[34] E. A. Wan, R. Van Der Merwe, and A. T. Nelson, *Dual estimation and the unscented transformation*, in Advances in Neural Information Processing Systems, 2000, pp. 666–672.
[35] D. Xiu and G. E. Karniadakis, *The Wiener–Askey polynomial chaos for stochastic differential equations*, SIAM J. Sci. Comput., 24 (2002), pp. 619–644.