Statistics and Computing (2022) 32:12
https://doi.org/10.1007/s11222-021-10069-9

Product-form estimators: exploiting independence to scale up Monte Carlo

Juan Kuntz[1,2] · Francesca R. Crucinio[1] · Adam M. Johansen[1,2]

Received: 6 April 2021 / Accepted: 6 November 2021

Abstract
We introduce a class of Monte Carlo estimators that aim to overcome the rapid growth of variance with dimension often observed for standard estimators by exploiting the target's independence structure. We identify the most basic incarnations of these estimators with a class of generalized U-statistics and thus establish their unbiasedness, consistency, and asymptotic normality. Moreover, we show that they obtain the minimum possible variance amongst a broad class of estimators, and we investigate their computational cost and delineate the settings in which they are most efficient. We exemplify the merger of these estimators with other well known Monte Carlo estimators so as to better adapt the latter to the target's independence structure and improve their performance. We do this via three simple mergers: one with importance sampling, another with importance sampling squared, and a final one with pseudo-marginal Metropolis–Hastings. In all cases, we show that the resulting estimators are well founded and achieve lower variances than their standard counterparts. Lastly, we illustrate the various variance reductions through several examples.

Keywords Dimensionality reduction, Importance sampling · Pseudo-marginal methods · Limit theorems · Product-form distributions · Conditional independence · U-statistics · Variance reduction

1 Introduction

Monte Carlo methods are sometimes said to overcome the curse of dimensionality because, regardless of the target's dimension, their rates of convergence are square root in the number of samples drawn. In practice, however, one encounters several problems when computing high-dimensional integrals using Monte Carlo, prominent among which is the issue that the constants present in the convergence rates typically grow rapidly with the target's dimension. Hence, even if we are able to draw independent samples from a high-dimensional target, the number of samples necessary to obtain estimates of a satisfactory accuracy is often prohibitively large (Silverman 1986; Snyder et al. 2008; Bengtsson et al. 2008; Agapiou et al. 2017). However, many of these targets possess strong independence structures [e.g., see Gelman and Hill (2006), Gelman (2006), Koller and Friedman (2009), Hoffman et al. (2013), Blei et al. (2003), and the many references therein]. In this paper, we investigate whether the rapid growth of the constants can be mitigated by exploiting these structures.

Variants of the following toy example are sometimes given to illustrate the issue [e.g., p. 95 in Chopin and Papaspiliopoulos (2020)]. Let μ be a K-dimensional isotropic Gaussian distribution with unit means and variances, and consider the basic Monte Carlo estimator for the mean ($\mu(\varphi) = 1$) of the product ($\varphi(x) := x_1 x_2 \dots x_K$) of its components $(x_1, \dots, x_K)$:

$$\mu^N(\varphi) := \frac{1}{N}\sum_{n=1}^{N} \varphi(X_1^n, \dots, X_K^n) = \frac{1}{N}\sum_{n=1}^{N} X_1^n \cdots X_K^n, \tag{1}$$

✉ Juan Kuntz
juan.kuntz-nussio@warwick.ac.uk

Francesca R. Crucinio
francesca.crucinio@warwick.ac.uk

Adam M. Johansen
a.m.johansen@warwick.ac.uk

[1] Department of Statistics, University of Warwick, Coventry CV4 7AL, UK

[2] Alan Turing Institute, 96 Euston Road, London NW1 2DB, UK

Published online: 21 December 2021

where $(X_1^n, \ldots, X_K^n)_{n=1}^N$ denote i.i.d. samples drawn from μ. Because the estimator's asymptotic variance equals $2^K - 1$, the number of samples required to obtain a reasonable estimate of $\mu(\varphi)$ grows exponentially with the target's dimension. Hence, it is impractical to use $\mu^N(\varphi)$ if K is even modestly large. For instance, if $K = 20$, we would require $\approx 10^{10}$ samples to obtain an estimate with standard deviation of $0.01 = \mu(\varphi)/100$, reaching the limits of most present-day personal computers, and if $K = 30$, we would require $\approx 10^{13}$ samples, exceeding these limits.

There is, however, a trivial way of overcoming the issue for the above example that does not require any knowledge about μ beyond the fact that it is product-form. Because μ is the product $\mu_1 \times \cdots \times \mu_K$ of K univariate unit-mean-and-variance Gaussian distributions $\mu_1, \ldots, \mu_K$ and φ is the product $\varphi_1 \cdots \varphi_K$ of K univariate functions $\varphi_1(x_1) = x_1, \ldots, \varphi_K(x_K) = x_K$, we can express $\mu(\varphi)$ as the product $\mu_1(\varphi_1) \cdots \mu_K(\varphi_K)$ of the corresponding K univariate means $\mu_1(\varphi_1), \ldots, \mu_K(\varphi_K)$. As we will see in Sect. 2.1, estimating each of these univariate means separately and taking the resulting product, we obtain an estimator for $\mu(\varphi)$ whose asymptotic variance is K:

$$\mu_\times^N(\varphi) := \frac{1}{N^K} \sum_{n_1=1}^{N} \cdots \sum_{n_K=1}^{N} \varphi(X_1^{n_1}, \ldots, X_K^{n_K}) = \left(\frac{1}{N} \sum_{n_1=1}^{N} X_1^{n_1} \right) \cdots \left(\frac{1}{N} \sum_{n_K=1}^{N} X_K^{n_K} \right). \tag{2}$$

Consequently, the number of samples necessary for $\mu_\times^N(\varphi)$ to yield a reasonable estimate of $\mu(\varphi)$ only grows linearly with the dimension, allowing us to practically deal with Ks in the millions.

The central expression in (2) makes sense regardless of whether φ is the product of univariate test functions. It defines a type of (unbiased, consistent, and asymptotically normal) Monte Carlo estimators for general φ and product-form μ which we refer to as *product-form estimators*. Their salient feature is that they achieve lower variances than the standard estimator (1) given the same number of samples from the target. The reason behind the variance reduction is simple: if $(X_1^n)_{n=1}^N, \ldots, (X_K^n)_{n=1}^N$ are independent sequences of samples drawn respectively from $\mu_1, \ldots, \mu_K$, then every 'permutation' of these samples has law μ, that is,

$$(X_1^{n_1}, \ldots, X_K^{n_K}) \sim \mu \quad \forall n_1, \ldots, n_K \leq N. \tag{3}$$

Hence, $\mu_\times^N(\varphi)$ in (2) averages over N^K tuples with law μ while its conventional counterpart (1) only averages over N such tuples. This increase in tuple number leads to a decrease in estimator variance, and we say that the product-form estimator is more *statistically efficient* than the standard one. Moreover, obtaining these N^K tuples does not require drawing any further samples from μ and, in this sense, product-form estimators make the most out of every sample available (indeed, we will show in Theorem 2 that they are minimum variance unbiased estimators, or MVUEs, for product-form targets). However, in contrast to the tuples in (1), those in (2) are not independent (the same components are repeated across several tuples). For this reason, product-form estimators achieve the same $\mathcal{O}(N^{-1/2})$ rate of convergence that the standard ones do and the variance reduction materializes only in lower proportionality constants (i.e., $\lim_{N\to\infty} \mathrm{Var}\,(\mu_\times^N(\varphi))/\mathrm{Var}\,(\mu^N(\varphi)) = C$ for some constant $C \leq 1$).

The space complexity of product-form estimators scales linearly with dimension: to utilize all N^K permuted tuples in (2) we need only store KN numbers,

$$X_1^1, \ldots, X_1^N; \ldots; X_K^1, \ldots, X_K^N.$$

However, unless the test function possesses special structure, the estimators' time complexity scales exponentially with dimension: brute-force computation of the sum in (2) requires[1] $\mathcal{O}(N^K)$ operations. Consequently, the use of product-form estimators for general φ proves to be a balancing act in which one must weigh the cost of acquiring new samples from μ (be it a computational one if the samples are obtained from simulations, or a real-life one if they are obtained from experiments) against the extra overhead required to evaluate these estimators, and it is limited to Ks no greater than ten.

If, however, the test function φ possesses some 'product structure,' then $\mu_\times^N(\varphi)$ can often be evaluated in far fewer than $\mathcal{O}(N^K)$ operations. The most extreme examples of such φ are functions that factorize fully and sums thereof (which we refer to as 'sums of products' or 'SOPs'), for which the evaluation cost is easily lowered to just $\mathcal{O}(KN)$. For instance, in the case of the toy Gaussian example above, we can evaluate the product-form estimator in $\mathcal{O}(KN)$ operations by expressing it as the product of the component-wise sample averages and computing each average separately (i.e., using the final expression in (2)). This cheaper approach just amounts to a dimensionality reduction technique: we re-write a high-dimensional integral as a polynomial of low-dimension integrals, estimate each of low-dimension integral

[1] *On our $\mathcal{O}$ notation:* The exact dependence on dimension of the estimators' evaluation costs depends on that of the test function φ. Hence, when discussing a generic φ, we say that the estimator's evaluation cost is $\mathcal{O}(N^d)$ for some d to mean that it is $\mathcal{O}(f(K)N^d)$ for some unspecified factor $f(K)$ factor independent of that accounts for φ's evaluation cost. When discussing classes of φ for which this factor is clear, we specify it. For example, we say that evaluation cost of the rightmost term in (2) is $\mathcal{O}(KN)$ rather than $\mathcal{O}(N)$.

separately, and plug the estimates back into the polynomial to obtain an estimate of the original integral. More generally, if the test function can be expressed as a sum of partially factorized functions, it is often possible to lower the cost to $\mathcal{O}(N^d)$ where $d < K$ depends on the amount of factorization, and taking this approach also amounts to a type of dimensionality reduction (this time featuring nested integrals).

This paper has two goals. First, to provide a comprehensive theoretical characterization of product-form estimators. Second, to illustrate their use for non-product-form targets when combined with, or embedded within, other more sophisticated Monte Carlo methodology. It is in these settings, where product-form estimators are deployed to tackle the aspects of the problem exhibiting product structure or conditional independences, that we believe these estimators find their greatest use. To avoid unnecessary technical distractions, and in the interest of accessibility, we achieve the second goal using simple examples. While we anticipate that the most useful such combinations or embeddings will not be so simple, we believe that the underlying ideas and guiding principles will be the same.

Relation to the literature In their basic form, product-form estimators (2) are a subclass of generalized U-statistics [see Lee (1990) or Korolyuk and Borovskich (1994) for comprehensive surveys]: multisample U-statistics with 'kernels' φ that take as arguments a *single* sample per distribution for several distributions ($K > 1$). Even though product-form estimators are unnatural examples of U-statistics because the original unisample U-statistics (Hoeffding 1948a) fundamentally involve symmetric kernels that take as arguments multiple samples from a single distribution ($K = 1$), the methods used to study either of these overlap significantly. The arguments required in the basic product-form case are simpler than those necessary for the most general case (multiple samples from multiple distributions) and, by focusing on the results that are of greatest interest from the Monte Carlo perspective, we are able to present readily accessible, intuitive, and compact proofs for the theoretical properties of (2). This said, whenever a result given here can be extracted from the U-statistics literature, we provide explicit references.

While U-statistics have been extensively studied since Hoeffding's seminal work (Hoeffding 1948a) and are commonly employed in a variety of statistical tests [e.g., independence tests (Hoeffding 1948b), two-sample tests (Gretton et al. 2012), goodness-of-fit tests (Liu et al. 2016), and more (Lee 1990; Kowalski and Tu 2007)] and learning tasks [e.g., regression (Kowalski and Tu 2007), classification (Clémençon et al. 2008), clustering (Clémençon 2011), and more (Clémençon et al. 2008, 2016)] where they arise as natural estimators, their use in Monte Carlo seems underexplored. Exceptions include Owen (2009) which cleverly applies unisample U-statistics to make the best possible use of a collection of genuine (and hence expensive to obtain and store) uniform random variables and Hall and Marron (1987) that uses them to obtain improved estimates for the integrated squared derivatives of a density.

Product-form estimators themselves can be found peppered throughout the Monte Carlo literature, with one exception (see below), always unnamed and specialized to particular contexts. First off, in the simplest setting of integrating fully factorized functions with respect to product-form measures, it is of course well known that better performance is obtained by separately approximating the marginal integrals and taking their product (although, we have yet to locate full variance expressions quantifying quite how much better, even for this near-trivial case). Beyond the fully factorized case, product-form estimators are found not in isolation but combined with other Monte Carlo methodology: Tran et al. (2013) embeds them within therein-defined importance sampling^2 (IS^2) to efficiently infer parameters of structured latent variable models, Schmon et al. (2020) employs them within pseudo-marginal MCMC to estimate intractable acceptance probabilities for similar models, Lindsten et al. (2017) and Kuntz et al. (2021) study their use within sequential Monte Carlo (SMC), and Aitchison (2019) builds on them to obtain tensor Monte Carlo (TMC), an extension of importance weighted variational autoencoders. The latter article is the aforementioned exception: its author defines the estimators in general and refers to them as 'TMC estimators,' but does not study them theoretically. To the best of our knowledge, there has been no previous systematic exploration of the estimators (2), their theoretical properties, and uses, a gap we intend to fill here. Furthermore, while in simple situations with fully, or almost-fully, factorized test functions [e.g., those in Tran et al. (2013) or Schmon et al. (2020)] it might be clear to most practitioners that employing a product-form estimator is the right thing to do, it may not be quite so immediately obvious how much of a difference this can make and that, in rather precise ways (cf. Theorems 2 and 4), judiciously using product-form estimators is the best thing one can do within Monte Carlo when tackling models with known independence structure but unknown conditional distributions (a common situation in practice). We aim to underscore these points through our analysis and examples.

Lastly, we remark that product-form estimators are reminiscent of classical product cubature rules (Stroud 1971). These are obtained by taking products of quadrature rules and, consequently, require computing sums over N^K points much like for product-form estimators [except for fully, or partially, factorized test functions φ where the cost can be similarly lowered, e.g., p. 24 in Stroud (1971)]. In fact, the high computational cost incurred by these rules for general φ partly motivated the development of more modern numerical integration techniques such as quasi-Monte Carlo (Dick et al. 2013), spare grid methods (Gerstner and Griebel 1998,

2003), and, of course, Monte Carlo itself. That said, we believe that these rules can be used to great effect if one is strategic in their application and the advent of the more modern methods has created many opportunities for such applications, something we intend to exemplify here using their Monte Carlo analogues: product-form estimators.

Paper structure This paper is divided into two main parts (Sects. 2 and 3), each corresponding to one of our two aims, and a discussion of our results, future research directions, and potential applications (Sect. 4).

Section 2 studies product-form estimators and their theoretical properties. In particular, we show that the estimators are strongly consistent, unbiased, and asymptotically normal, and we give expressions for their finite sample and asymptotic variances (Sect. 2.1). We argue that they are more statistically efficient than their conventional counterparts in the sense that they achieve lower variances given the same number of samples (Sect. 2.2). Lastly, we consider their computational cost (Sect. 2.3) and explore the circumstances in which they prove most computationally efficient (Sect. 2.4).

Section 3 gives simple examples illustrating how one may embed product-form estimators within standard Monte Carlo methodology and extend their use beyond product-form targets. In particular, we combine them with importance sampling and obtain estimators applicable to targets that are absolutely continuous with respect to fully factorized distributions (Sect. 3.1) and partially factorized ones (Sect. 3.2), and we consider their use within pseudo-marginal MCMC (Sect. 3.3). We then examine the numerical performance of these extensions on a simple hierarchical model (Sect. 3.4).

This paper has six appendices (provided in the supplementary material). The first five contain proofs: Appendix A those for the basic properties of product-form estimators, Appendix B that for their MVUE property, Appendix C those for the basic properties of the 'partially product-form' estimators introduced in Sect. 3.2, Appendix D that for the latter's MVUE property, and Appendix E that for the statistical efficiency (vis-à-vis their non-product counterparts) of the product-form pseudo-marginal MCMC estimators considered in Sect. 3.3. Appendix F contains an additional, simple extension of product-form estimators (to targets that are mixtures of product-form distributions), omitted from the main text in the interest of brevity.

2 Product-form estimators

Consider the basic Monte Carlo problem: given a probability distribution μ on a measurable space $(S, \mathcal{S})$ and a function φ belonging to the space L^2_μ of square μ-integrable real-valued functions on S, estimate the average

$$\mu(\varphi) := \int \varphi(x)\mu(dx).$$

Throughout this section, we focus on the question 'by exploiting the product-form structure of a target μ, can we design estimators of $\mu(\varphi)$ that are more efficient than the usual ones?'. By product-form, we mean that μ is the product of $K > 1$ distributions $\mu_1, \ldots, \mu_K$ on measurable spaces $(S_1, \mathcal{S}_1), \ldots, (S_K, \mathcal{S}_K)$ satisfying $S = S_1 \times \cdots \times S_K$ and $\mathcal{S} = \mathcal{S}_1 \times \cdots \times \mathcal{S}_K$, where the latter denotes the product sigma-algebra. Furthermore, if A is a non-empty subset of $[K] := \{1, \ldots, K\}$, then we use $\mu_A := \prod_{k\in A} \mu_k$ to denote the product of the μ_ks indexed by ks in A and $\mu_A(\varphi)$ to denote the measurable function on $\prod_{k\notin A} S_k$ obtained by integrating the arguments of φ indexed by ks in A with respect to μ_A:

$$\mu_A(\varphi)(x_{A^c}) := \int \varphi(x_A, x_{A^c})\mu_A(dx_A)$$

for all x_{A^c} in $\prod_{k\in A^c} S_k$, where $A^c := [K]\backslash A$ denotes A's complement, under the assumption that these integrals are well defined. If A is empty, we set $\mu_A(\varphi) := \varphi$.

2.1 Theoretical characterization

Suppose that we have at our disposal N i.i.d. samples $X^1, \ldots, X^N$ drawn from μ. We can view these samples as N tuples

$$(X^1_1, \ldots, X^1_K), \quad \ldots, \quad (X^N_1, \ldots, X^N_K)$$

of i.i.d. samples $X^1_1, \ldots, X^N_1, \ldots, X^1_K, \ldots, X^N_K$ independently drawn from $\mu_1, \ldots, \mu_k$, respectively. As we will see in Sect. 2.2, the *product-form estimator*,

$$\mu^N_\times(\varphi) := \frac{1}{N^K} \sum_{\boldsymbol{n}\in[N]^K} \varphi(X^{\boldsymbol{n}}) \tag{4}$$

where $X^{\boldsymbol{n}}$ with $\boldsymbol{n} = (n_1, \ldots, n_K)$ denotes the 'permuted' tuple $(X^{n_1}_1, \ldots, X^{n_K}_K)$ (i.e., a tuple obtained as one of the N^K component-wise permutations of the original samples), yields lower variance estimates for $\mu(\varphi)$ than the conventional choice using the same samples,

$$\mu^N(\varphi) := \frac{1}{N} \sum_{n=1}^{N} \varphi(X^n), \tag{5}$$

regardless of whether the test function φ possesses any sort of product structure. The conventional estimator directly

approximates the target with the samples' empirical distribution,

$$\mu \approx \frac{1}{N}\sum_{n=1}^{N}\delta_{X^n} =: \mu^N. \tag{6}$$

The product-form estimator instead first approximates the marginals $\mu_1, \ldots, \mu_K$ of the target with the corresponding component-wise empirical distributions,

$$\mu_1^N := \frac{1}{N}\sum_{n=1}^{N}\delta_{X_1^n}, \quad \ldots, \quad \mu_K^N := \frac{1}{N}\sum_{n=1}^{N}\delta_{X_K^n},$$

and then takes the product of these to obtain an approximation of μ,

$$\mu \approx \prod_{k=1}^{K}\left(\frac{1}{N}\sum_{n=1}^{N}\delta_{X_k^n}\right) = \frac{1}{N^K}\sum_{\boldsymbol{n}\in[N]^K}\delta_{X^{\boldsymbol{n}}} =: \mu_\times^N. \tag{7}$$

The built-in product structure in $\mu_\times^N$ makes it a better suited approximation to the product-form target μ than the non-product-form μ^N. Before pursuing this further, we take a moment to show that $\mu_\times^N(\varphi)$ is a well founded estimator for $\mu(\varphi)$ and obtain expressions for its variance.

Theorem 1 *If φ is μ-integrable, then $\mu_\times^N(\varphi)$ in (4) is unbiased:*

$$\mathbb{E}\left[\mu_\times^N(\varphi)\right] = \mu(\varphi) \quad \forall N > 0.$$

If, furthermore, φ belongs to L^2_μ, then $\mu_{A^c}(\varphi)$ belongs to $L^2_{\mu_A}$ for all subsets A of $[K]$. The estimator's variance is given by

$$\begin{aligned}&\mathrm{Var}(\mu_\times^N(\varphi))\\ &\quad= \sum_{\emptyset\neq A\subseteq[K]}\frac{1}{N^{|A|}}\sum_{B\subseteq A}(-1)^{|A|-|B|}\sigma^2_{A,B}(\mu_{A^c}(\varphi)),\end{aligned} \tag{8}$$

for every $N > 0$, where $|A|$ and $|B|$ denote the cardinalities of A and B and

$$\sigma^2_{A,B}(\psi) := \mu_B([\mu_{A\setminus B}(\psi) - \mu_A(\psi)]^2) \tag{9}$$

for all ψ in $L^2_{\mu_A}$ and $B \subseteq A \subseteq [K]$. Furthermore, $\mu_\times^N(\varphi)$ is strongly consistent and asymptotically normal:

$$\lim_{N\to\infty}\mu_\times^N(\varphi) = \mu(\varphi) \ \text{ almost surely}, \tag{10}$$

$$N^{1/2}[\mu_\times^N(\varphi) - \mu(\varphi)] \Rightarrow \mathcal{N}(0, \sigma_\times^2(\varphi)) \ \text{ as } N\to\infty, \tag{11}$$

where $\sigma_\times^2(\varphi) := \sum_{k=1}^K \sigma_k^2(\varphi)$ with

$$\sigma_k^2(\varphi) := \mu_k([\mu_{\{k\}^c}(\varphi) - \mu(\varphi)]^2) \quad \forall k \in [K]$$

and $\Rightarrow$ denotes convergence in distribution.

As mentioned in Sect. 1, product-form estimators are special cases of multisample U-statistics and Theorem 1 can be pieced together from various results in the U-statistics literature. For example, within Korolyuk and Borovskich (1994) one can find the unbiasedness (p. 35), variance expressions (p. 38), consistency (which also holds for μ-integrable φ; Theorem 3.2.1), and asymptotic normality (Theorem 4.5.1). To keep the paper self-contained we include a simple proof of Theorem 1, specially adapted for product-form estimators, in Appendix A. It has two key ingredients, the first being the following decomposition expressing the 'global approximation error' $\mu_\times^N - \mu$ as a sum of products of 'marginal approximation errors' $\mu_1^N - \mu_1, \ldots, \mu_K^N - \mu_K$:

$$\begin{aligned}\mu_\times^N - \mu &= \prod_{k=1}^K \mu_k^N - \mu = \prod_{k=1}^K[(\mu_k^N - \mu_k) + \mu_k] - \mu\\ &= \sum_{\emptyset\neq A\subseteq[K]}\left(\prod_{k\in A}[\mu_k^N - \mu_k]\right)\times\mu_{A^c},\end{aligned} \tag{12}$$

The other is the following expression for the L^2 norm of a generic product of marginal errors [see p. 152 in Korolyuk and Borovskich (1994) for its multisample U-statistics analogue]. It tells us that the product of l of these errors has $\mathcal{O}(N^{-l/2})$ norm, as one would expect given that the errors are independent and that classical theory [e.g., p. 168 in Chopin and Papaspiliopoulos (2020)] tells us that the norm of each is $\mathcal{O}(N^{-1/2})$.

Lemma 1 *If A is a non-empty subset of $[K]$, ψ belongs to $L^2_{\mu_A}$, and $\sigma^2_{A,B}(\psi)$ is as in (9), then*

$$\begin{aligned}&\mathbb{E}\left[\left[\left(\prod_{k\in A}[\mu_k^N - \mu_k]\right)(\psi)\right]^2\right]\\ &= \frac{1}{N^{|A|}}\sum_{B\subseteq A}(-1)^{|A|-|B|}\sigma^2_{A,B}(\psi) \quad \forall N > 0.\end{aligned}$$

Proof This lemma follows from the equation

$$\begin{aligned}&\left(\prod_{k\in A}[\mu_k^N - \mu_k]\right)(\psi)\\ &= \sum_{B\subseteq A}(-1)^{|A|-|B|}\mu_B^N(\mu_{A\setminus B}(\psi))\\ &= \mu_A^N\left(\sum_{B\subseteq A}(-1)^{|A|-|B|}\mu_{A\setminus B}(\psi)\right) =: \mu_A^N(\psi_A)\end{aligned} \tag{13}$$

which, together with (12), is known as *Hoeffding's canonical decomposition* in the U-statistics literature [e.g., p. 38

in Korolyuk and Borovskich (1994)] and *ANOVA-like* elsewhere (Efron and Stein 1981). Similar decomposition are commonplace in the quasi-Monte Carlo literature, e.g., Appendix A in Owen (2013). See Appendix A for the details. □

2.2 Statistical efficiency

The product-form estimator $\mu_\times^N(\varphi)$ in (4) yields the best unbiased estimates of $\mu(\varphi)$ that can be achieved using only the knowledge that μ is product-form and N i.i.d. samples drawn from μ.

Theorem 2 *For any given measurable real-valued function φ on $(S, \mathcal{S})$, $\mu_\times^N(\varphi)$ is an MVUE for $\mu(\varphi)$: if f is a measurable real-valued function on $(S^N, \mathcal{S}^N)$ such that*

$$\mathbb{E}\left[f(X^1, \ldots, X^N)\right] = \mu(\varphi)$$

whenever $X^1, \ldots, X^N$ are i.i.d. with law μ, for all product-form μ on $(S, \mathcal{S})$ satisfying $\mu(|\varphi|) < \infty$, then

$$\mathrm{Var}\left(f(X^1, \ldots, X^N)\right) \geq \mathrm{Var}\left(\mu_\times^N(\varphi)\right).$$

Proof See Appendix B. □

While it is well known that unisample U-statistics are MVUEs [e.g., see Clémençon et al. (2016)], we have been unable to locate an explicit proof that covers the general multisample case and, in particular, that of product-form estimators. Instead, we adapt the argument given in Chapter 1 of Lee (1990) (whose origins trace back to Halmos (1946)) for unisample U-statistics and prove Theorem 2 in Appendix B

Theorem 2 implies that product-form estimators achieve lower variances than their conventional counterparts:

Corollary 1 *If φ belongs to L^2_μ and $\sigma^2(\varphi) := \mu([\varphi - \mu(\varphi)]^2)$ denotes $\mu^N(\varphi)$'s asymptotic variance,*

$$\mathrm{Var}\left(\mu_\times^N(\varphi)\right) \leq \frac{\sigma^2(\varphi)}{N} = \mathrm{Var}\left(\mu^N(\varphi)\right) \quad \forall N > 0,$$
$$\sigma_\times^2(\varphi) \leq \sigma^2(\varphi).$$

Proof See Appendix B. □

In other words, product-form estimators are more statistically efficient than their standard counterparts: using the same number of independent samples drawn from the target, $\mu_\times^N(\varphi)$ achieves a lower variance than $\mu^N(\varphi)$. The reason behind this variance reduction was outlined in Sect. 1: the product-form estimator uses the empirical distribution of the collection $(X^{\boldsymbol{n}})_{\boldsymbol{n}\in[N]^K}$ of permuted tuples as an approximation to μ. Because μ is product-form, each of these permuted tuples is as much a sample drawn from μ as any of the original unpermuted tuples $(X^n)_{n=1}^N$. Hence, product-form estimators transform N samples drawn from μ into N^K samples and, consequently, lower the variance. However, the permuted tuples are not independent and we get a diminishing returns effect: the more permutations we make, the greater the correlations among them, and the less 'new information' each new permutation affords us. For this reason, the estimator variance remains $\mathcal{O}(N^{-1})$, cf. (8), instead of $\mathcal{O}(N^{-K})$ as would be the case for the standard estimator using N^K independent samples. As we discuss in Sect. 4, there is also a pragmatic middle ground here: use $N < M < N^K$ permutations instead of all N^K possible ones. In particular, by choosing these M permutations to be as uncorrelated as possible (e.g., so that they have few overlapping entries), it might be feasible to retain most of the variance reduction while avoiding the full $\mathcal{O}(N^K)$ cost (cf. Kong and Zheng (2021) and references therein for similar feats in the U-statistics literature).

Given that the variances of both estimators are (asymptotically) proportional to each other, we are now faced with the question 'how large might the proportionality constant be?'. If the test function is linear or constant, e.g., $S_1 = \cdots = S_K = \mathbb{R}$ and

$$\varphi(x) = \sum_{k=1}^K x_k, \tag{14}$$

then the two estimators trivially coincide, no variance reduction is achieved, and the constant is one. However, these are the cases in which the standard estimator performs well [e.g., for (14), $\mu^N(\varphi)$'s variance breaks down into a sum of K univariate integrals and, consequently, grows slowly with the dimension K]. However, if the test function includes dependencies between the components, then the proportionality constant can be arbitrarily large and the variance reduction unbounded as the following example illustrates.

Example 1 If $K = 2$, $\mu_1 = \mu_2 = \mathcal{N}(0, 1)$, and $\varphi(x) := 1_{\{\min(x_1, x_2) \geq \alpha\}}(x)$, then

$$\mu(\varphi) = \mu(\varphi^2) = [1 - \Phi(\alpha)]^2,$$
$$\mu_1(\varphi)(x_2) = 1_{\{x_2 \geq \alpha\}}[1 - \Phi(\alpha)],$$

where Φ denotes the CDF of a standard normal, and similarly for $\mu_2(\varphi)(x_1)$. In addition,

$$\mu_1(\mu_2(\varphi)^2) = \mu_2(\mu_1(\varphi)^2) = [1 - \Phi(\alpha)]^3.$$

It then follows that

$$\frac{\sigma^2(\varphi)}{\sigma_\times^2(\varphi)} = \frac{2 - \Phi(\alpha)}{2[1 - \Phi(\alpha)]} \to \infty \quad \text{as} \quad \alpha \to \infty.$$

It is not difficult to glean some intuition as to why the product-form estimator yields far more accurate estimates

than its standard counterpart for large α. In these cases, unpermuted tuples with *both* components greater than α are extremely rare (they occur with probability $[1 - \Phi(\alpha)]^2$) and, until one arises, the standard estimator is stuck at zero (a relative error of 100%). On the other hand, for the product-form estimator to return a nonzero estimate, we only require unpermuted tuples with a single component greater than α, which are generated much more frequently (with probability $1 - \Phi(\alpha)$).

Of particular interest is the case of high-dimensional targets (i.e., large K) for which obtaining accurate estimates of $\mu(\varphi)$ proves challenging. Even though the exact manner in which the variance reduction achieved by the product-form estimator scales with dimension of course depends on the precise target and test function, it is straightforward to gain some insight by revisiting our starting example.

Example 2 Suppose that

$$S_1 = \cdots = S_K, \;\; \mathcal{S}_1 = \cdots = \mathcal{S}_K, \;\; \mu_1 = \cdots = \mu_K = \rho,$$
$$\varphi = \prod_{k=1}^{K} \varphi_k, \;\; \varphi_1 = \cdots = \varphi_K = \psi,$$

for some univariate distribution ρ and test function ψ satisfying $\rho(\psi) \neq 0$. In this case,

$$\begin{aligned}\sigma^2(\varphi) &= \mu(\varphi^2) - \mu(\varphi)^2 = \rho(\psi^2)^K - \rho(\psi)^{2K},\\ \sigma^2_\times(\varphi) &= K\rho([\rho(\psi)^{K-1}[\psi - \rho(\psi)]]^2)\\ &= K\rho(\psi)^{2(K-1)}\rho([\psi - \rho(\psi)]^2)\\ &= CV^2 K \rho(\psi)^{2K}, \end{aligned} \tag{15}$$

where $CV := \sqrt{\rho([\psi - \rho(\psi)]^2)}\big/ |\rho(\psi)|$ denotes the coefficient of variation of ψ w.r.t. ρ. Hence,

$$\begin{aligned}\frac{\sigma^2(\varphi)}{\sigma^2_\times(\varphi)} &= \frac{(\rho(\psi^2)/\rho(\psi)^2)^K - 1}{CV^2 K}\\ &= \frac{(1+CV^2)^K - 1}{CV^2 K} = \frac{1}{K}\sum_{k=0}^{K-1}\binom{K}{k+1} CV^{2k}, \end{aligned} \tag{16}$$

and we see that the reduction in variance grows exponentially with the dimension K.

At first glance, (15) might appear to imply that the number of samples required for $\mu^N_\times(\varphi)$ to yield a reasonable estimate of $\mu(\varphi)$ grows exponentially with K if $|\rho(\psi)| > 1$. However, what we deem a 'reasonable estimate' should take into account the magnitude of the average $\mu(\varphi)$ we are estimating. In particular, it is natural to ask for the standard deviation of our estimates to be $\varepsilon\,|\mu(\varphi)|$ for some prescribed relative tolerance $\varepsilon > 0$. In this case, we find that the number of samples required by the product-form estimator is approximately

$$\sigma^2_\times(\varphi)/(\varepsilon^2\mu(\varphi)^2) = CV^2 K \varepsilon^{-2}.$$

In the case of the conventional estimator $\mu^N(\varphi)$, the number required to achieve the same accuracy is instead

$$\sigma^2(\varphi)/(\varepsilon^2\mu(\varphi)^2) = \varepsilon^{-2}((1+CV^2)^K - 1).$$

That is, the number of samples necessary to obtain a reasonable estimate grows linearly with dimension for $\mu^N_\times(\varphi)$ and exponentially for $\mu^N(\varphi)$.

Notice that the univariate coefficient of variation CV features heavily in Example 2's analysis: the greater it is, the greater the variance reduction, and the difference gets amplified exponentially with the dimension K. This observation might be explained as follows: if μ is highly peaked (so that the coefficient is close to zero), then the unpermuted tuples are clumped together around the peak (Fig. 1a), permuting their entries only yields further tuples around the peak (Fig. 1b), and the empirical average changes little. If, on the other hand, μ is spread out (so that the coefficient is large), then the unpermuted pairs are scattered across the space (Fig. 1c), permuting their entries reveals unexplored regions of the space (Fig. 1d), and the estimates improve. Of course, how spread out the target is must be measured in terms of the test function and we end up with the coefficient of variation in (16).

2.3 Computational efficiency

As shown in the previous section, product-form estimators are always at least as statistically efficient as their conventional counterparts: the variances of the former are bounded above by those of the latter. These gains in statistical efficiency come at a computational cost: even though both conventional and product-form estimators share the same $\mathcal{O}(N)$ memory needs, the latter requires evaluating the test function N^K times, while the former requires only N evaluations. For this reason, the question of whether product-form estimators are more *computationally efficient* than their conventional counterparts (i.e., achieve smaller errors given the same computational budget) is not as straightforward. In short, sometimes but not always.

One way to answer the computational efficiency question is to compare the cost incurred by each estimator in order to achieve a desired given variance σ^2. To do so, we approximate the variance of $\mu^N_\times(\varphi)$ with its asymptotic variance divided by the sample number (as justified by Theorem 1). The number of samples required for the variance to equal σ^2 is $N := \sigma^2(\varphi)/\sigma^2$ for the conventional estimator and (approximately) $N_\times := \sigma^2_\times(\varphi)/\sigma^2$ for the product-form one. The costs of evaluating the former with N samples

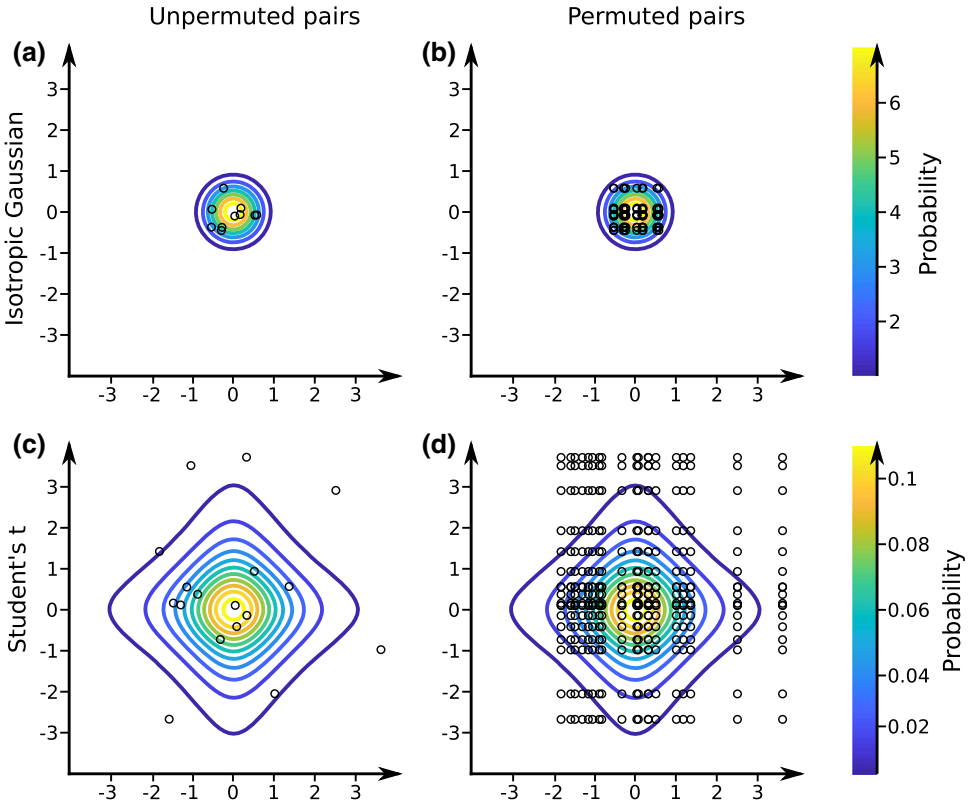

Fig. 1 Ensembles of unpermuted (**a**, **c**) and permuted (**b**, **d**) pairs for a peaked target (**a**, **b**) and heavy tailed one (**c**, **d**). **a** 10 pairs (dots) independently drawn from a two-dimensional isotropic Gaussian (contours) with mean zero and variance 0.1. **b** The 100 pairs (dots) obtained by permuting the pairs in **a**. **c** 20 pairs (dots) independently draw from the product of two student-t distributions (contours) with 1.5 degrees of freedom. **d** The 400 permuted pairs (dots) obtained by permuting the pairs in **c**

and the latter with $N_\times$ samples are $NC_\varphi + NC_X + N$ and $N_\times^K C_\varphi + N_\times C_X + N_\times^K$, respectively, where C_φ and C_X are the costs, relative to that of a single elementary arithmetic operation, of evaluating φ and generating a sample from μ, respectively, and the rightmost N and $N_\times^K$ terms account for the cost of computing the corresponding sample average once all evaluations of φ are carried out. It follows that $\mu_\times^N(\varphi)$ is (asymptotically) at least as computationally efficient as $\mu^N(\varphi)$ if and only if the ratio of their respective costs is no smaller than one or, after some re-arranging,

$$\frac{\sigma^2(\varphi)}{\sigma_\times^2(\varphi)} \geq \frac{(\sigma_\times^2(\varphi)/\sigma^2)^{K-1} C_r + 1}{C_r + 1}, \tag{17}$$

where $C_r := (C_\varphi + 1)/C_X$ denotes the relative cost of evaluating the test function and drawing samples. Our first observation here is that above is always satisfied in the limit $C_r \to 0$ because $\sigma^2(\varphi) \geq \sigma_\times^2(\varphi)$ (Corollary 1). This corresponds the case where the cost of acquiring the samples dwarfs the overhead of evaluating the sample average (for instance, if the samples are obtained from long simulations or real-life experiments). If so, we do really want to make the most of the samples we have and product-form estimators enable us to do so. Conversely, if samples are cheap to generate and the test function is expensive to evaluate (i.e., $C_r \to \infty$), then we are better off using the basic estimator.

To investigate the case where the costs of generating samples and evaluating the test function are comparable ($C_r \approx 1$), note that the variance approximation $\mathrm{Var}\,(\mu_\times^{N_\times}(\varphi)) \approx \sigma_\times^2(\varphi)/N_\times$ and, consequently, (17) are valid only if $\sigma_\times^2(\varphi) > \sigma^2$. Otherwise, $N_\times = 1$ and the product-form estimator simply equals $\varphi(X^1)$ with variance $\sigma^2(\varphi)$. In the high-dimensional (i.e., large K) case which is of particular interest, (17) then (approximately) reduces to

$$\frac{\sigma^2(\varphi)}{\sigma_\times^2(\varphi)} \geq \frac{1}{2}\left(\frac{\sigma_\times^2(\varphi)}{\sigma^2}\right)^{K-1}. \tag{18}$$

To gain insight into whether it is reasonable to expect the above to hold, we revisit Example 2.

Example 3 Setting once again our desired standard deviation to be proportional to the magnitude of the target average (i.e., $\sigma = \varepsilon\,|\mu(\varphi)| = \varepsilon|\rho(\psi)|^K$) and calling on (15, 16), we re-write (18) as

$$\frac{(1+CV^2)^K - 1}{CV^2K} \geq \frac{(CV^2K\varepsilon^{-2})^{K-1}}{2}$$
$$\Leftrightarrow \frac{(1+CV^2)^K - 1}{CV^{2K}} \geq \frac{\varepsilon^2}{2}\left(\frac{K}{\varepsilon^2}\right)^K.$$

The expression shows that, in this full $\mathcal{O}(N^K)$ cost case, $\mu^N(\varphi)$ outperforms $\mu_\times^N(\varphi)$ in computational terms for large

dimensions K (and, even more so, for small relative tolerances ε).

In summary, unless the cost of generating samples is significantly larger than that of evaluating φ, we expect the basic estimator to outperform the product-form one. Simply put, independent samples are more valuable for estimation than correlated permutations thereof. Hence, if independent samples are cheap to generate, then we are better off drawing further independent samples rather than permuting the ones we already have.

That is, unless we can find a way to evaluate the product-form estimator that does not require summing over all N^K permutations. Indeed, the above analysis is out of place for Example 3 because, in this case, we can express the product-form estimator as the product

$$\mu_\times^N(\varphi) = \prod_{k=1}^K \left(\frac{1}{N} \sum_{n=1}^N \psi(X_k^n) \right) = \prod_{k=1}^K \mu_k^N(\psi) \tag{19}$$

of the univariate sample averages $\mu_1^N(\psi), \ldots, \mu_K^N(\psi)$ and evaluate each of these separately at a total $\mathcal{O}(KN)$ cost. Given that the number of samples required for $\mu_\times^N(\varphi)$ to yield a reasonable estimate scales linearly with dimension (Example 2), it follows that the cost incurred by computing such an estimate scales quadratically with dimension. In the case of $\mu^N(\varphi)$, the number of samples required, and hence the cost, scales exponentially with dimension; making the product-form estimator the clear choice for this simple case. This type of trick significantly expands the usefulness of product-form estimators, as we see in the following section.

2.4 Efficient computation

Recall our starting example from Sect. 1. In that case, the product-form estimator trivially breaks down into the product of K sample averages (2) and, consequently, we can evaluate it in $\mathcal{O}(KN)$ operations. We can exploit this trick whenever the test function possesses product-like structure: if φ is a sum

$$\varphi = \sum_{j=1}^J \varphi^j \text{ of products } \varphi^j := \prod_{k=1}^K \varphi_k^j \tag{20}$$

of univariate functions $(\varphi_k^j : S_k \to \mathbb{R})_{j\in[J],k\in[K]}$, the product-form estimator decomposes into a sum of products (SOP) of univariate averages,

$$\mu_\times^N(\varphi) = \sum_{j=1}^J \prod_{k=1}^K \mu_k^N(\varphi_k^j),$$

where

$$\mu_k^N(\varphi_k^j) := \frac{1}{N} \sum_{n=1}^N \varphi_k^j(X_k^n) \;\; \forall j \in [J], \;\; k \in [K],$$

and we are able to evaluate $\mu_\times^N(\varphi)$ in $\mathcal{O}(KN)$ operations. (Of course, 'univariate' need not mean that the function is defined on $\mathbb{R}$ and we can be strategic in our choice of component spaces $S_1, \ldots, S_K$; e.g., if $\varphi(x_1, x_2, x_3) = \varphi_1(x_1, x_2)\varphi_2(x_3)$ for some functions $\varphi_1 : \mathbb{R}^2 \to \mathbb{R}$ and $\varphi_2 : \mathbb{R} \to \mathbb{R}$, we could pick $K := 2$, $S_1 := \mathbb{R}^2$, and $S_2 := \mathbb{R}$.) In these cases, the use of product-form estimators amounts to nothing more than a dimensionality-reduction technique: we exploit the independence of the target to express our K-dimensional integral in terms of an SOP of one-dimensional integrals,

$$\mu(\varphi) = \sum_{j=1}^J \prod_{k=1}^K \mu_k(\varphi_k^j) =: f(\{\mu_k(\varphi_k^j)\}_{j\in[J],k\in[K]}),$$

estimate each of these separately,

$$\mu_k(\varphi_k^j) \approx \mu_k^N(\varphi_k^j) \quad \forall j \in [J], \;\; k \in [K],$$

and replace the one-dimensional integrals in the SOP with their estimates to obtain an estimate for the K-dimensional integral:

$$\mu(\varphi) \approx f(\{\mu_k^N(\varphi_k^j)\}_{j\in[J],k\in[K]}) = \mu_\times^N(\varphi).$$

By so exploiting the structure in μ and φ, the product-form estimator achieves a lower variance than the standard estimator (Corollary 1). Moreover, evaluating each univariate sample average $\mu_k^N(\varphi_k^j)$ requires only $\mathcal{O}(N)$ operations and, consequently the computational complexity of $\mu_\times^N(\varphi)$ is $\mathcal{O}(KN)$. The running time can be further reduced by calculating the univariate sample averages in parallel.

Similar considerations apply if the test function φ is a product of low-dimensional functions (and sums thereof) instead of univariate ones, e.g., $\varphi(x) = \prod_{i=1}^I \varphi_i((x_k)_{k\in A_i})$ for a collection of factors $\varphi_1, \ldots, \varphi_I$ with arguments indexed by subsets $A_1, \ldots, A_I$ of $[K]$. As with the SOP case, one should aim to swap as many summation and product signs in

$$\mu_\times^N(\varphi) = \frac{1}{N^K} \sum_{n_1=1}^N \cdots \sum_{n_K=1}^N \prod_{i=1}^I \varphi_i((X_k^{n_k})_{k\in A_i})$$

as the factors permit. Exactly how best to do this is obvious for simple situations such as that in Example 5 in Sect. 3.1. For more complicated ones, we advice using the 'variable elimination' algorithm [cf. Chapter 9 in Koller and Friedman (2009)] commonly employed for inference in discrete graphical models. The complexity of the resulting procedure

essentially depends on the order in which one attempts the swapping (however, it is easy to find bounds thereon; for instance, it is bounded below by both the maximum cardinality of $A_1, \ldots, A_I$ and half the length of the longest cycle in φ's factor graph). While finding the ordering with lowest complexity for general partially factorized φ itself proves to be a problem whose worst-case complexity is exponential in K, good suboptimal orderings can often be found using cheap heuristics [cf. Sect. 9.4.3 in Koller and Friedman (2009)].

For general φ lacking any sort of product structure, we are sometimes able to extend the linear-cost approach by approximating φ with SOPs (e.g., using truncated Taylor expansions for analytic φ). The idea is that if $\varphi \approx f$ for some SOP f, then

$$\mu(\varphi) \approx \mu(f), \quad \mathrm{Var}\,(\mu_\times^N(\varphi)) \approx \mathrm{Var}\,(\mu_\times^N(f)),$$

and we can use $\mu_\times^N(f) \approx \mu(f)$ as a linear-cost estimator for $\mu(\varphi)$ without significantly affecting the variance reduction. This, of course, comes at the expense of introducing a bias in our estimates, albeit one that can often be made arbitrarily small by using more and more refined approximations [these biases may in principle be removed using multilevel randomization, see McLeish (2011) or Rhee and Glynn (2015)]. The choice of approximation quality itself proves to be a balancing act as more refined approximations typically incur higher evaluation costs. If these costs are high enough, then any potential computational gains afforded by the reduction in variance are lost. In summary, this SOP approximation approach is most beneficial for test functions (a) that are effectively approximated by SOP functions (so that the bias is low), (b) whose SOP approximations are relatively cheap to evaluate (so that the cost is low), and (c) that have a high-dimensional product-form component to them (so that the variance reduction is large, cf. Sect. 2.2). In these cases, the gains in performance can be substantial as illustrated by the following toy example.

Example 4 Let $\mu_1, \ldots, \mu_K$ be uniform distributions on the interval $[0, a]$ of length $a > 1$ and consider the function $\varphi(x) := e^{x_1 \cdots x_K}$. The integral can be expressed in terms of the generalized hypergeometric function ${}_pF_q$,

$$\begin{aligned}\mu(\varphi) &= \sum_{j=0}^{\infty} \frac{\mu_1(x_1^j)\ldots\mu_K(x_K^j)}{j!} = \sum_{j=0}^{\infty} \frac{1}{j!}\left[\frac{a^j}{(j+1)}\right]^K \\ &= {}_K F_K(1,\ldots,1;2,\ldots,2;a^K),\end{aligned}$$

and grows super-exponentially with the dimension K (see Fig. 2a). Because

$$\varphi(x) = e^{x_1\cdots x_K} \approx \sum_{j=0}^{J} \frac{[x_1\ldots x_K]^j}{j!}$$

$$= 1 + \sum_{j=1}^{J} \frac{x_1^j \ldots x_K^j}{j!} =: \varphi_J(x)$$

for large enough truncation cutoffs J, we have that

$$\mu_\times^N(\varphi) \approx \mu_\times^N(\varphi_J) = 1 + \sum_{j=1}^{J} \frac{\mu_1^N(x_1^j)\ldots\mu_K^N(x_K^j)}{j!}.$$

Using $\mu_\times^N(\varphi_J)$ instead of $\mu_\times^N(\varphi)$ as an estimator for $\mu(\varphi)$, we lower the computational cost from $\mathcal{O}(N^K)$ to $\mathcal{O}(KN)$. In exchange, we introduce a bias:

$$\begin{aligned}\mathbb{E}\left[\mu_\times^N(\varphi_J)\right] - \mu(\varphi) &= \mu(\varphi_J) - \mu(\varphi) = \mu(\varphi_J - \varphi) \\ &= \mu\left(\sum_{j=J+1}^{\infty} \frac{x_1^j\ldots x_K^j}{j!}\right) \\ &= \sum_{j=J+1}^{\infty} \frac{\mu_1(x_1^j)\ldots\mu_K(x_K^j)}{j!} \\ &= \sum_{j=J+1}^{\infty} \frac{1}{j!}\left[\frac{a^j}{j+1}\right]^K.\end{aligned}$$

As $\sum_{j=J+1}^{\infty} \frac{a^{jK}}{j!} = o(a^{JK}/J!)$, the bias decays super-exponentially with the cutoff J, at least for sufficiently large J. In practice, we found it to be significant for Js smaller than $0.8a^K$ and negligible for Js larger than $1.2a^K$ (Fig. 2b). In particular, the cutoff J necessary for $\mu_\times^N(\varphi_J)$ to yield estimates with small bias grows exponentially with the dimension K.

Similar manipulations to those above reveal that

$$\begin{aligned}\sigma^2(\varphi) &= {}_K F_K\left(1,\ldots,1;2,\ldots,2;2a^K\right) \\ &\quad - {}_K F_K(1,\ldots,1;2,\ldots,2;a^K)^2 \\ \sigma_\times^2(\varphi_J) &= K\sum_{i=0}^{J}\sum_{j=0}^{J} \frac{1}{i!j!}\frac{ij}{i+j+1}\left(\frac{a^{i+j}}{(i+1)(j+1)}\right)^K\end{aligned}$$

and we find that the variance reduction achieved by $\mu_\times^N(\varphi_J)$ far outpaces the growth in K of the cutoff (and, thus, the computational cost of $\mu_\times^N(\varphi_J)$) necessary to achieve a small bias (Fig. 2c). Indeed, the asymptotic-standard-deviation-to-mean ratio, $\sigma(\varphi)/\mu(\varphi)$, rapidly diverges with K in the case of the standard estimator (Fig. 2d, solid). In that of the biased product-form estimator, the ratio, $\sigma_\times(\varphi_J)/\mu(\varphi)$, also diverges with K but at a much slower rate (Fig. 2d, dashed). For this reason, the number of samples necessary for obtain a, say, 1% accuracy estimate of $\mu(\varphi)$ using $\mu_\times^N(\varphi_J)$ remains manageable for a substantially larger range of as and Ks than in the case of $\mu^N(\varphi)$, even after factoring in the

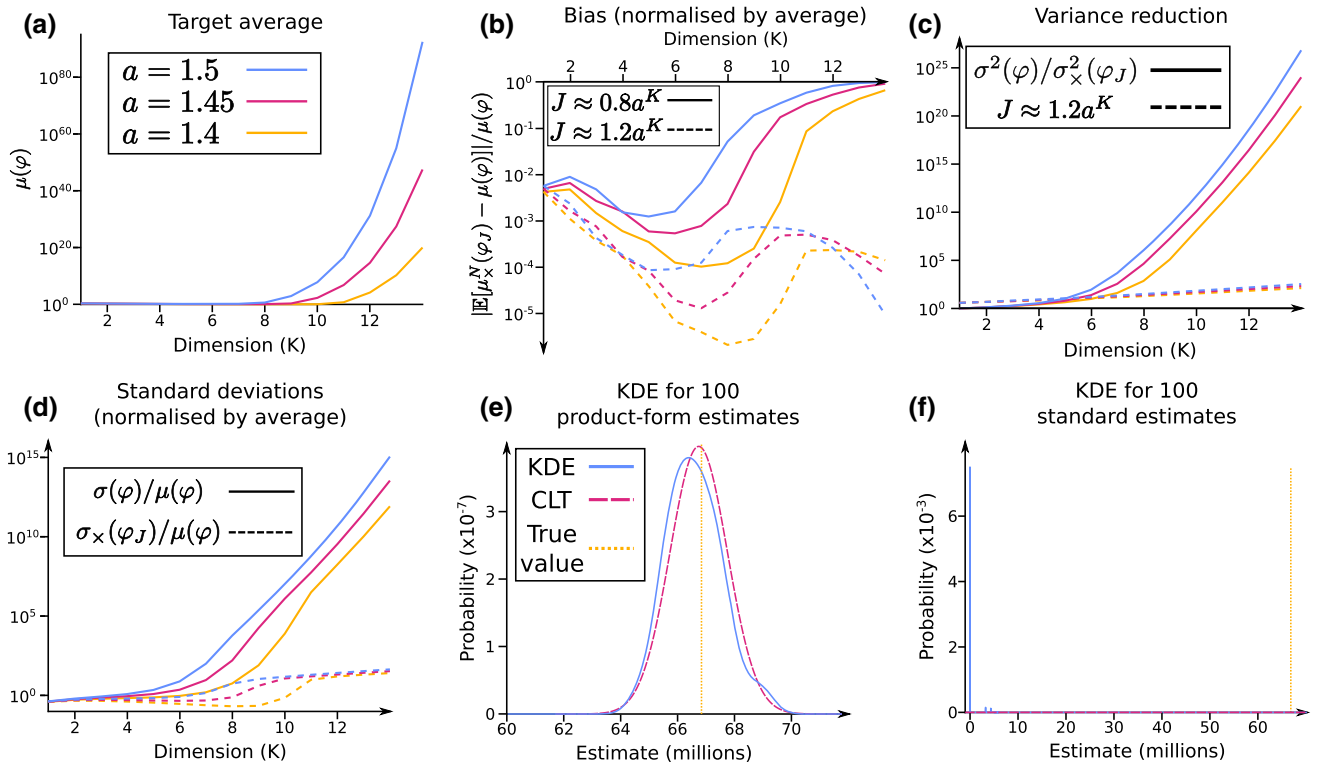

Fig. 2 **a–d** Plots generated for three values of a: $a = 1.4$ (blue), $a = 1.45$ (magenta), and $a = 1.5$ (yellow). **a** Target average $\mu(\varphi)$ as a function of dimension K. **b** Bias of product-form estimator (normalized by target average) as a function of K with truncation cut-offs $J = \lceil 0.8a^K \rceil + 2$ (solid) and $J = \lceil 1.2a^K \rceil + 2$ (dashed). We added the $+2$ to avoid trivial cutoffs for low values of a^K. **c** Ratio of asymptotic variances $\sigma^2(\varphi)/\sigma^2_\times(\varphi_J)$ (solid) with $J = \lceil 1.2a^K \rceil + 2$ (dashed) as a function of K. **d** Asymptotic standard deviation (normalized by target average) for conventional (solid) and biased product-form (dashed, with $J = \lceil 1.2a^K \rceil + 2$) estimators as a function of K. **e** Kernel density estimator with plug-in bandwidth (Wand and Jones 1994) (blue) obtained with $a = 1.5$, $K = 10$, $J = 70$, and 100 repeats of $\mu_\times^N(\varphi_J)$ each involving $N = 10^6$ samples is a good match to the corresponding sampling distribution (magenta) predicted by the CLT in Theorem 1. Comparing with the target average (yellow), we find a mean absolute error across repeats of $6.73 \times 10^5 \approx \mu(\varphi)/100$. **f** As in **e** but for $\mu^N(\varphi)$. This time, the predicted sampling distribution is extremely wide (with a standard deviation of 6.7×10^{14}) and a poor match to the kernel density estimator (almost a Dirac delta close to zero). The mean absolute error is $6.67 \times 10^7 \approx \mu(\varphi)$. The estimator's failure stems from the extreme rarity of samples X^n achieving very large values of $\varphi(X^n)$ (i.e., those with components that are all close to a). Because the components are independent, they are extremely unlikely to simultaneously be close to a and the aforementioned samples are not observed for realistic ensemble sizes N. The product-form estimator avoids this issue by averaging over each component separately. (colour figure online)

extra cost required to evaluate $\mu_\times^N(\varphi_J)$ for J's large enough that the bias is insignificant. For instance, with an interval length of 1.5 and ten dimensions, a cutoff of seventy, one million samples, and less than one minute of computation time suffices for $\mu_\times^N(\varphi_J)$ to produce a 1% accuracy estimate of $\mu(\varphi) \approx 6.68 \times 10^7$ (Fig. 2e). Using the same one million samples and the standard estimator, we obtain very poor estimates (Fig. 2f). Indeed, $\mu^N(\varphi)$'s asymptotic variance equals 4.45×10^{29} and, so, we would need approximately 10^{18} samples for it to yield 1% accuracy estimates, something far beyond current computational capabilities.

3 Extensions to non-product-form targets

While interesting product-form distributions can be found throughout the applied probability literature—ranging from the stationary distributions of Jackson queues (Jackson 1957; Kelly 1979) and complex-balanced stochastic reaction networks (Anderson et al. 2010; Cappelletti and Wiuf 2016) to the mean-field approximations used in variational inference (Ranganath et al. 2014; Blei et al. 2017)—most target distributions encountered in practice are not product-form. In this section, we demonstrate how to combine product-form estimators with other Monte Carlo methodology and expand their utility beyond the product-form case.

We consider three simple extensions: one to targets that are absolutely continuous with respect to fully factorized distributions (Sect. 3.1), resulting in a product-form vari-

ant of importance sampling [e.g., see Chapter 8 in Chopin and Papaspiliopoulos (2020)]; another to targets that are absolutely continuous with respect to partially factorized distributions (Sect. 3.2), resulting in a product-form version of importance sampling squared (Tran et al. 2013); and a final one to targets with intractable densities arising from latent variable models (Sect. 3.3), resulting in a product-form variant of pseudo-marginal MCMC (Schmon et al. 2020). In all cases, we show theoretically that the product-form variants achieve smaller variances than their standard counterparts. We then investigate their performance numerically by applying them to a simple hierarchical model (Sect. 3.4).

A further extension, this time to targets that are mixtures of product-form distributions, can be found in Appendix F. Because many distributions may be approximated with these mixtures, this extension potentially opens the door to tackling still more complicated targets (at the expense of introducing some bias).

3.1 Importance sampling

Suppose that we are given an unnormalized (but finite) unsigned target measure γ that is absolutely continuous with respect to the product-form distribution μ in Sect. 2, and let $w := d\gamma/d\mu$ be the corresponding Radon–Nikodym derivative. Instead of the usual important sampling (IS) estimator, $\gamma^N(\varphi) := \mu^N(w\varphi)$ with μ^N as in (6), for $\gamma(\varphi)$, we consider its product-form variant, $\gamma_\times^N(\varphi) := \mu_\times^N(w\varphi)$ with $\mu_\times^N$ as in (7). The results of Sect. 2 immediately give us the following.

Corollary 2 *If φ is γ-integrable, then $\gamma_\times^N(\varphi)$ is an unbiased estimator for $\gamma(\varphi)$. If, furthermore, $w\varphi$ lies in L^2_μ, then $\gamma_\times^N(\varphi)$ is strongly consistent, asymptotically normal, and its finite sample and asymptotic variances are bounded above by those of $\gamma^N(\varphi)$:*

$$\begin{aligned}\mathrm{Var}\,(\gamma_\times^N(\varphi)) &= \mathrm{Var}\,(\mu_\times^N(w\varphi))\\ &\le \mathrm{Var}\,(\mu^N(w\varphi))\\ &= \mathrm{Var}\,(\gamma^N(\varphi))\ \forall N > 0,\\ \sigma^2_{\gamma,\times}(\varphi) = \sigma^2_\times(w\varphi) &\le \sigma^2(w\varphi) = \sigma^2_\gamma(\varphi),\end{aligned}$$

where $\mathrm{Var}\,(\mu_\times^N(w\varphi))$ *and* $\sigma^2_\times(w\varphi)$ *are as in Theorem 1.*

Proof Replace φ with $w\varphi$ in Theorem 1 and Corollary 1. □

Corollary 2 tells us that $\gamma_\times^N(\varphi)$ is more statistically efficient than the conventional IS estimator $\gamma^N(\varphi)$ regardless of whether the target γ is product-form or not. In a nutshell, $\mu_\times^N$ is a better approximation to the proposal μ than μ^N and, consequently, $\gamma_\times^N(dx) = w(x)\mu_\times^N(dx)$ is a better approximation to $\gamma(dx) = w(x)\mu(dx)$ than $\gamma^N(dx) = w(x)\mu^N(dx)$. Indeed, by constructing all N^K permutations of the tuples $X^1, \ldots, X^N$, we explore other areas of the state space. This can be particularly useful when the proposal and target are mismatched as it can amplify the number of tuples landing in the target's high probability regions (i.e., achieving high weights w) and, consequently, substantially improve the quality of the finite sample approximation (Fig. 3).

Similarly, the self-normalized version $\pi_\times^N(\varphi) := \gamma_\times^N(\varphi)/\gamma_\times^N(S)$ of the product-form IS estimator $\gamma_\times^N(\varphi)$ is a consistent and asymptotically normal estimator for averages $\pi(\varphi)$ with respect to the normalized target $\pi := \gamma/\gamma(S)$. As in the case of the standard self-normalized importance sampling (SNIS) estimator $\pi^N(\varphi) := \gamma^N(\varphi)/\gamma^N(S)$, the ratio in $\pi_\times^N(\varphi)$'s definition introduces an $\mathcal{O}(N^{-1})$ bias and stops us from obtaining analytical expression for the finite sample variance (that the bias is $\mathcal{O}(N^{-1})$ follows from an argument similar to that given for standard SNIS in p. 35 of Liu (2001) and requires making assumptions on the higher moments of $\varphi(X^1)$). Otherwise, $\pi_\times^N(\varphi)$'s theoretical properties are analogous to those of the product-form estimator $\mu_\times^N(\varphi)$ and its importance sampling extension $\gamma_\times^N(\varphi)$:

Corollary 3 *If $w\varphi$ lies in L^2_μ, then $\pi_\times^N(\varphi)$ is strongly consistent, asymptotically normal, and its asymptotic variance is bounded above by that of $\pi^N(\varphi)$:*

$$\begin{aligned}\sigma^2_{\pi,\times}(\varphi) &= \sigma^2_\times(\gamma(S)^{-1}w[\varphi - \pi(\varphi)])\\ &\le \sigma^2(\gamma(S)^{-1}w[\varphi - \pi(\varphi)]) = \sigma^2_\pi(\varphi),\end{aligned}$$

where $\sigma^2_\times(\gamma(S)^{-1}w[\varphi - \pi(\varphi)])$ is as in Theorem 1.

Proof Given Theorem 1 and Corollary 1, the arguments here follow closely those for standard SNIS. In particular, because $\pi_\times^N(\varphi) = \gamma_\times^N(\varphi)/\gamma_\times^N(S) = \mu_\times^N(w\varphi)/\mu_\times^N(w)$ and $\mu(w) = \gamma(S)$,

$$\begin{aligned}\pi_\times^N(\varphi) - \pi(\varphi) &= \frac{\mu_\times^N(w\varphi)}{\mu_\times^N(w)} - \pi(\varphi)\\ &= \frac{\mu(w)}{\mu_\times^N(w)}\mu_\times^N\left(\frac{w[\varphi - \pi(\varphi)]}{\gamma(S)}\right)\\ &= \frac{\mu(w)}{\mu_\times^N(w)}\mu_\times^N(w^\pi[\varphi - \pi(\varphi)]).\end{aligned}$$

Given that $\mu_\times^N(w)$ tends to $\mu(w)$ almost surely (and, hence, in probability) as N approaches infinity (Theorem 1), the strong consistency and asymptotic normality of $\pi_\times^N(\varphi)$ then follow from those of $\mu_\times^N(\gamma(S)^{-1}w[\varphi - \pi(\varphi)])$ (Theorem 1) and Slutsky's theorem. The asymptotic variance bound follows from that in Corollary 1. □

This type of approach is best suited for targets π possessing at least some product structure. The structure manifest itself in partially factorized weight functions w and substantially lowers the evaluation costs of $\gamma_\times^N(\varphi)$ and $\pi_\times^N(\varphi)$ for

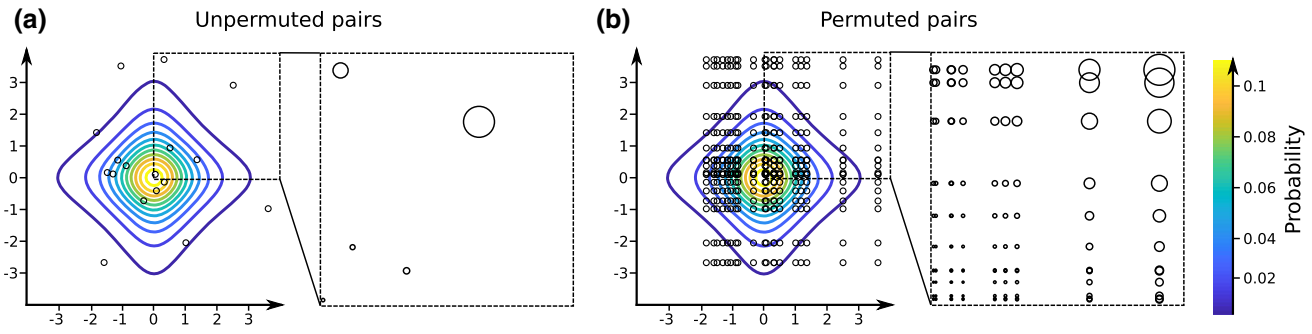

Fig. 3 Product-form approximations improve state space exploration. The target, a uniform distribution on $[0, 4]^2$ (dashed square), and the proposal, the product of two student-t distributions with 1.5 degrees of freedom (contours), are mismatched. Consequently, only 5 of 20 pairs independently drawn from the proposal land within the target's support (**a**) and the corresponding weighted sample approximation (**a**, inset, dot diameter proportional to sample weight) is poor. By permuting these pairs, we improve the coverage of the state space (**b**), increase the number of pairs lying within the target's support, and obtain a much better weighted sample approximation (**b**, inset, dot diameter proportional to sample weight)

simple test functions φ, as the following example illustrates.

Example 5 (A simple hierarchical model) Consider the following basic hierarchical model:

$$Y_k \sim \mathcal{N}(X_k, 1) \quad X_k \sim \mathcal{N}(0, \theta), \quad \forall k \in [K]. \tag{21}$$

It has a single unknown parameter, the variance θ of the latent variables $X_1, \ldots, X_K$, which we infer using a Bayesian approach. That is, we choose a prior $p(d\theta)$ on θ and draw inferences from the corresponding posterior,

$$\begin{aligned}\pi(d\theta, dx) &:= p(d\theta, dx|y) \\ &\propto p(d\theta) \prod_{k=1}^{K} \mathcal{N}(y_k; x_k, 1)\mathcal{N}(dx_k; 0, \theta) =: \gamma(d\theta, dx),\end{aligned} \tag{22}$$

where $y = (y_1, \ldots, y_K)$ denotes the vector of observations. For most priors, no analytic expressions for the normalizing constant can be found and we are forced to proceed numerically. One option is to choose the proposal

$$\mu(d\theta, dx) := p(d\theta) \prod_{k=1}^{K} \mathcal{N}(dx_k; 0, 1), \tag{23}$$

in which case

$$w_{IS}(\theta, x) := \frac{d\gamma}{d\mu}(\theta, x) = \prod_{k=1}^{K} \frac{\mathcal{N}(y_k; x_k, 1)\mathcal{N}(x_k; 0, \theta)}{\mathcal{N}(x_k; 0, 1)}.$$

(Were we to be using standard IS instead of product-form variant, the proposal

$$\mu(d\theta, dx) := p(d\theta) \prod_{k=1}^{K} \mathcal{N}(dx_k; 0, \theta) \tag{24}$$

would be the natural choice, a point we return to after the example.) Hence, to estimate the normalizing constant or any integral w.r.t. to a univariate marginal of the posterior, we need to draw samples from μ and evaluate the product-form estimator $\mu_{\times}^{N}(\varphi)$ for a test function of the form $\varphi(\theta, x) = f(\theta) \prod_{k=1}^{K} g_k(\theta, x_k)$, the cost of which totals $\mathcal{O}(KN^2)$ operations because

$$\mu_{\times}^{N}(\varphi) = \frac{1}{N^{K+1}} \sum_{m=1}^{N} f(\theta^m) \prod_{k=1}^{K} \sum_{n_k=1}^{N} g_k(\theta^m, x_k^{n_k}).$$

We return to this in Sect. 3.4, where we will make use of the following expression for the (unnormalized) posteriors's θ-marginal available due to the Gaussianity in (21):

$$\gamma(d\theta) = p(d\theta) \prod_{k=1}^{K} \mathcal{N}(y_k; 0, \theta + 1). \tag{25}$$

Clearly, the above expression opens the door to simpler and more effective methods for computing integrals with respect to this marginal than estimators targeting the full posterior. However, the estimators we discuss can be applied analogously to the many commonplace hierarchical models [e.g., see Gelman and Hill (2006), Gelman (2006), Koller and Friedman (2009), Hoffman et al. (2013), Blei et al. (2003), and the many references therein] for which such expressions are not available.

When applying IS, or extensions thereof like SMC, one should choose the proposal to be as close as possible to the target [e.g., see Agapiou et al. (2017)]. In this regard, the product-form IS approach is not entirely satisfactory for the above example: by definition, the proposal must be fully factorized while the target, π in (22), is only partially so (the latent variables are independent only when conditioned on the parameter variable). As we show in the next section, it is straightforward to adapt this product-form IS approach to match such partially factorized targets.

3.2 Partially factorized targets and proposals

Consider a target or proposal μ over a product space $(\Theta \times S, \mathcal{T} \times \mathcal{S})$ with the same partial product structure as the target in Example 5:

$$\mu(d\theta, dx) = (\mu_0 \otimes \mathcal{M})(d\theta, dx) := \mu_0(d\theta) \prod_{k=1}^{K} \mathcal{M}_k(\theta, dx_k), \tag{26}$$

where, for each k in [K], ... $\theta \mapsto \mathcal{M}_k(\theta, dx_k)$ denotes a Markov kernel mapping from $(\Theta, \mathcal{T})$ to $(S_k, \mathcal{S}_k)$. Suppose that we are given M i.i.d. samples $\theta^1, \ldots, \theta^M$ drawn from μ_0 and, for each of these, N (conditionally) i.i.d. samples $X^{m,1}, \ldots, X^{m,N}$ drawn from the product kernel $\mathcal{M}(\theta, dx) := \prod_{k=1}^{K} \mathcal{M}_k(\theta, dx_k)$ evaluated at θ^m. Given a test function φ on $\Theta \times S$, consider the following 'partially product-form' estimator for $\mu(\varphi)$:

$$\mu_{\times}^{M,N}(\varphi) := \frac{1}{M} \sum_{m=1}^{M} \left(\frac{1}{N^K} \sum_{\boldsymbol{n} \in [N]^K} \varphi(\theta^m, X^{m,\boldsymbol{n}}) \right) = \frac{1}{MN^K} \sum_{m=1}^{M} \sum_{\boldsymbol{n} \in [N]^K} \varphi(\theta^m, X^{m,\boldsymbol{n}}) \tag{27}$$

for all $M, N > 0$. It is well founded (for simplicity, we only consider the estimator's asymptotics as $M \to \infty$ with N fixed, but other limits can be studied by combining the approaches in Appendices A and C.

Theorem 3 *If φ is μ-integrable with μ as in (26), then $\mu_{\times}^{M,N}(\varphi)$ in (27) is unbiased and strongly consistent: for all $N > 0$,*

$$\mathbb{E}\left[\mu_{\times}^{M,N}(\varphi)\right] = \mu(\varphi) \quad \forall M > 0,$$
$$\lim_{M \to \infty} \mu_{\times}^{M,N}(\varphi) = \mu(\varphi) \text{ almost surely.}$$

If, furthermore, φ belongs to L^2_{μ}, then $\mathcal{M}_{[K]\setminus A}(\varphi)$ belongs to $L^2_{\mu_0 \otimes \mathcal{M}_A}$ for all subsets A of $[K]$, where $\mathcal{M}_A(\theta, dx_A) := \prod_{k \in A} \mathcal{M}_k(\theta, dx_k)$, and the estimator is asymptotically normal: for all $N > 0$, and as $M \to \infty$,

$$M^{1/2}[\mu_{\times}^{M,N}(\varphi) - \mu(\varphi)] \Rightarrow \mathcal{N}(0, \sigma^2_{\times,N}(\varphi)), \tag{28}$$

where $\Rightarrow$ denotes convergence in distribution and

$$\sigma^2_{\times,N}(\varphi) := \mu_0([\mathcal{M}\varphi - \mu(\varphi)]^2) + \sum_{\emptyset \neq A \subseteq [K]} \sum_{B \subseteq A} \frac{(-1)^{|A|-|B|} \mu_0(\mathcal{M}_B[\mathcal{M}_{[K]\setminus B}\varphi - \mathcal{M}\varphi]^2)}{N^{|A|}}.$$

For any $N, M > 0$, the estimator's variance is given by $\mathrm{Var}\,(\mu_{\times}^{M,N}(\varphi)) = \sigma^2_{\times,N}(\varphi)/M$.

Proof See Appendix C. □

The partially product-form estimator (27) is more statistically efficient than its standard counterpart.

Corollary 4 *For any φ belonging to L^2_{μ} and $N > 0$,*

$$\mathrm{Var}\,(\mu_{\times}^{M,N}(\varphi)) \leq \mathrm{Var}\,(\mu^{M,N}(\varphi)) \ \forall M > 0,$$
$$\sigma^2_{\times,N}(\varphi) \leq \sigma^2_N(\varphi),$$

where $\mu^{M,N}(\varphi) := \frac{1}{MN} \sum_{m=1}^{M} \sum_{n=1}^{N} \varphi(\theta^n, X^{m,n})$ and $\sigma^2_N(\varphi)$ denotes its asymptotic (in M) variance.

Proof See Appendix C. □

In fact, modulo a small caveat (cf. Remark 1 below), $\mu_{\times}^{M,N}(\varphi)$ yields the best unbiased estimates of $\mu(\varphi)$ achievable using only the knowledge that μ is partially factorized and M i.i.d. samples drawn from $\mu_0 \otimes \mathcal{M}^N$: a perhaps unsurprising fact given that it is the composition of two minimum variance unbiased estimators (Theorem 2).

Theorem 4 *Suppose that $\mathcal{T}$ contains all singleton sets (i.e., $\{\theta\}$ for all θ in Θ). For any given measurable real-valued function φ on $\Theta \times S$, $\mu_{\times}^{M,N}(\varphi)$ is a minimum variance unbiased estimator for $\mu(\varphi)$: if f is a measurable real-valued function on $(\Theta \times S^N)^M$ such that*

$$\mathbb{E}\left[f((\theta^m, X^{m,1}, \ldots, X^{m,N})_{m=1}^{M}) \right] = \mu(\varphi)$$

whenever $(\theta^m, X^{m,1}, \ldots, X^{m,N})_{m=1}^{M}$ is an i.i.d. sequence drawn from $\mu_0 \otimes \mathcal{M}^N$, for all partially factorized $\mu = \mu_0 \otimes \mathcal{M}$ on $\Theta \times S$ satisfying $\mu(|\varphi|) < \infty$ and

$$\mu_0(\{\theta\}) = 0 \quad \forall \theta \in \Theta, \tag{29}$$

then

$$\mathrm{Var}\,(f((\theta^m, X^{m,1}, \ldots, X^{m,N})_{m=1}^{M})) \geq \mathrm{Var}\,(\mu_{\times}^{M,N}(\varphi)).$$

Proof See Appendix D. □

Remark 1 (*The importance of* (29)) Consider the extreme scenario that μ_0 is a Dirac delta at some θ^*, so that $\theta^1 = \cdots = \theta^M = \theta^*$ with probability one and

$$\mu_\times^{M,N}(\varphi) = \frac{1}{M}\sum_{m=1}^{M}\frac{1}{N^K}\sum_{\boldsymbol{n}\in[N]^K}\varphi(\theta^*, X^{m,\boldsymbol{n}}) \quad \text{a.s.}$$

In this case, we are clearly better off (at least in terms estimator variance) stacking all of our X samples into one big ensemble and replacing the partially product-form estimator with the (fully) product-form estimator,

$$\mu_\times^{MN}(\varphi) = \frac{1}{(MN)^K}\sum_{\boldsymbol{l}\in[MN]^K}\varphi(\theta^*, \tilde{X}^{\boldsymbol{l}}),$$

where $(\tilde{X}^l)_{l\in[MN]}$ denotes $(X^{m,n})_{m\in[M],n\in[N]}$ in vectorized form (indeed Theorem 2 implies that $\mu_\times^{MN}(\varphi)$ is a minimum variance unbiased estimator in this situation). More generally, note that, because

$$\begin{aligned}\mu_0^2(\{\theta^1 = \theta^2\}) &= \int 1_{\{\theta^1=\theta^2\}}\mu_0^2(d\theta^1, d\theta^2)\\ &= \int\left(\int 1_{\{\theta^1=\theta^2\}}\mu_0(d\theta^1)\right)\mu_0(d\theta^2)\\ &= \int \mu_0(\{\theta\})\mu_0(d\theta),\end{aligned}$$

μ_0 not possessing atoms, i.e., (29), is equivalent to $\mu_0^2(\{\theta^1 = \theta^2\}) = 0$. It is then straightforward to argue that (29) is equivalent to the impossibility of several θ^m coinciding or, in other words, to

$$\mu_0^M(\{\theta^i \neq \theta^j \ \forall i \neq j\}) = 1. \tag{30}$$

Were this not to be the case, the estimator in (27) would not possess the MVUE property. To recover it, we would need to amend the estimator as follows: 'if several θ^ms take the same value, first stack their corresponding $X^{m,1}, \ldots, X^{m,N}$ samples, and then apply a product-form estimator to the stacked samples.' However, to not overly complicate this section's exposition and Theorem 4's proof, we restrict ourselves to distributions satisfying (29).

We are now in a position to revisit Example 5 and better adapt the proposal to the target. This leads to a special case of an algorithm known as 'importance sampling squared' or 'IS^2', cf. Tran et al. (2013).

Example 6 (*A simple hierarchical model, revisited*) Consider again the model in Example 5. Recall that our previous choice of proposal did not quite capture the conditional independence structure in the target π: the former was fully factorized while the latter is only partially so. It seems more natural to instead use the proposal in (24) which is also easy to sample from but both mirrors π's independence structure and leads to further cancellations in the weight function (in particular, it no longer depends on θ):

$$w_{IS^2}(x) := \prod_{k=1}^{K}\mathcal{N}(y_k; x_k, 1) = \frac{d\gamma}{d\mu}(\theta, x).$$

It follows that, to estimate the normalizing constant or any integral w.r.t. to a univariate marginal of the posterior, we need to draw samples from $\mu_0 \otimes \mathcal{M}^N$ and evaluate the partially product-form estimator $\mu_\times^{M,N}(\varphi)$ for a test function of the form $\varphi(\theta, x) = f(\theta)\prod_{k=1}^{K} g_k(x_k)$. Because

$$\mu_\times^{M,N}(\varphi) = \frac{1}{MN^K}\sum_{m=1}^{M} f(\theta^m)\prod_{k=1}^{K}\sum_{n_k=1}^{N} g_k(X_k^{m,n_k}),$$

the total cost then reduces to $\mathcal{O}(KMN)$. We also return to this is Sect. 3.4.

3.3 Grouped independence Metropolis–Hastings

As a further example of how one may embed product-form estimators within more sophisticated Monte Carlo methodology and exploit the independence structure present in the problem, we revisit Beaumont's Grouped Independence Metropolis–Hastings [GIMH (Beaumont 2003)], a simple and well known pseudo-marginal MCMC sampler (Andrieu and Roberts 2009). Like many of these samplers, it is intended to tackle targets whose densities cannot be evaluated pointwise but are marginals of higher-dimensional distributions whose densities can be evaluated pointwise. Our inability to evaluate the target's density precludes us from directly applying the Metropolis–Hastings algorithm (MH, e.g., see Chapter XIII in Asmussen and Glynn (2007)) as we cannot compute the necessary acceptance probabilities. For instance, in the case of a target $\pi(d\theta)$ on a space $(\Theta, \mathcal{T})$ and an MH proposal $Q(\theta, d\tilde{\theta})$ with respective densities $\pi(\theta)$ and $Q(\theta, \tilde{\theta})$, we would need to evaluate

$$1 \wedge \frac{\pi(\tilde{\theta})Q(\theta, \tilde{\theta})}{\pi(\theta)Q(\tilde{\theta}, \theta)}$$

where θ denotes the chain's current state and $\tilde{\theta} \sim Q(\theta, \cdot)$ the proposed move. GIMH instead replaces the intractable $\pi(\theta)$ and $\pi(\tilde{\theta})$ in the above with importance sampling estimates thereof: if $\pi(\theta, x)$ denotes the density of the higher-dimensional distribution $\pi(d\theta, dx)$ whose θ-marginal is

$\pi(d\theta)$, and $w(\theta, x) := \pi(\theta, x)/\mathcal{M}(\theta, x)$ for a given Markov kernel $\mathcal{M}(\theta, dx)$ with density $\mathcal{M}(\theta, x)$,

$$\pi^N(\theta) = \frac{1}{N}\sum_{n=1}^{N} w(\theta, X^n),$$
$$\pi^N(\tilde{\theta}) = \frac{1}{N}\sum_{n=1}^{N} w(\tilde{\theta}, \tilde{X}^n), \tag{31}$$

where $X^1, \ldots, X^N$ and $\tilde{X}^1, \ldots, \tilde{X}^N$ are i.i.d. samples drawn from $\mathcal{M}(\theta, \cdot)$ and $\mathcal{M}(\tilde{\theta}, \cdot)$, respectively Key in Beaumont's approach is that the samples are recycled from one iteration to another: if $Z^1, \ldots, Z^N$ and $\tilde{Z}^1, \ldots \tilde{Z}^N$ denote the i.i.d. samples used in the previous iteration, then $(X^1, \ldots, X^N) := (Z^1, \ldots, Z^N)$ if the previous move was rejected and $(X^1, \ldots, X^M) := (\tilde{Z}^1, \ldots, \tilde{Z}^N)$ if it was accepted.

As explained in Andrieu and Roberts (2009) [see also Andrieu and Vihola (2015)], the algorithm's correctness does not require the density estimates to be generated by (31), only for them to be unbiased. In particular, if the estimates are unbiased, GIMH may be interpreted as an MH algorithm on an expanded state space with an extension of $\pi(d\theta)$ as its invariant distribution. Consequently, provided that the density estimator is suitably well behaved, GIMH returns consistent and asymptotically normal estimates of the target under conditions comparable to those for standard MH algorithms [e.g., the GIMH chain is uniformly ergodic whenever the associated 'marginal' chain is and the estimator is uniformly bounded (Andrieu and Roberts 2009); see Andrieu and Vihola (2015) for further refinements]. Consequently, if the kernel is product-form (i.e., $\mathcal{M}(\theta, dx)$ is product-form for each θ), we may replace the estimators in (31) with their product-form counterparts:

$$\pi_{\times}^N(\theta) = \frac{1}{N^K}\sum_{\boldsymbol{n}\in[N]^K} w(\theta, X^{\boldsymbol{n}}),$$
$$\pi_{\times}^N(\tilde{\theta}) = \frac{1}{N^K}\sum_{\boldsymbol{n}\in[N]^K} w(\tilde{\theta}, \tilde{X}^{\boldsymbol{n}}), \tag{32}$$

where K denotes the dimensionality of the x-variables (the unbiasedness follows from $X^{\boldsymbol{n}}$ and $\tilde{X}^{\boldsymbol{n}}$ having respective laws $\mathcal{M}(\theta, dx)$ and $\mathcal{M}(\tilde{\theta}, d\tilde{x})$ for any $\boldsymbol{n}$ in $[N]^K$). Thanks to the results in Andrieu and Vihola (2016), it is straightforward to show that this choice leads to lower estimator variances, at least asymptotically.

Corollary 5 *Let $(\theta^{m,N})_{m=1}^{\infty}$ and $(\theta_{\times}^{m,N})_{m=1}^{\infty}$ denote the GIMH chains generated using (31) and (32), respectively, and the same proposal $Q(\theta, d\theta)$. If φ belongs to L_{π}^2, then*

$$\lim_{M\to\infty} Var\left(\frac{1}{\sqrt{M}}\sum_{m=1}^{M}\varphi(\theta_{\times}^{m,N})\right)$$
$$\leq \lim_{M\to\infty} Var\left(\frac{1}{\sqrt{M}}\sum_{m=1}^{M}\varphi(\theta^{m,N})\right) \quad \forall N > 0.$$

Proof See Appendix E. □

Given the argument used in the proof, the results of Andrieu and Vihola (2016), Theorem 10 in particular, imply much more than the variance bound in the corollary's statement. For instance, if the target is not concentrated on points, then the spectral gap of $(\theta_{\times}^{m,N})_{m=1}^{\infty}$ is bounded below by that of $(\theta^{m,N})_{m=1}^{\infty}$. We finish the section by returning to our running example.

Example 7 (*A simple hierarchical model, re-visited*) Here, we follow Sect. 5.1 in Schmon et al. (2020). Consider once again the model in Example 5 and suppose we are interested only in the posterior's θ-marginal $\pi(d\theta)$. Choosing

$$\mathcal{M}(\theta, dx) := \prod_{k=1}^{K}\mathcal{N}(dx_k; 0, \theta),$$

the weight function factorizes,

$$w_{GIMH}(\theta, x) = \frac{\pi(\theta, x)}{\mathcal{M}(\theta, x)} = p(\theta)\prod_{k=1}^{K}\mathcal{N}(y_k; x_k, 1);$$

resulting in an evaluation cost of $\mathcal{O}(KN)$ for (31, 32) and, regardless of which density estimates we use, a total cost of $\mathcal{O}(KMN)$ where M denotes the number of steps we run the chain for. We return to this in the following section.

3.4 Numerical comparison

Here, we apply the estimators discussed throughout Sects. 3.1–3.3 to the simple hierarchical model introduced in Example 5 and we examine their performance. To benchmark the latter, we choose the prior to be conditionally conjugate to the model's likelihood: $p(d\theta)$ is the Inv-Gamma$(\alpha/2, \alpha\beta/2)$ distribution, in which case

$$X_k|y_k, \theta \sim \mathcal{N}\left(\frac{y_k}{\theta^{-1}+1}, \frac{1}{\theta^{-1}+1}\right) \ \forall k \in [K],$$
$$\theta|y, X \sim \text{Inv-Gamma}\left(\frac{\alpha+K}{2}, \frac{\alpha\beta+\sum_{k=1}^{K}X_k^2}{2}\right);$$

and we can alternatively approximate the posterior, $\pi(d\theta, dx)$ in (22), using a Gibbs' sampler. Note that the

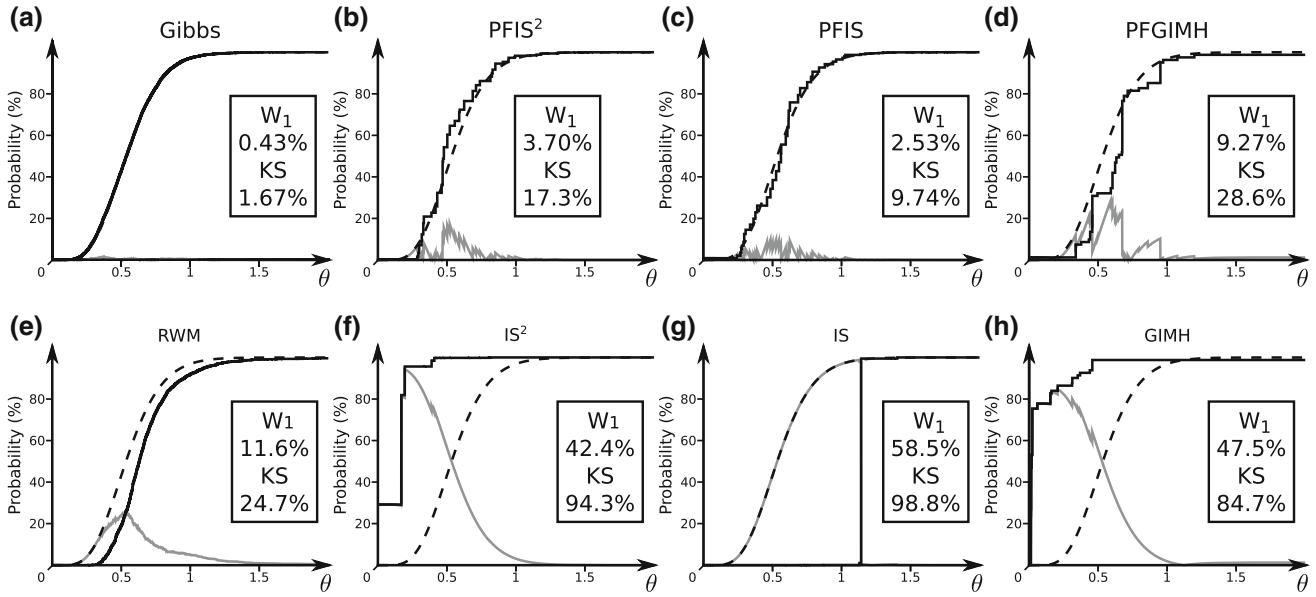

Fig. 4 Empirical cumulative density functions for $\pi(d\theta)$ obtained using the eight approximations discussed in the text (black solid lines). As guides, we also plot a high-quality approximation $\pi_{REF}(d\theta)$ using (25) and quadrature, and the pointwise absolute difference between π_{REF} and the eight approximations (grey lines). *Insets* Wasserstein-1 distance (W_1) between π_{REF} and the panel's approximation (area under the grey line, e.g., see p. 64 in Shorack and Wellner (2009)) and corresponding Kolmogorov–Smirnov statistic (KS, maximum of the grey line)

above expressions are unnecessary for the evaluation of the estimators in Sects. 3.1–3.3. To compare with standard methodology that also does not requires such expressions, we also approximate the posterior using Random Walk Metropolis (RWM) with the proposal variance tuned so that the mean acceptance probability (approximately) equals 25%. To keep the comparison honest, we run these two chains for N^2 steps and set $M = N$ for the estimators in Sects. 3.2 and 3.3 ; in which case all estimators incur a similar $\mathcal{O}(KN^2)$ cost. We further fix $K := 100$, $\alpha := 1$, $\beta := 1$, and $N := 100$ and generate artificial observations $y_1, \ldots, y_{100}$ by running (21) with $\theta := 1$.

Figure 4 shows approximations to the posteriors's θ-marginal $\pi(d\theta)$ obtained using a Gibbs sampler, RWM, IS (Sect. 3.1), IS^2 (Sect. 3.2), GIMH (Sect. 3.3), and the last three's product-form variants (PFIS, PFIS^2, and PFGIMH, respectively). In the cases of Gibbs, RWM, GIMH, and PFGIMH, we used a 20% burn-in period and approximated the marginal with the empirical distribution of the θ-components of the states visited by the chain. For GIMH and PFGIMH, we also used a random walk proposal with its variance tuned so that the mean acceptance probability hovered around 25%. For IS, PFIS, IS^2, and PFIS^2, we used the proposals specified in Examples 5 and 6 and computed the approximations using

$$\pi_{IS}^{N^2}(d\theta) := \frac{\sum_{n=1}^{N^2} w_{IS}(\theta^n, X^n)\delta_{\theta^n}}{\sum_{n=1}^{N^2} w_{IS}(\theta^n, X^n)},$$

$$\pi_{PFIS}^{N}(d\theta) := \frac{\sum_{n=1}^{N}\left(\sum_{\boldsymbol{n}\in[N]^K} w_{IS}(\theta^n, X^{\boldsymbol{n}})\right)\delta_{\theta^n}}{\sum_{n=1}^{N}\sum_{\boldsymbol{n}\in[N]^K} w_{IS}(\theta^n, X^{\boldsymbol{n}})},$$

$$\pi_{IS^2}^{N,N}(d\theta) := \frac{\sum_{m=1}^{N}\left(\sum_{n=1}^{N} w_{IS^2}(X^{m,n})\right)\delta_{\theta^m}}{\sum_{m=1}^{N}\sum_{n=1}^{N} w_{IS^2}(X^{m,n})},$$

$$\pi_{PFIS^2}^{N,N}(d\theta) := \frac{\sum_{m=1}^{N}\left(\sum_{\boldsymbol{n}\in[N]^K} w_{IS^2}(X^{m,\boldsymbol{n}})\right)\delta_{\theta^m}}{\sum_{m=1}^{N}\sum_{\boldsymbol{n}\in[N]^K} w_{IS^2}(X^{m,\boldsymbol{n}})}.$$

(Note that for IS, we are using N^2 samples instead of N so that its cost is also $\mathcal{O}(KN^2)$.)

Our first observation is that the approximations produced by IS, IS^2, and GIMH are very poor. The first two exhibit severe weight degeneracy (in either case, a single particle had over 50% of the probability mass and three had over 90%), something unsurprising given the target's moderately high dimension of 101.[2] The third possesses a pronounced spurious peak close to zero (with over 70% of the mass) caused by large numbers of rejections in that vicinity. Replacing the

[2] One may wonder whether in the case of IS, the degeneracy could instead be due to our use of the proposal (23) rather than the more natural choice (24). It is not: the average W_1 distance and KS statistic (see Fig. 4's caption for definitions) across 100 replicates of the $\pi(d\theta)$'s approximation obtained using (24) and IS (with N^2 samples) were 32.7 and 82.3%, respectively. In other words, a modest improvement over IS with proposal (23) (compare with Table 1), but not one sufficient to break the degeneracy: 83 approximations (out of 100) had at least 50% of their mass concentrated in 2 particles (out of 10, 000) and all but 7 had over 80% of their mass concentrated in 10 particles.

Table 1 Average-across repeats W_1 error and KS statistic for the approximations of $\pi(d\theta)$, and average absolute errors for the corresponding mean and standard deviation estimates, obtained using each of the eight methods

	Gibbs (%)	PFIS^2 (%)	PFIS (%)	PFGIMH (%)	RWM (%)	IS^2 (%)	IS (%)	GIMH (%)
W_1	0.73	4.41	5.33	7.95	9.99	32.7	36.0	39.9
KS	1.93	17.9	19.2	24.9	23.3	82.3	79.9	71.8
Mean error	1.08	4.51	7.84	8.00	16.2	53.6	60.4	70.6
Standard deviation error	1.51	8.07	9.01	17.1	21.0	64.4	64.0	26.4

Table 2 Total absolute error for the mean and standard deviation estimates of $\pi(dx)$'s univariate marginals

	Gibbs	PFIS^2	PFIS	RWM	IS^2	IS
Mean	56	169	511	1209	3295	5727
Standard deviation	40	124	332	663	2874	3340

Note that no results are given for GIMH and PFGIMH since these algorithms directly target the θ-marginal $\pi(d\theta)$

i.i.d. estimators embedded within these algorithms with their product-form counterparts removes both the weight degeneracy and the spurious peak; PFIS, PFIS^2, and PFGIMH return much improved approximations. The best approximation is the one returned by the Gibbs sampler: an expected outcome given that the sampler's use of the conditional distributions makes it the estimator most 'tailored' or 'well adapted' to the target. However, these distributions are not available for most models (precluding application of these samplers to such models) and even just taking the, usually obvious, independence structure into account can make a substantial difference: the quality of the approximations returned by PFIS and PFIS^2 exceeds the quality of that returned by the common, or even default, choice of RWM. Note that this is the case even though the proposal variance in RWM was tuned, while that in the other two was simply set to 1 (a reasonable choice given that $\theta = 1$ was used to generate the data, but likely not the optimal one). In fact, for this simple model, it is easy to sensibly incorporate observations into the PFIS and PFIS^2 proposals [e.g., use $p(d\theta)\prod_{k=1}^{K}\mathcal{N}(dx_k; y_k, 1)$ for PFIS and $p(d\theta)\prod_{k=1}^{K}\mathcal{N}(dx_k; y_k\theta[1+\theta]^{-1}, \theta[1+\theta]^{-1})$ for PFIS^2] and potentially improve their performance.

To benchmark the approaches more thoroughly, we generated $R := 100$ replicates of the eight full posterior approximations and computed various error metrics (Tables 1 and 2). For the θ-component, we used the high-quality reference approximation π_{REF} described in Fig. 4's caption to obtain the average (across repeats) W_1 distance and KS statistic (as described in the caption), and the average absolute error of the posterior mean and standard deviation estimates normalized by the true mean or standard deviation (i.e., $M_\theta^{-1}R^{-1}\sum_{r=1}^{R}\left|M_\theta^r - M_\theta\right|$ for the posterior mean estimates, where M_θ denotes the true mean and M_θ^r the r^{th} estimate thereof, and similarly for the standard deviation estimates). For the x-components, we instead used high-accuracy estimates for the component-wise means and standard deviations (obtained by running a Gibbs sampler for $N^4 = 10^8$ steps) to compute the corresponding total absolute errors across replicates and components ($\sum_{k=1}^{K}\sum_{r=1}^{R}\left|M_k^r - M_k\right|$, where M_k denotes the true mean for the k^{th} x-component and M_k^r the rth estimate thereof, and similarly for the standard deviation estimates).

Once again, the product-form estimators far outperformed their i.i.d. counterparts. Moreover, they perform just as well or better than RWM. PFIS^2's estimates are particularly accurate: a fact that does not surprise us given that its proposal has the same partially factorized structure as the target, in this sense making it the best adjusted estimator to the problem. That is, best except for the Gibbs sampler which exploits the conditional distributions (encoding more information than this structure). We conclude with an interesting detail: PFIS^2 and PFIS perform similarly when approximating the θ-marginal (cf. Table 1), but PFIS^2 outperforms PFIS when approximating the latent variable marginals (cf. Table 2). This is perhaps not too surprising because, in the case of the θ-marginal approximation, both PFIS^2 and PFIS employ the same number N of θ-samples, while, in that of k^{th} latent variable, PFIS^2 uses N^2 x_k-samples and PFIS uses only N such samples.

4 Discussion

The main message of this paper is that when using Monte Carlo estimators to tackle problems possessing some sort of product structure, one should endeavor to exploit this structure and improve the estimators' performance. The resulting product-form estimators are not a panacea for the curse of dimensionality in Monte Carlo, but they are a useful and sometimes overlooked tool in the practitioner's arsenal and make certain problems solvable when they otherwise would not be. More specifically, whenever the target, or proposal, we are drawing samples from is product-form, these estimators achieve a smaller variance than their conventional counterparts. In our experience (e.g., Examples 2 and 4), the gap in variance grows exponentially with dimension

whenever the integrand does not decompose into a sum of low-dimensional functions like in the trivial case (14). For the reasons given in Sect. 2.2, we expect the variance reduction to be further accentuated by targets that are 'spread out' rather than peaked.

The gains in statistical efficiency come at a computational price: in the absence of exploitable structure in the test function, product-form estimators incur an $\mathcal{O}(N^K)$ cost limiting their applicability targets of dimension $K \leq 10$, while conventional estimators only carry an $\mathcal{O}(N)$ cost (although in practice the cost of obtaining reasonable estimates using the latter often scales poorly with K, with the effect hidden in the proportionality constant, e.g., Examples 2 and 4). Hence, for general test functions, product-form estimators are of most use when the variance reduction is particular pronounced or when samples are expensive to acquire (both estimators require drawing the same number N of samples) or store [as, for example, when one employs physical random numbers and requires reproducibility Owen (2009)]. In the latter case, product-form estimators enable us to extract the most possible from the samples we have gathered so far: by permuting the samples' components, the estimators artificially generate further samples. Of course, the more permutations we make, the more correlated our sample ensemble becomes and we get a diminishing returns effect that results in an $\mathcal{O}(N^{-1/2})$ rate of convergence instead of the $\mathcal{O}(N^{-K/2})$ rate we would achieve using N^K independent samples. There is a middle ground here that remains unexplored: using $N < M < N^K$ permutations instead of all N^K possible, so lowering the cost to $\mathcal{O}(M)$ at the expense of some of the variance reduction [see Lin et al. (2005) or Lindsten et al. (2017) for similar ideas in the Monte Carlo literature]. In particular, by choosing the M permutations so that the correlations among them are minimized (e.g., the M permutations with least overlap among their components), it might be possible to substantially reduce the cost without sacrificing too much of the variance reduction. Indeed, by setting the number M of permutations to be such that M evaluations of the test function incurs a cost comparable to that of generating the N unpermuted tuples, one can ensure that the overall cost of the resulting estimator never greatly exceeds that of the conventional estimator. This type of approach has been studied in the sparse grid literature (Gerstner and Griebel 1998) and is closely related to the theory of incomplete U-statistics (cf. Chapter 4.3 in Lee (1990)), an area in which there are ongoing efforts directed at designing good reduced-cost estimators [e.g., see Kong and Zheng (2021)].

There are, however, settings in which product-form estimators should be applied without hesitation: if the integrand is a sum of products (SOP) of univariate functions, the cost comes down to $\mathcal{O}(N)$ without affecting the variance reduction (Sect. 2.4). For instance, when estimating ELBO gradients to optimize mean-field approximations (Ranganath et al. 2014) of posteriors e^v with SOP potentials v. More generally, if the test function is a sum of partially factorized functions, the estimators' evaluation costs can often be substantially reduced (see also Sect. 2.4) so that the variance reduction far outweighs the more mild increases in cost. For instance, as we saw with the applications of importance sampling and its product-form variant in Sect. 3.4.

For integrands lacking this sort of structure, and at the expense of introducing some bias, these types of cost reductions can sometimes be retained if one is able to find a good SOP approximation to the integrand (Example 4). How to construct these approximations for generic functions (or for function classes of interest in given applications) is an open question upon whose resolution the success of this type of approach hinges. In reality, combining product-form estimators with SOP approximations amounts to nothing more than an approximate dimensionality reduction technique: we approximate a high-dimensional integral with a linear combination of products of low-dimension integrals, estimate each of the latter separately, and plug the estimates back into the linear combination to obtain an estimate of the original integral. It is certainly not without precedents: for instance, Rahman and Xu (2004), Ma and Zabaras (2009), Gershman et al. (2012), and Braun and McAuliffe (2010) all propose, in rather different contexts, similar approximations except that the low-dimensional integrals are computed using closed-form expressions or quadrature (for a very well known example, see the delta method for moments in Oehlert (1992)). In practice, the best option will likely involve a mix of these: use closed-form expressions where available, quadrature where possible, and Monte Carlo (or Quasi Monte Carlo) for everything else.

About the computational resources required to evaluate product-form estimators, and the allocation thereof, we ought to mention one interesting variant of the estimators that we omitted from the main text to keep the exposition simple. Throughout we assumed that the same number of samples are drawn from each marginal $\mu_1, \ldots, \mu_K$ of the product-form target or proposal μ. This need not be the case: straightforward extensions of our arguments show that the estimator

$$\mu_{\times}^{N_1,\ldots,N_K}(\varphi) := \frac{1}{\prod_{k=1}^{K} N_k} \sum_{n_1=1}^{N_1} \cdots \sum_{n_K=1}^{N_K} \varphi(X_1^{n_1}, \ldots, X_K^{n_K})$$

behaves much as (4) does, even if a different number of samples N_k are used per marginal μ_k. This variant potentially allows us to concentrate our computational budget on 'the most important dimensions,' an idea that has found significant success in other areas of numerical integration [e.g., see Gerstner and Griebel (1998, 2003) or Owen (1998)]. In our case, this could be done using the pertinent generalizations of the variance expressions in Theorem 1, which are identi-

cal except that $N^{|A|}$ therein must be replaced by $\prod_{k\in A} N_k$ (these can be obtained by retracing the steps in the theorem's proof). In particular, one could estimate the terms in these expressions and adjust the sample sizes so that the estimator variance is minimized, potentially in an iterative manner leading to an adaptive scheme.

Combining product-form estimators with other Monte Carlo methodology expands their utility beyond product-form targets. We illustrated this in Sect. 3 by describing the three simplest and most readily accessible combinations we could think of: their merger with importance sampling applicable to targets that are absolutely continuous with respect to fully factorized distributions (Sect. 3.1), that with importance sampling squared applicable to targets that are absolutely continuous with respect to partially factorized distributions (Sect. 3.2, see also Tran et al. (2013)), and that with pseudo-marginal MCMC applicable to targets with intractable densities (Sect. 3.3, see also Schmon et al. (2020)). In all of these cases, we demonstrated theoretically that the resulting estimators are more statistically efficient than their standard counterparts (Corollaries 2–5). Many other extensions are possible. For instance, one can embed product-form estimators within random weight particle filters (Rousset and Doucet 2006; Fearnhead et al. 2008, 2010)—and, more generally, algorithms reliant on unbiased estimation—much the same way we did for IS^2 and GIMH in Sects. 3.2–3.3. For an example of a slightly different vein, see Appendix F where we consider 'mixture-of-product-form' estimators applicable to targets which are mixtures of product-form distributions and, by combining these with importance sampling, we obtain a product-form version of (stratified) mixture importance sampling estimators (Oh and Berger 1993; Hesterberg 1995) that is particularly appropriate for multimodal targets. For further examples, see the divide-and-conquer SMC algorithm (Lindsten et al. 2017; Kuntz et al. 2021) obtained by combining product-form estimators with SMC and Tensor Monte Carlo (Aitchison 2019) obtained by merging the estimators with variational autoencoders.

When choosing among the resulting (and at times bewildering) constellation of estimators, we recommend following one simple principle: pick estimators that somehow 'resemble' or 'mirror' the target. Good examples of this are well parametrized Gibbs samplers which generate new samples using the target's exact conditional distributions and, consequently, often outperform other Monte Carlo algorithms (e.g., Sect. 3.4). While for many targets these conditional distributions cannot be obtained (nor are good parametrizations known), their (conditional) independence structure is usually obvious [e.g., see Gelman and Hill (2006), Gelman (2006), Koller and Friedman (2009), Hoffman et al. (2013), Blei et al. (2003), and the many references therein] and can be mirrored using product-form estimators within one's methodology of choice. Indeed, in the case of the simple hierarchical model (Example 5), it was the $PFIS^2$ estimator utilizing samples with exactly the same independence structure as the model's that performed best (besides the Gibbs sampler). Of course, this model's independence structure was particularly simple, and so were the resulting estimators. However, we believe that broadly the same considerations apply to models with more complex structures and that product-form estimators can be adapted to such structures by following analogous steps.

To summarize, we believe that product-form estimators are of greatest use not on their own, but embedded within more complicated Monte Carlo routines to tackle the aspects of the problem exhibiting product structure. There remains much work to be done in this direction.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11222-021-10069-9.

Acknowledgements JK and AMJ acknowledge support from the Engineering and Physical Sciences Research Council (EPSRC; Grant # EP/T004134/1) and the Lloyd's Register Foundation Programme on Data-Centric Engineering at the Alan Turing Institute. FRC acknowledges support from the EPSRC and the Medical Research Council OXWASP Centre for Doctoral Training (Grant # EP/L016710/1). FRC and AMJ acknowledge further support from the EPSRC (Grant # EP/R034710/1).

References

Agapiou, S., Papaspiliopoulos, O., Sanz-Alonso, D., et al.: Importance sampling: intrinsic dimension and computational cost. Stat. Sci. **32**(3), 405–431 (2017). https://doi.org/10.1214/17-STS611

Aitchison, L.: Tensor Monte Carlo: particle methods for the GPU era. Adv. Neural Inf. Process. Syst. **32**, 7148–7157 (2019)

Anderson, D.F., Craciun, G., Kurtz, T.G.: Product-form stationary distributions for deficiency zero chemical reaction networks. Bull. Math. Biol. **72**(8), 1947–1970 (2010). https://doi.org/10.1007/s11538-010-9517-4

Andrieu, C., Roberts, G.O.: The pseudo-marginal approach for efficient Monte Carlo computations. Ann. Stat. **37**(2), 697–725 (2009). https://doi.org/10.1214/07-AOS574

Andrieu, C., Vihola, M.: Convergence properties of pseudo-marginal Markov chain Monte Carlo algorithms. Ann. Appl. Probab. **25**(2), 1030–1077 (2015). https://doi.org/10.1214/14-AAP1022

Andrieu, C., Vihola, M.: Establishing some order amongst exact approximations of MCMCs. Ann. Appl. Probab. **26**(5), 2661–2696 (2016). https://doi.org/10.1214/15-AAP1158

Asmussen, S., Glynn, W.: Stochastic Simulation: Algorithms and Analysis. Springer, New York (2007). https://doi.org/10.1007/978-0-387-69033-9

Beaumont, M.A.: Estimation of population growth or decline in genetically monitored populations. Genetics **164**(3), 1139–1160 (2003). https://doi.org/10.1093/genetics/164.3.1139

Bengtsson, T., Bickel, P., Li, B.: Curse-of-dimensionality revisited: collapse of the particle filter in very large scale systems. In: Probability and Statistics: Essays in Honor of David A. Freedman, vol. 2, pp. 316–334. Institute of Mathematical Statistics (2008). https://doi.org/10.1214/193940307000000518

Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)

Blei, D.M., Kucukelbir, A., McAuliffe, J.D.: Variational inference: a review for statisticians. J. Am. Stat. Assoc. **112**(518), 859–877 (2017). https://doi.org/10.1080/01621459.2017.1285773

Braun, M., McAuliffe, J.: Variational inference for large-scale models of discrete choice. J. Am. Stat. Assoc. **105**(489), 324–335 (2010). https://doi.org/10.1198/jasa.2009.tm08030

Cappelletti, D., Wiuf, C.: Product-form Poisson-like distributions and complex balanced reaction systems. SIAM J. Appl. Math. **76**(1), 411–432 (2016). https://doi.org/10.1137/15M1029916

Chopin, N., Papaspiliopoulos, O.: An Introduction to Sequential Monte Carlo. Springer, Cham, (2020). https://doi.org/10.1007/978-3-030-47845-2

Clémençon, S.: On U-processes and clustering performance. Adv. Neural. Inf. Process. Syst. **24**, 37–45 (2011)

Clémençon, S., Lugosi, G., Vayatis, N.: Ranking and empirical minimization of U-statistics. Ann. Stat. **36**(2), 844–874 (2008). https://doi.org/10.1214/009052607000000910

Clémençon, S., Colin, I., Bellet, A.: Scaling-up empirical risk minimization: optimization of incomplete U-statistics. J. Mach. Learn. Res. **17**(76), 1–36 (2016)

Dick, J., Kuo, F.Y., Sloan, I.H.: High-dimensional integration: the quasi-Monte Carlo way. Acta Numer. **22**, 133–288 (2013). https://doi.org/10.1017/S0962492913000044

Efron, B., Stein, C.: The Jackknife estimate of variance. Ann. Stat. **9**(3), 586–596 (1981). https://doi.org/10.1214/aos/1176345462

Fearnhead, P., Papaspiliopoulos, O., Roberts, G.O.: Particle filters for partially observed diffusions. J. R. Stat. Soc. Ser. B Methodol. **70**(4), 755–777 (2008). https://doi.org/10.1111/j.1467-9868.2008.00661.x

Fearnhead, P., Papaspiliopoulos, O., Roberts, G.O., et al.: Random-weight particle filtering of continuous time processes. J. R. Stat. Soc. Ser. B Methodol. **72**(4), 497–512 (2010). https://doi.org/10.1111/j.1467-9868.2010.00744.x

Gelman, A.: Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). Bayesian Anal. **1**(3), 515–534 (2006). https://doi.org/10.1214/06-BA117A

Gelman, A., Hill, J.: Data Analysis using Regression and Multilevel/Hierarchical Models. Cambridge University Press (2006). https://doi.org/10.1017/CBO9780511790942

Gershman, S.J., Hoffman, M.D., Blei, D.M.: Nonparametric variational inference. In: Proc. 29th Int. Conf. Mach. Learn., pp. 235–242 (2012)

Gerstner, T., Griebel, M.: Numerical integration using sparse grids. Numer. Algorithms **18**(3), 209–232 (1998). https://doi.org/10.1023/A:1019129717644

Gerstner, T., Griebel, M.: Dimension-adaptive tensor-product quadrature. Computing **71**(1), 65–87 (2003). https://doi.org/10.1007/s00607-003-0015-5

Gretton, A., Borgwardt, K.M., Rasch, M.J., et al.: A kernel two-sample test. J. Mach. Learn. Res. **13**(25), 723–773 (2012)

Hall, P., Marron, J.S.: Estimation of integrated squared density derivatives. Stat. Probab. Lett. **6**(2), 109–115 (1987). https://doi.org/10.1016/0167-7152(87)90083-6

Halmos, P.R.: The theory of unbiased estimation. Ann. Math. Stat. **17**(1), 34–43 (1946). https://doi.org/10.2307/2235902

Hesterberg, T.: Weighted average importance sampling and defensive mixture distributions. Technometrics **37**(2), 185–194 (1995). https://doi.org/10.1080/00401706.1995.10484303

Hoeffding, W.: A class of statistics with asymptotically normal distribution. Ann. Math. Stat. **19**(3), 293–325 (1948). https://doi.org/10.1214/aoms/1177730196

Hoeffding, W.: A non-parametric test of independence. Ann. Math. Stat. **19**(4), 546–557 (1948). https://doi.org/10.2307/2236021

Hoffman, M.D., Blei, D.M., Wang, C., et al.: Stochastic variational inference. J. Mach. Learn. Res. **14**(1), 1303–1347 (2013)

Jackson, J.R.: Networks of waiting lines. Oper. Res. **5**(4), 518–521 (1957). https://doi.org/10.1287/opre.5.4.518

Kelly, F.P.: Reversibility and Stochastic Networks, 1st edn. Wiley, Chichester (1979)

Koller, D., Friedman, N.: Probabilistic Graphical Models: Principles and Techniques. The MIT Press (2009)

Kong, X., Zheng, W.: Design based incomplete U-statistics. Stat. Sin. **31**(3), 1593–1618 (2021). https://doi.org/10.5705/ss.202019.0098

Korolyuk, V.S., Borovskich, Y.V.: Theory of U-Statistics. Springer (1994)

Kowalski, J., Tu, X.M.: Modern Applied U-Statistics. Wiley-Blackwell (2007). https://doi.org/10.1002/9780470186466

Kuntz J, Crucinio FR, Johansen AM (2021) The divide-and-conquer sequential Monte Carlo algorithm: theoretical properties and limit theorems. ArXiv preprint arXiv:2110.15782

Lee, A.J.: U-Statistics: Theory and Practice. CRC Press (1990)

Lin, M.T., Zhang, J.L., Cheng, Q., et al.: Independent particle filters. J. Am. Stat. Assoc. **100**(472), 1412–1421 (2005). https://doi.org/10.1198/016214505000000349

Lindsten, F., Johansen, A.M., Naesseth, C.A., et al.: Divide-and-Conquer with sequential Monte Carlo. J. Comput. Graph. Stat. **26**(2), 445–458 (2017). https://doi.org/10.1080/10618600.2016.1237363

Liu, J.S.: Monte Carlo Strategies in Scientific Computing. Springer, New York (2001). https://doi.org/10.1007/978-0-387-76371-2

Liu, Q., Lee, J., Jordan, M.: A kernelized Stein discrepancy for goodness-of-fit tests. In: Proc. 33rd Int. Conf. Mach. Learn., pp. 276–284 (2016)

Ma, X., Zabaras, N.: An adaptive hierarchical sparse grid collocation algorithm for the solution of stochastic differential equations. J. Comput. Phys. **228**(8), 3084–3113 (2009). https://doi.org/10.1016/j.jcp.2009.01.006

McLeish, D.: A general method for debiasing a Monte Carlo estimator. Monte Carlo Methods Appl. **17**(4), 301–315 (2011). https://doi.org/10.1515/mcma.2011.013

Oehlert, G.W.: A note on the delta method. Am. Stat. **46**(1), 27–29 (1992). https://doi.org/10.1080/00031305.1992.10475842

Oh, M.S., Berger, J.O.: Integration of multimodal functions by Monte Carlo importance sampling. J. Am. Stat. Assoc. **88**(422), 450–456 (1993). https://doi.org/10.1080/01621459.1993.10476295

Owen, A.B.: Monte Carlo extension of quasi-Monte Carlo. In: Proc. 1998 Winter Simul. Conf., pp. 571–577 (1998). https://doi.org/10.1109/WSC.1998.745036

Owen, A.B.: Recycling physical random numbers. Electron. J. Stat. **3**, 1531–1541 (2009). https://doi.org/10.1214/09-EJS541

Owen, A.B.: Monte Carlo theory, methods and examples (2013). https://statweb.stanford.edu/~owen/mc/

Rahman, S., Xu, H.: A univariate dimension-reduction method for multi-dimensional integration in stochastic mechanics. Probab. Eng. Mech. **19**(4), 393–408 (2004). https://doi.org/10.1016/j.probengmech.2004.04.003

Ranganath, R., Gerrish, S., Blei, D.: Black box variational inference. In: Proc. 17th Int. Conf. Artif. Intell. Stat., pp. 814–822 (2014)

Rhee, C.H., Glynn, P.W.: Unbiased estimation with square root convergence for SDE models. Oper. Res. **63**(5), 1026–1043 (2015). https://doi.org/10.1287/opre.2015.1404

Rousset, M., Doucet, A.: Discussion of "Exact and computationally efcient likelihood-based estimation for discretely observed diffusion processes" by Beskos, Papaspiliopoulos, Roberts and Fearnhead. J. R. Stat. Soc. Ser. B Methodol. **68**(3), 375–376 (2006). https://doi.org/10.1111/j.1467-9868.2006.00552.x

Schmon, S.M., Deligiannidis, G., Doucet, A., et al.: Large-sample asymptotics of the pseudo-marginal method. Biometrika **108**(1), 37–51 (2020). https://doi.org/10.1093/biomet/asaa044

Shorack, G.R., Wellner, A.W.: Empirical processes with applications to statistics. SIAM (2009). https://doi.org/10.1137/1.9780898719017

Silverman, B.W.: Density estimation for statistics and data analysis, Monographs on Statistics and Applied Probability, vol. 26. CRC Press (1986)

Snyder, C., Bengtsson, T., Bickel, P., et al.: Obstacles to high-dimensional particle filtering. Mon. Weather Rev. **136**(12), 4629–4640 (2008). https://doi.org/10.1175/2008MWR2529.1

Stroud, A.H.: Approximate Calculation of Multiple Integrals. Prentice-Hall (1971)

Tran MH, Scharth M, Pitt MK, et al (2013) Importance sampling squared for Bayesian inference in latent variable models. ArXiv preprint arXiv:1309.3339

Wand, M.P., Jones, M.C.: Multivariate plug-in bandwidth selection. Comput. Stat. **9**(2), 97–116 (1994)