On Nesting Monte Carlo Estimators

Tom Rainforth [1] **Robert Cornish** [1 2] **Hongseok Yang** [3] **Andrew Warrington** [2] **Frank Wood** [4]

Abstract

Many problems in machine learning and statistics involve nested expectations and thus do not permit conventional Monte Carlo (MC) estimation. For such problems, one must nest estimators, such that terms in an outer estimator themselves involve calculation of a separate, nested, estimation. We investigate the statistical implications of nesting MC estimators, including cases of multiple levels of nesting, and establish the conditions under which they converge. We derive corresponding rates of convergence and provide empirical evidence that these rates are observed in practice. We further establish a number of pitfalls that can arise from naïve nesting of MC estimators, provide guidelines about how these can be avoided, and lay out novel methods for reformulating certain classes of nested expectation problems into single expectations, leading to improved convergence rates. We demonstrate the applicability of our work by using our results to develop a new estimator for discrete Bayesian experimental design problems and derive error bounds for a class of variational objectives.

1 Introduction

Monte Carlo (MC) methods are used throughout the quantitative sciences. For example, they have become a ubiquitous means of carrying out approximate Bayesian inference (Doucet et al., 2001; Gilks et al., 1995). The convergence of MC estimation has been considered extensively in the literature (Durrett, 2010). However, the implications arising from the *nesting* of MC schemes, where terms in the integrand depend on the result of separate, nested, MC estimators, is generally less well known. This paper examines the convergence of such nested Monte Carlo (NMC) methods.

Nested expectations occur in wide variety of problems from portfolio risk management (Gordy & Juneja, 2010) to stochastic control (Belomestny et al., 2010). In particular, simulations of agents that reason about other agents often include nested expectations. Tackling such problems requires some form of nested estimation scheme like NMC.

A common class of nested expectations is doubly-intractable inference problems (Murray et al., 2006; Liang, 2010), where the likelihood is only known up to a parameter-dependent normalizing constant. Some problems are even multiply-intractable, such that they require multiple levels of nesting to encode (Stuhlmüller & Goodman, 2014). This can occur, for example, when nesting probabilistic programs (Mantadelis & Janssens, 2011; Le et al., 2016). Our results can be used to show that changes are required to the approaches currently employed by probabilistic programming systems to ensure consistent estimation for such problems (Rainforth, 2017; 2018).

The expected information gain used in Bayesian experimental design (Chaloner & Verdinelli, 1995) requires the calculation of an entropy of a marginal distribution and therefore the expectation of the logarithm of an expectation. By extension, any Kullback-Leibler divergence where one of the terms is a marginal distribution also involves a nested expectation. Hence, our results have important implications for relaxing mean-field assumptions, or using different bounds, in variational inference (Hoffman & Blei, 2015; Naesseth et al., 2017; Maddison et al., 2017) and deep generative models (Burda et al., 2015; Le et al., 2018).

Certain nested estimation problems can be tackled by pseudo-marginal methods (Beaumont, 2003; Andrieu & Roberts, 2009; Andrieu et al., 2010). These consider inference problems where the likelihood is intractable, but can be estimated unbiasedly. From a theoretical perspective, they reformulate the problem in an extended space with auxiliary variables that are used to represent the stochasticity in the likelihood computation, enabling the problem to be expressed as a single expectation.

Our work goes beyond this by considering cases in which a non-linear mapping is applied to the output of the inner expectation, (e.g. the logarithm in the experimental design example), prohibiting such reformulation. We demonstrate that the construction of consistent NMC algorithms is possible, establish convergence rates, and provide empirical evi-

[1] Department of Statistics, University of Oxford [2] Department of Engineering, University of Oxford [3] School of Computing, KAIST [4] Department of Computer Science, University of British Columbia. Correspondence to: Tom Rainforth <rainforth@stats.ox.ac.uk>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018. Copyright 2018 by the author(s).

dence that these rates are observed in practice. Our results show that whenever an outer estimator depends non-linearly on an inner estimator, then the number of samples used in *both* the inner and outer estimators must, in general, be driven to infinity for convergence. We extend our results to cases of repeated nesting and show that the optimal NMC convergence rate is $O(1/T^{\frac{2}{D+2}})$ where T is the total number of samples used in the estimator and D is the nesting depth (with $D = 0$ being conventional MC), whereas naïve approaches only achieve a rate of $O(1/T^{\frac{1}{D+1}})$. We further lay out methods for reformulating certain classes of nested expectation problems into a single expectation, allowing usage of conventional MC estimation schemes with superior convergence rates than naïve NMC. Finally, we use our results to make application-specific advancements in Bayesian experimental design and variational auto-encoders.

1.1 Related Work

Though the convergence of NMC has previously received little attention within the machine learning literature, a number of special cases having been investigated in other fields, sometimes under the name of *nested simulation* (Longstaff & Schwartz, 2001; Belomestny et al., 2010; Gordy & Juneja, 2010; Broadie et al., 2011). While most of this literature focuses on particular application-specific non-linear mappings, a convergence bound for a wider range of problems was shown by Hong & Juneja (2009) and recently revisited in the context of rare-event problems by Fort et al. (2017). The latter paper further considers the case where samples in the outer estimator originate from a Markov chain. Compared to this previous work, ours is the first to consider multiple levels of nesting, applies to a wider range of non-linear mappings, and provides more precise convergence rates. By introducing new results, outlining special cases, providing empirical assessment, and examining specific applications, we provide a unified investigation and practical guide nesting MC estimators in a machine learning context. We begin to realize the potential significance of this by using our theoretical results to make advancements in a number of specific application areas.

Another body of literature related to our work is in the study of the convergence of Markov chains with approximate transition kernels (Rudolf & Schweizer, 2015; Alquier et al., 2016; Medina-Aguayo et al., 2016). The analysis in this work is distinct, but complementary, to our own, focusing on the impact of a known bias on an MCMC chain, whereas our focus is more on the quantifying this bias. Also related is the study of techniques for variance reduction, such as multilevel MC (Heinrich, 2001; Giles, 2008), and bias reduction, such as the multi-step Richardson-Romberg method (Pages, 2007; Lemaire et al., 2017) and Russian roulette sampling (Lyne et al., 2015), many of which are applicable in a NMC context and can improve performance.

2 Problem Formulation

The key idea of MC is that the expectation of an arbitrary function $\lambda \colon \mathcal{Y} \to \mathcal{F} \subseteq \mathbb{R}$ under a probability distribution $p(y)$ for its input $y \in \mathcal{Y}$ can be approximated using:

$$I = \mathbb{E}_{y \sim p(y)} \left[\lambda(y) \right] \tag{1}$$

$$\approx \frac{1}{N} \sum_{n=1}^{N} \lambda(y_n) \quad \text{where} \quad y_n \overset{i.i.d.}{\sim} p(y). \tag{2}$$

In this paper, we consider the case that λ is itself intractable, defined only in terms of a functional mapping of an expectation. Specifically, $\lambda(y) = f(y, \gamma(y))$ where we can evaluate $f \colon \mathcal{Y} \times \Phi \to \mathcal{F}$ exactly for a given y and $\gamma(y)$, but $\gamma(y)$ is the output of the following intractable expectation of another variable $z \in \mathcal{Z}$:

$$\text{either} \quad \gamma(y) = \mathbb{E}_{z \sim p(z|y)} \left[\phi(y, z) \right] \tag{3a}$$

$$\text{or} \quad \gamma(y) = \mathbb{E}_{z \sim p(z)} \left[\phi(y, z) \right] \tag{3b}$$

depending on the problem, with $\phi \colon \mathcal{Y} \times \mathcal{Z} \to \Phi \subseteq \mathbb{R}$. All our results apply to both cases, but we will focus on (3a) for clarity. Estimating I involves computing an integral over z for each value of y in the outer integral. We refer to the approach of tackling both integrations using MC as *nested Monte Carlo* (NMC):

$$I = \mathbb{E}\left[f(y, \gamma(y)) \right] \approx I_{N,M} = \frac{1}{N} \sum_{n=1}^{N} f(y_n, (\hat{\gamma}_M)_n) \tag{4a}$$

where $y_n \overset{i.i.d.}{\sim} p(y)$ and

$$(\hat{\gamma}_M)_n = \frac{1}{M} \sum_{m=1}^{M} \phi(y_n, z_{n,m}) \tag{4b}$$

where each $z_{n,m} \sim p(z|y_n)$ are independently sampled. In Section 3 we will build on this further by considering cases with multiple levels of nesting, where calculating $\phi(y, z)$ involves computation of an intractable (nested) expectation.

3 Convergence of Nested Monte Carlo

We now show that approximating $I \approx I_{N,M}$ is in principle possible, at least when f is well-behaved. In particular, we establish a convergence rate of the mean squared error of $I_{N,M}$ and prove a form of almost sure convergence to I. We further generalize our convergence rate to apply to the case of multiple levels of estimator nesting.

Before providing a formal examination of the convergence of NMC, we first provide intuition about how we might expect to construct a convergent

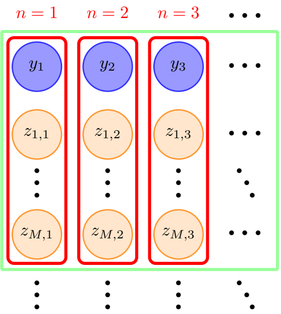


Figure 1. Informal convergence representation

NMC estimator. Consider the diagram shown in Figure 1, and suppose that we want our error to be less than some arbitrary ε. Assume that f is sufficiently smooth that we can choose M large enough to make $|I - \mathbb{E}[f(y_n, (\hat{\gamma}_M)_n)]| < \varepsilon$ (we will characterize the exact requirements for this later). For this fixed M, we have a standard MC estimator on an extended space $y, z_1, \dots, z_M$ such that each sample constitutes one of the red boxes. As we take $N \to \infty$, i.e. taking all the samples in the green box, this estimator converges such that $I_{N,M} \to \mathbb{E}[f(y_n, (\hat{\gamma}_M)_n)]$ as $N \to \infty$ for fixed M. As we can make ε arbitrarily small, we can also achieve an arbitrarily small error.

More formally, convergence bounds for NMC have previously been shown by Hong & Juneja (2009). Under the assumptions that each $(\hat{\gamma}_M)_n$ is Gaussian distributed (which is often reasonable due to the central limit theorem) and that f is thrice differentiable other than at some finite number of points, they show that it is possible to achieve a converge rate of $O(1/N + 1/M^2)$. We now show that these assumptions can be relaxed to only requiring f to be Lipschitz continuous, at the expense of weakening the bound.

Theorem 1. *If f is Lipschitz continuous and $f(y_n, \gamma(y_n)), \phi(y_n, z_{n,m}) \in L^2$, the mean squared error of $I_{N,M}$ converges to 0 at rate $O\left(1/N + 1/M\right)$.*

Proof. The theorem follows as a special case of Theorem 3. For exposition, a more accessible proof for this particular result is also provided in Appendix A in the supplement. □

Inspection of the convergence rate above shows that, given a total number of samples $T = MN$, our bound is tightest when $N \propto M$, with a corresponding rate $O(1/\sqrt{T})$ (see Appendix G). When the additional assumptions of Hong & Juneja (2009) apply, this rate can be lowered to $O(1/T^{2/3})$ by setting $N \propto M^2$. We will later show that this faster convergence rate can, in fact, be achieved whenever f is continuously differentiable, see also (Fort et al., 2017).

These convergence rates suggest that, for most f, it is necessary to increase not only the total number of samples, T, but also the number of samples used for each evaluation of the inner estimator, M, to achieve convergence. Further, as we show in Appendix B, the estimates produced by NMC are, in general, biased. This is perhaps easiest to see by noting that as $N \to \infty$, the variance of the estimator must tend to zero by the law of large numbers, but our bounds remain non-zero for any finite M, implying a bias.

3.1 Minimum Continuity Requirements

We next consider the question of what is the minimal requirement on f to ensures some form of convergence? For a given y_1, we have that $(\hat{\gamma}_M)_1 = \frac{1}{M}\sum_{m=1}^{M} \phi(y_1, z_{1,m}) \to \gamma(y_1)$ almost surely as $M \to \infty$, because the left-hand side is a MC estimator. If f is continuous around y_1, this also implies $f(y_1, (\hat{\gamma}_M)_1) \to f(y_1, \gamma(y_1))$. Our candidate requirement is that this holds in expectation, i.e. that it holds when we incorporate the effect of the outer estimator. More precisely, we define $(\epsilon_M)_n = |f(y_n, (\hat{\gamma}_M)_n) - f(y_n, \gamma(y_n))|$ and require that $\mathbb{E}[(\epsilon_M)_1] \to 0$ as $M \to \infty$ (noting that $(\epsilon_M)_n$ are i.i.d. and so $\mathbb{E}[(\epsilon_M)_1] = \mathbb{E}[(\epsilon_M)_n], \forall n \in \mathbb{N}$). Informally, this "expected continuity" requirement is weaker than uniform continuity (and much weaker than Lipschitz continuity) as it allows (potentially infinitely many) discontinuities in f. More formally we have the following result.

Theorem 2. *For $n \in \mathbb{N}$, let*

$$(\epsilon_M)_n = |f(y_n, (\hat{\gamma}_M)_n) - f(y_n, \gamma(y_n))|.$$

Assume that $\mathbb{E}[(\epsilon_M)_1] \to 0$ as $M \to \infty$. Let Ω be the sample space of our underlying probability space, so that $I_{\tau_\delta(M),M}$ forms a mapping from Ω to $\mathbb{R}$. Then, for every $\delta > 0$, there exists a measurable $A_\delta \subseteq \Omega$ with $\mathbb{P}(A_\delta) < \delta$, and a function $\tau_\delta : \mathbb{N} \to \mathbb{N}$ such that, for all $\omega \notin A_\delta$,

$$I_{\tau_\delta(M),M}(\omega) \overset{a.s.}{\to} I \quad \text{as} \quad M \to \infty.$$

Proof. See Appendix C. □

As well as providing proof of a different form of convergence to any existing results, this result is particularly important because many, if not most, functions are not Lipschitz continuous due to their behavior in the limits. For example, even the function $f(y, \gamma(y)) = (\gamma(y))^2$ is not Lipschitz continuous because the derivative is unbounded as $|\gamma(y)| \to \infty$, whereas the vast majority of problems will satisfy our weaker requirement of $\mathbb{E}[(\epsilon_M)_1] \to 0$.

3.2 Repeated Nesting and Exact Bounds

We next consider the case of multiple levels of nesting. This case is particularly important for analyzing probabilistic programming languages. To formalize what we mean by arbitrary nesting, we first assume some fixed integral depth $D > 0$, and real-valued functions $f_0, \cdots, f_D$. We then define

$$\gamma_D\left(y^{(0:D-1)}\right) = \mathbb{E}\left[f_D\left(y^{(0:D)}\right) \middle| y^{(0:D-1)}\right] \quad \text{and}$$

$$\gamma_k(y^{(0:k-1)}) = \mathbb{E}\left[f_k\left(y^{(0:k)}, \gamma_{k+1}\left(y^{(0:k)}\right)\right) \middle| y^{(0:k-1)}\right],$$

for $0 \leq k < D$, where $y^{(k)} \sim p\left(y^{(k)} | y^{(0:k-1)}\right)$. Note that our single nested case corresponds to the setting of $D = 1$, $f_0 = f$, $f_1 = \phi$, $y^{(0)} = y$, $y^{(1)} = z$, $\gamma_0 = I$, and $\gamma_1 = \gamma$. Our goal is to estimate $\gamma_0 = \mathbb{E}\left[f_0\left(y^{(0)}, \gamma_1\left(y^{(0)}\right)\right)\right]$. To do so we will use the following NMC scheme:

$$I_D\left(y^{(0:D-1)}\right) = \frac{1}{N_D}\sum_{n=1}^{N_D} f_D\left(y^{(0:D-1)}, y_n^{(D)}\right) \quad \text{and}$$

$$I_k\left(y^{(0:k-1)}\right)$$

$$= \frac{1}{N_k}\sum_{n=1}^{N_k} f_k\left(y^{(0:k-1)}, y_n^{(k)}, I_{k+1}\left(y^{(0:k-1)}, y_n^{(k)}\right)\right)$$

for $0 \leq k \leq D-1$, where each $y_n^{(k)} \sim p\left(y^{(k)}|y^{(0:k-1)}\right)$ is drawn independently. Note that there are multiple values of $y_n^{(k)}$ for each possible $y^{(0:k-1)}$ and that $I_k\left(y^{(0:k-1)}\right)$ is still a random variable given $y^{(0:k-1)}$.

We are now ready to provide our general result for the convergence bounds that applies to cases of repeated nesting, provides constant factors (rather than just using big O notation), and shows how the bound can be improved if the additional assumption of continuous differentiability holds.

Theorem 3. *If $f_0, \cdots, f_D$ are all Lipschitz continuous in their second input with Lipschitz constants*

$$K_k := \sup_{y^{(0:k)}} \left| \frac{\partial f_k\left(y^{(0:k)}, \gamma_{k+1}(y^{(0:k)})\right)}{\partial \gamma_{k+1}} \right|,$$

for all $k \in 0, \dots, D-1$ and if

$$\varsigma_k^2 := \mathbb{E}\left[\left(f_k\left(y^{(0:k)}, \gamma_{k+1}\left(y^{(0:k)}\right)\right) - \gamma_k\left(y^{(0:k-1)}\right)\right)^2\right] < \infty \quad \forall k \in 0, \dots, D$$

then

$$\mathbb{E}\left[\left(I_0 - \gamma_0\right)^2\right] \leq \frac{\varsigma_0^2}{N_0} + \sum_{k=1}^{D}\left(\prod_{\ell=0}^{k-1} K_\ell^2\right)\frac{\varsigma_k^2}{N_k} + O(\epsilon) \quad (5)$$

where $O(\epsilon)$ represents asymptotically dominated terms.

If $f_0, \cdots, f_D$ are also continuously differentiable with second derivative bounds

$$C_k := \sup_{y^{(0:k)}} \left| \frac{\partial^2 f_k\left(y^{(0:k)}, \gamma_{k+1}(y^{(0:k)})\right)}{\partial \gamma_{k+1}^2} \right|$$

then this mean square error bound can be tightened to

$$\mathbb{E}\left[\left(I_0 - \gamma_0\right)^2\right] \leq \frac{\varsigma_0^2}{N_0} + \left(\frac{C_0 \varsigma_1^2}{2N_1} + \sum_{k=0}^{D-2}\left(\prod_{d=0}^{k} K_d\right)\frac{C_{k+1}\varsigma_{k+2}^2}{2N_{k+2}}\right)^2 + O(\epsilon). \quad (6)$$

For a single nesting, we can further characterize $O(\epsilon)$ giving

$$\mathbb{E}\left[\left(I_0 - \gamma_0\right)^2\right] \leq \frac{\varsigma_0^2}{N_0} + \frac{4K_0^2\varsigma_1^2}{N_0 N_1} + \frac{2K_0\varsigma_0\varsigma_1}{N_0\sqrt{N_1}} + \frac{K_0^2\varsigma_1^2}{N_1} \quad (7)$$

$$\mathbb{E}\left[\left(I_0 - \gamma_0\right)^2\right] \leq \frac{\varsigma_0^2}{N_0} + \frac{C_0^2\varsigma_1^4}{4N_1^2}\left(1 + \frac{1}{N_0}\right) + \frac{K_0^2\varsigma_1^2}{N_0 N_1} + \frac{2K_0\varsigma_1}{N_0\sqrt{N_1}}\sqrt{\varsigma_0^2 + \frac{C_0^2\varsigma_1^4}{4N_1^2}} + O\left(\frac{1}{N_1^3}\right) \quad (8)$$

for when the continuous differentiability assumption does not hold and holds respectively.

Proof. See Appendix D. □

These results give a convergence rate of $O(\sum_{k=0}^{D} 1/N_k)$ when only Lipschitz continuity holds and $O(1/N_0 + (\sum_{k=1}^{D} 1/N_k)^2)$ when all the f_k are also continuously differentiable. As estimation requires drawing $O(T)$ samples where $T = \prod_{k=0}^{D} N_k$, the convergence rate will rapidly diminish with repeated nesting. More precisely, as shown in Appendix G, the optimal convergence rates are $O(1/T^{\frac{1}{D+1}})$ and $O(1/T^{\frac{2}{D+2}})$ respectively for the two cases, both of which imply that the rate diminishes exponentially with D.

4 Special Cases

We now outline some special cases where it is possible to achieve a convergence rate of $O(1/N)$ in the mean square error (MSE) as per conventional MC estimation. Establishing these cases is important because it identifies for which problems we can use conventional results, when we can achieve an improved convergence rate, and what precautions we must take to ensure this. We will focus on single nesting instances, but note that all results still apply to repeated nesting scenarios because they can be used to "collapse" layers and thereby reduce the depth of the nesting.

4.1 Linear f

Our first special case is that f is linear in its second argument, i.e. $f(y, \alpha v + \beta w) = \alpha f(y, v) + \beta f(y, w)$. Here the problem can be rearranged to a single expectation, a well-known result which forms the basis for pseudo-marginal, nested sequential MC (Naesseth et al., 2015), and certain ABC methods (Csilléry et al., 2010). Namely we have

$$\begin{aligned} I &= \mathbb{E}_{y \sim p(y)}\left[f\left(y, \mathbb{E}_{z \sim p(z|y)}\left[\phi(y,z)\right]\right)\right] \\ &= \mathbb{E}_{y \sim p(y)}\left[\mathbb{E}_{z \sim p(z|y)}\left[f(y, \phi(y,z))\right]\right] \\ &\approx \frac{1}{N}\sum_{n=1}^{N} f(y_n, \phi(y_n, z_n)) \end{aligned} \quad (9)$$

where $(y_n, z_n) \sim p(y)p(z|y)$ if $\gamma(y)$ is of the form of (3a) and $y_n \sim p(y)$ and $z_n \sim p(z)$ are independently drawn if $\gamma(y)$ is of the form of (3b).

4.2 Finite Possible Realizations of y

Our second case is if y must take one of finitely many values $y_1, \cdots, y_C$, then it is possible to use another approach to ensure the same convergence rate as standard MC. The key observation is to note that in this case we can convert the nested problem (2) into C separate non-nested problems

$$I = \sum_{c=1}^{C} P(y = y_c)\, f(y_c, \gamma(y_c)) \quad (10)$$

which can then be estimated using

$$I_N = \sum_{c=1}^{C} (\hat{P}_N)_c\, (\hat{f}_N)_c \quad \text{where} \quad (11)$$

$$P(y = y_c) \approx (\hat{P}_N)_c = \frac{1}{N}\sum_{n=1}^{N} \mathbb{I}(y_n = y_c) \quad (12)$$

$$f(y_c, \gamma(y_c)) \approx (\hat{f}_N)_c = f\left(y_c, \frac{1}{N}\sum_{n=1}^{N} \phi(y_c, z_{n,c})\right) \quad (13)$$

with $y_n \overset{i.i.d.}{\sim} p(y)$ and $z_{n,c} \sim p(z|y_c)$ (or $z_{n,c} \sim p(z)$ if using the formulation in (3b)). Note the critical point that each $z_{n,c}$ is independent of y_n as each y_c is a constant. We can now show the following result which, though intuitively straightforward, requires care to formally prove.

Theorem 4. *If f is Lipschitz continuous, then the mean squared error of $I_N = \sum_{c=1}^{C} (\hat{P}_N)_c \, (\hat{f}_N)_c$ as an estimator for I as per* (10) *converges at rate $O(1/N)$.*

Proof. See Appendix E. □

4.3 Products of Expectations

We next consider the scenario, occurring for many latent variables models and probabilistic programming problems, where $\gamma(y)$ is equal to the product of multiple expectations, rather than just a single expectation as per (3a). That is,

$$I = \mathbb{E}_{y \sim p(y)} \left[f \left(y, \prod_{\ell=1}^{L} \mathbb{E}_{z_\ell \sim p(z_\ell|y)} \left[\psi_\ell(y, z_\ell) \right] \right) \right]. \quad (14)$$

Because the z_ℓ will not in general be independent, we cannot trivially rearrange (14) to a standard nested estimation by moving the product within the expectation. Our insight is that the required rearrangement can instead be achieved by introducing new random variables $\{z'_\ell\}_{\ell=1:L}$ such that each $z'_\ell|y \sim p(z_\ell|y)$ and the z'_ℓ are independent of one another. This can be achieved by, for example, taking L independent samples from the joint $Z_\ell \overset{i.i.d.}{\sim} p(z_{1:L}|y)$ and using the ℓ^{th} such draw for the ℓ^{th} dimension of z', i.e. setting $z'_\ell = \{Z_\ell\}_\ell$. For every $y \in \mathcal{Y}$ we now have

$$\prod_{\ell=1}^{L} \mathbb{E}_{z_\ell \sim p(z_\ell|y)}[\psi_\ell(y, z_\ell)] = \prod_{\ell=1}^{L} \mathbb{E}_{z'_\ell \sim p(z'_\ell|y)}[\psi_\ell(y, z'_\ell)]$$
$$= \mathbb{E}_{\{z'_\ell\}_{\ell=1:L} \sim p(\{z'_\ell\}_{\ell=1:L}|y)} \left[\prod_{\ell=1}^{L} \psi_\ell(y, z'_\ell) \right] \quad (15)$$

which is a single expectation on an extended space and shows that (14) fits the NMC formulation. Furthermore, we can now show that if f is linear, the MSE of the NMC estimator (14) converges at the standard MC rate $O(1/N)$, provided that M remains fixed.

Theorem 5. *Consider the NMC estimator*

$$I_N = \frac{1}{N} \sum_{n=1}^{N} f \left(y_n, \prod_{\ell=1}^{L} \frac{1}{M_\ell} \sum_{m=1}^{M_\ell} \psi_\ell(y_n, z'_{n,\ell,m}) \right)$$

where each $y_n \in \mathcal{Y}$ and $z'_{n,\ell,m} \in \mathcal{Z}_\ell$ are independently drawn from $y_n \sim p(y)$ and $z'_{n,\ell,m}|y_n \sim p(z_\ell|y_n)$, respectively. If f is linear, the estimator converges almost surely to I, with a convergence rate of $O(1/N)$ in the mean square error for any fixed choice of $\{M_\ell\}_{\ell=1:L}$.

Proof. See Appendix F. □

As this result holds in the case $L = 1$, an important consequence is that whenever f is linear, the same convergence rate is achieved regardless of whether we reformulate the problem to a single expectation or not, provided that the number of samples used by the inner estimator is fixed.

4.4 Polynomial f

Perhaps surprisingly, whenever f is of the form

$$f(y, \gamma(y)) = g(y)\, \gamma(y)^\alpha \quad (16)$$

where $\alpha \in \mathbb{Z}_{\geq 0}$, then it is also possible to construct a standard MC estimator by building on the ideas introduced in Section 4.3 and those of (Goda, 2016). The key idea is

$$\left(\mathbb{E}\left[z\right]\right)^2 = \mathbb{E}\left[z\right] \mathbb{E}\left[z'\right] = \mathbb{E}\left[zz'\right] \quad (17)$$

where z and z' are i.i.d. Therefore, assuming appropriate integrability requirements, we can construct the following non-nested MC estimator:

$$\mathbb{E}\left[g(y)\, \gamma(y)^\alpha\right] = \mathbb{E}\left[g(y) \prod_{\ell=1}^{\alpha} \mathbb{E}_{z_\ell \sim p(z|y)} \left[\phi(y, z_\ell)|y\right] \right]$$
$$= \mathbb{E}\left[g(y) \prod_{\ell=1}^{\alpha} \phi(y, z_\ell) \right] \approx \frac{1}{N} \sum_{n=1}^{N} g(y_n) \prod_{\ell=1}^{\alpha} \phi(y_n, z_{n,\ell})$$

where we independently draw each $z_{n,\ell}|y_n \sim p(z|y_n)$.

5 Empirical Verification

The convergence rates proven in Section 3 are only *upper bounds* on the worst-case performance. We will now examine whether these convergence rates are tight in practice, investigate what happens when our guidelines are not followed, and outline some applications of our results.

5.1 Simple Analytic Model

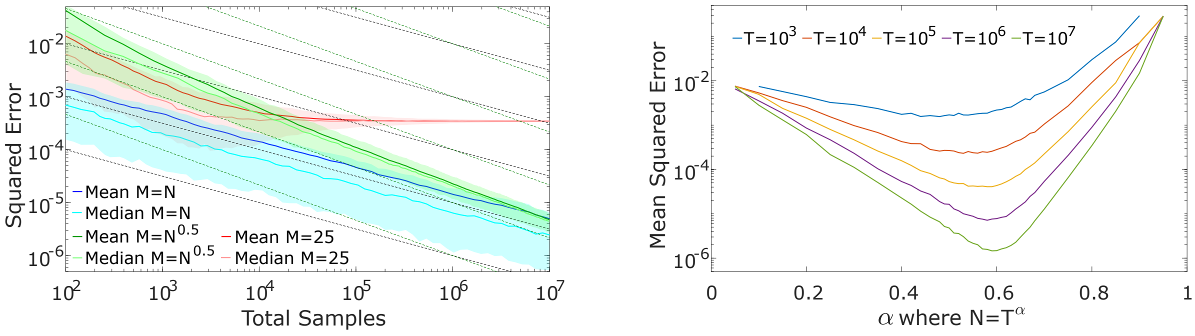
We start with the following analytically calculable problem

$$y \sim \text{Uniform}(-1, 1), \quad (18a)$$
$$z \sim \mathcal{N}(0, 1), \quad (18b)$$
$$\phi(y, z) = \sqrt{2/\pi} \exp\left(-2(y - z)^2\right), \quad (18c)$$
$$f(y, \gamma(y)) = \log(\gamma(y)) = \log(\mathbb{E}_z[\phi(y, z)]). \quad (18d)$$

for which $I = \frac{1}{2} \log\left(\frac{2}{5\pi}\right) - \frac{2}{15}$. Figure 2a shows the corresponding empirical convergence obtained by applying (4) to (18) directly. It shows that, for this problem, the theoretical convergence rates from Theorem 3 are indeed realized. The figure also demonstrates the danger of not increasing M with N, showing that the NMC estimator converges to an incorrect solution when M is held constant. Figure 2b shows the effect of varying N and M for various fixed sample budgets T and demonstrates that the asymptotically optimal strategy can be suboptimal for finite budgets.

5.2 Planning Cancer Treatment

We now introduce a real-world example to show the applicability of NMC in a scenario where the solution is not analytically tractable and conventional MC is insufficient. Consider a treatment center assessing a new policy for plan-



(a) Convergence with increasing sample budget.

(b) Final error for different T and N.

Figure 2. Empirical convergence of NMC for (18). [Left] convergence in total samples for different ways of setting M and N. Results are averaged over 1000 independent runs, while shaded regions give the 25%-75% quantiles. The theoretical convergence rates, namely $O(1/\sqrt{T})$ and $O(1/T^{2/3})$ for setting $N \propto M$ and $N \propto M^2$ respectively, are observed (see the dashed black and green lines respectively for reference). The fixed M case converges at the standard MC error rate of $O(1/T)$ but to a biased solution. [Right] final error for different total sample budgets as a function of α where $N = T^\alpha$ and $M = T^{1-\alpha}$ iterations are used for the outer and inner estimators respectively. This shows that even though $\alpha = \frac{2}{3}$ is the asymptotically optimal allocation strategy, this is not the optimal solution for finite T. Nonetheless, as T increases, the optimum value of α increases, starting around 0.5 for $T = 10^3$ and reaching around 0.6 for $T = 10^7$.

ning cancer treatments, subject to a budget. Clinicians must decide on a patient-by-patient basis whether to administer chemotherapy in the hope that their tumor will reduce in size sufficiently to be able to perform surgery at a later date. A treatment is considered to have been successful if the size of the tumor drops below a threshold value in a fixed time window. The clinicians have at their disposal a simulator for the evolution of tumors with time, parameterized by both observable values, y, such as tumor size, and unobservable values, z, such as the patient-specific response to treatment. Given a set of input parameters, the simulator deterministically returns a binary response $\phi(y, z) \in \{0, 1\}$, with 1 indicating a successful treatment. To estimate the probability of a successful treatment for a given patient, the clinician must calculate the expected success over these unobserved variables, namely $\mathbb{E}_{z\sim p(z|y)}[\phi(y, z)]$ where $p(z|y)$ represents a probabilistic model for the unobserved variables, which could, for example, be constructed based on empirical data. The clinician then decides whether to go ahead with the treatment for that patient based on whether the calculated probability of success exceeds a certain threshold T_{treat}.

The treatment center wishes to estimate the expected number of patients that will be treated for a given T_{treat} so that it can minimize this threshold without exceeding its budget. To do this, it calculates the expectation of the clinician's decisions to administer treatment, giving the complete nested expectation for calculating the number of treated patients as

$$I(T_{\text{treat}}) = \mathbb{E}\left[\mathbb{I}\left(\mathbb{E}_{z\sim p(z|y)}[\phi(y, z)] > T_{\text{treat}}\right)\right], \quad (19)$$

where the step function $\mathbb{I}(\cdot > T_{\text{treat}})$ imposes a non-linear mapping, preventing conventional MC estimation. Full details on ϕ, $p(y)$, and $p(z|y)$ are given in Appendix H.

To verify the convergence rate, we repeated the analysis from Section 5.1 for (19) at a fixed value of $T_{\text{treat}} = 0.35$. The results, shown in Figure 3, again verify the theoretical rates. By further testing different values of T_{treat}, we found $T_{\text{treat}} = 0.125$ to be optimal under the budget.

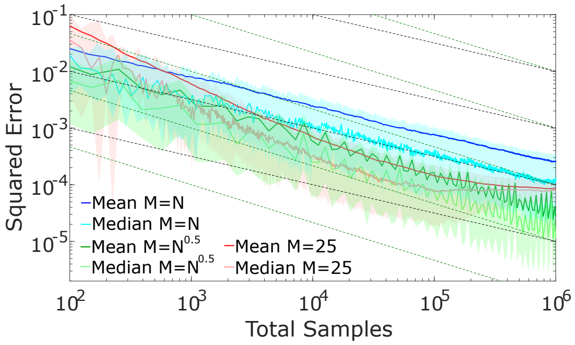


Figure 3. Convergence of NMC for cancer simulation. A ground truth estimate was calculated using a single run with $M = 10^5$ and $N = 10^5$. Experimental setup and conventions are as per Figure 2a and we again observe the expected convergence rates. When $M = \sqrt{N}$ an interesting fluctuation behavior is observed. Further testing suggests that this originates because the bias of the estimator depends in a fluctuating manner on the value of M as the binary output of $\phi(y, z)$ creates a quantization effect on the possible estimates for $\hat{\gamma}$. This effect is also observed for the $M = N$ case but is less pronounced.

5.3 Repeated Nesting

We next consider some simple models with multiple levels of nesting, starting with

$$y^{(0)} \sim \text{Uniform}(0, 1), \;\; y^{(1)} \sim \mathcal{N}(0, 1), \;\; y^{(2)} \sim \mathcal{N}(0, 1),$$

$$f_0\left(y^{(0)}, \gamma_1\left(y^{(0)}\right)\right) = \log \gamma_1\left(y^{(0)}\right) \quad (20a)$$

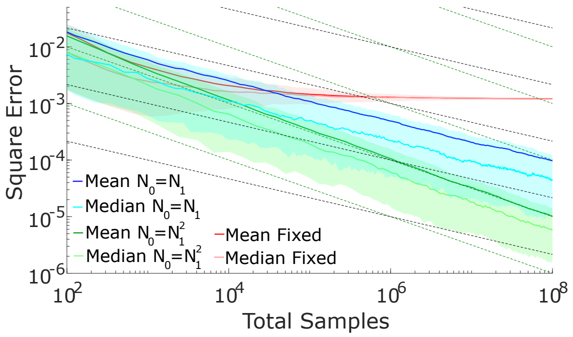


Figure 4. Empirical convergence of NMC to (20) for an increasing total sample budget $T = N_0N_1N_2$. Setup and conventions as per Figure 2a. Shown in red is the convergence with a fixed $N_2 = 5$ and $N_0 = N_1^2$, which we see gives a biased solution. Shown in blue is the convergence when setting $N_0 = N_1 = N_2$, which we see converges at the expected $O(T^{-1/3})$ rate. Shown in green is the convergence when setting $N_0 = N_1^2 = N_2^2$ which we see again gives the theoretical convergence rate, namely $O(T^{-1/2})$.

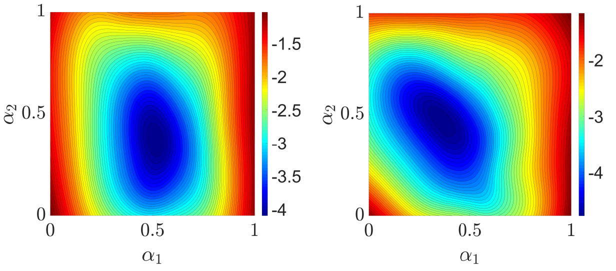
$$f_1\left(y^{(0:1)}, \gamma_2\left(y^{(0:1)}\right)\right) = \exp\left(-\frac{1}{2}\left(y^{(0)} - y^{(1)} - \log\gamma_2\left(y^{(0:1)}\right)\right)\right) \tag{20b}$$

$$f_2\left(y^{(0:2)}\right) = \exp\left(y^{(2)} - \frac{y^{(0)} + y^{(1)}}{2}\right) \tag{20c}$$

which has analytic solution $I = -3/32$. The convergence plot shown in Figure 4 demonstrates that the theoretically expected convergence behaviors are observed for different methods of setting N_0, N_1, and N_2.

We further investigated the empirical performance of different strategies for choosing N_0, N_1, N_2 under a finite fixed budget $T = N_0N_1N_2$. In particular, we looked to establish the optimal empirical setting under the fixed budget $T = 10^6$ for the model described in (20) and a slight variation where $y^{(0)}$ is replaced with $y^{(0)}/10$, for which the ground truth is now $I = 39/160$. Defining α_1 and α_2 such that $N_0 = T^{\alpha_1}$, $N_1 = T^{\alpha_2(1-\alpha_1)}$, and $N_2 = T^{(1-\alpha_1)(1-\alpha_2)}$, we ran a Bayesian optimization algorithm, namely BOPP (Rainforth et al., 2016), to optimize the log MSE, $\log_{10}\left(\mathbb{E}\left[(I_0(\alpha_1, \alpha_2) - \gamma_0)^2\right]\right)$, with respect to (α_1, α_2). For each tested (α_1, α_2), the MSE was estimated using 1000 independently generated samples of I_0 and we allowed a total of 200 such tests. We found respective optimal values for (α_1, α_2) of $(0.53, 0.36)$ and $(0.38, 0.45)$. By comparison, the asymptotically optimal setup suggested by our theoretical results is $(0.5, 0.5)$, showing that the finite budget optimal allocation can vary significantly from the asymptotically optimal solution and that it does so in a problem dependent manner.

As a byproduct, BOPP also produced Gaussian process approximations to the log MSE variations, as shown in Figure 5. We see that the two problems lead to distinct performance variations. Based on the (unshown) uncertainty estimates of these Gaussian processes, we believe these approximations are a close representation of the truth.



(a) Original

(b) Modification

Figure 5. Contour plots of $\log_{10}\left(\mathbb{E}\left[(I_0 - \gamma_0)^2\right]\right)$ produced by BOPP for different allocations of the sample budget $T = 10^6$ for the problem shown in (20) and its modified variant.

6 Applications

6.1 Bayesian Experimental Design

In this section, we show how our results can be used to derive an improved estimator for the problem of Bayesian experimental design (BED) in the case where the experiment outputs are discrete. A summary of our approach is provided here, with full details provided in Appendix I.

Bayesian experimental design provides a framework for designing experiments in a manner that is optimal from an information-theoretic viewpoint (Chaloner & Verdinelli, 1995; Sebastiani & Wynn, 2000). Given a prior $p(\theta)$ on parameters θ and a corresponding likelihood $p(y|\theta, d)$ for experiment outcomes y given a design d, the Bayesian optimal design d^* is given by maximizing the mutual information between θ and y defined as follows

$$\bar{U}(d) = \int_{\mathcal{Y}}\int_{\Theta} p(y, \theta|d) \log\left(\frac{p(\theta|y, d)}{p(\theta)}\right) d\theta dy. \tag{21}$$

Estimating d^* is challenging as $p(\theta|y, d)$ is rarely known in closed-form. However, appropriate algebraic manipulation shows that (21) is consistently estimated by

$$\hat{U}_{\text{NMC}}(d) = \frac{1}{N}\sum_{n=1}^{N}\Bigg[\log(p(y_n|\theta_{n,0}, d)) - \log\left(\frac{1}{M}\sum_{m=1}^{M} p(y_n|\theta_{n,m}, d)\right)\Bigg] \tag{22}$$

where $\theta_{n,m} \sim p(\theta)$ for each $(m, n) \in \{0, \dots, M\} \times \{1, \dots, N\}$, and $y_n \sim p(y|\theta = \theta_{n,0}, d)$ for each $n \in \{1, \dots, N\}$. This naïve NMC estimator has been implicitly used by (Myung et al., 2013) amongst others and gives a convergence rate of $O(1/N + 1/M^2)$ as per Theorem 3.

When y can only take on finitely many realizations $y_1, \dots, y_c$, we use the ideas introduced in Section 4.2 to

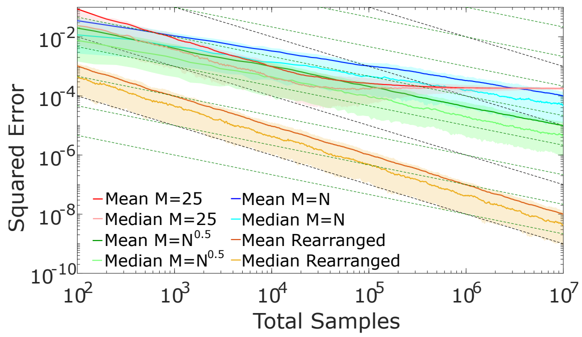


Figure 6. Convergence of NMC (i.e. (22)) and our reformulated estimator (23) for the BED problem. Experimental setup and conventions are as per Figure 2a, with a ground truth estimate made using a single run of the reformulated estimator with 10^{10} samples. We see that the theoretical convergence rates are observed, with the advantages of the reformulated estimator particularly pronounced.

derive the following improved estimator

$$\hat{U}_R(d) = \frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C} p(y_c|\theta_n, d)\log\left(p(y_c|\theta_n, d)\right) \tag{23}$$
$$-\sum_{c=1}^{C}\left[\left(\frac{1}{N}\sum_{n=1}^{N} p(y_c|\theta_n, d)\right)\log\left(\frac{1}{N}\sum_{n=1}^{N} p(y_c|\theta_n, d)\right)\right]$$

where $\theta_n \sim p(\theta), \forall n \in \{1, \dots, N\}$. As C is fixed, (23) converges at the standard MC error rate of $O(1/N)$. This constitutes a substantially faster convergence as (22) requires a total of MN samples compared to N for (23).

We finish by showing that the theoretical advantages of this reformulation also lead to empirical gains. For this we consider a model used in psychology experiments introduced by (Vincent, 2016), details of which are given in Appendix I. Figure 6 demonstrates that the theoretical convergence rates are observed while results given in Appendix I show that this leads to significant practical gains in estimating $\bar{U}(d)$.

6.2 Variational Autoencoders

To give another example of the applicability of our results, we now use Theorem 3 to directly derive a new result for the importance weighted autoencoder (IWAE) (Burda et al., 2015). Both the IWAE and the standard variational autoencoder (VAE) (Kingma & Welling, 2013) use lower bounds on the model evidence as objectives for train deep generative models and employ estimators of the form

$$I_{N,M} = \frac{1}{N}\sum_{n=1}^{N}\log\left(\frac{1}{M}\sum_{m=1}^{M} w_{n,m}(\theta)\right) \tag{24}$$

for some given θ upon which the random $w_{n,m}(\theta)$ depend. The IWAE sets $N = 1$ and the VAE $M = 1$. We can view (24) as a (biased) NMC estimator for the log evidence $\log \mathbb{E}\left[w_{1,1}(\theta)\right]$, which is the target one actually wishes to optimize (for the generative network). We can now assess the MSE of this biased estimator using (8), noting that this is a special case where $\varsigma_0^2 = 0$, giving $\mathbb{E}\left[\left(I_{N,M} - I\right)^2\right] \le \frac{C_0^2\varsigma_1^4}{4M^2}\left(1 + \frac{1}{N}\right) + \frac{K_0^2\varsigma_1^2}{NM} + \frac{C_0K_0\varsigma_1^3}{NM^{3/2}} + O(\frac{1}{M^3})$. For a fixed budget $T = NM$ this becomes $O\left(\frac{1}{M^2} + \frac{1}{T} + \frac{1}{T\sqrt{M}}\right)$. Given T is fixed, we thus see that the higher M is, the lower the error bound. Therefore, the lowest MSE is achieved by setting $N = 1$ and $M = T$, as is done by the IWAE. As we show in Rainforth et al. (2018), these results further carry over to the reparameterized derivative estimates $\nabla_\theta I_{N,M}$.

6.3 Nesting Probabilistic Programs

Probabilistic programming systems (PPSs) (Goodman et al., 2008; Wood et al., 2014) provide a strong motivation for the study of NMC methods because many PPSs allow for arbitrary nesting of models (or queries, as they are known in the PPS literature), such that it is easy to define and run nested inference problems, including cases with multiple layers of nesting (Stuhlmüller & Goodman, 2012; 2014). Though this ability to nest queries has started to be exploited in application-specific work (Ouyang et al., 2016; Le et al., 2016), the resulting nested inference problems fall outside the scope of conventional convergence proofs and so the statistical validity of the underlying inference engines has previously been an open question in the field.

As we show in Rainforth (2017; 2018), the results presented here can be brought to bear on assessing the relative correctness of the different ways PPSs allow model nesting. In particular, the correctness of sampling from the conditional distribution of one query within another follows from Theorem 3, but only if the computation for each call to the inner query increases the more times that query is called. This requirement is not satisfied by current systems. Meanwhile, Theorem 5 can be used to the assert that observing the output of one query inside another leads to convergence at the standard MC rate, provided that the computation of the inner query instead remains fixed.

7 Conclusions

We have introduced a formal framework for NMC estimation and shown that it can be used to yield a consistent estimator for problems that cannot be tackled with conventional MC alone. We have derived convergence rates and considered what minimal continuity assumptions are required for convergence. However, we have also highlighted a number of potential pitfalls for naïve application of NMC and provided guidelines for avoiding these, e.g. highlighting the importance of increasing the number of samples in both the inner and the outer estimators to ensure convergence. We have further introduced techniques for converting certain classes of NMC problems to conventional MC ones, providing improved convergence rates. Our work has implications throughout machine learning and we hope it will provide the foundations for exploring this plethora of applications.

Acknowledgements

Tom Rainforth's research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) ERC grant agreement no. 617071. However, the majority of this work was undertaken while he was in the Department of Engineering Science, University of Oxford, and was supported by a BP industrial grant. Robert Cornish is supported by an NVIDIA scholarship. Hongseok Yang is supported by an Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIP) (No.R0190-16-2011, Development of Vulnerability Discovery Technologies for IoT Software Security). Andrew Warrington is supported by a Keble College scholarship. Frank Wood is supported under DARPA PPAML through the U.S. AFRL under Cooperative Agreement FA8750-14-2-0006, Sub Award number 61160290-111668.

References

Alquier, P., Friel, N., Everitt, R., and Boland, A. Noisy Monte Carlo: Convergence of Markov chains with approximate transition kernels. *Statistics and Computing*, 26(1-2):29–47, 2016.

Andrieu, C. and Roberts, G. O. The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics*, pp. 697–725, 2009.

Andrieu, C., Doucet, A., and Holenstein, R. Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2010.

Beaumont, M. A. Estimation of population growth or decline in genetically monitored populations. *Genetics*, 164 (3):1139–1160, 2003.

Belomestny, D., Kolodko, A., and Schoenmakers, J. Regression methods for stochastic control problems and their convergence analysis. *SIAM Journal on Control and Optimization*, 2010.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. Variational inference: A review for statisticians. *arXiv preprint arXiv:1601.00670*, 2016.

Broadie, M., Du, Y., and Moallemi, C. C. Efficient risk estimation via nested sequential simulation. *Management Science*, 2011.

Burda, Y., Grosse, R., and Salakhutdinov, R. Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*, 2015.

Chaloner, K. and Verdinelli, I. Bayesian experimental design: A review. *Statistical Science*, 1995.

Csilléry, K., Blum, M. G., Gaggiotti, O. E., and François, O. Approximate Bayesian computation (ABC) in practice. *Trends in ecology & evolution*, 25(7):410–418, 2010.

Doucet, A., De Freitas, N., and Gordon, N. An introduction to sequential Monte Carlo methods. In *Sequential Monte Carlo methods in practice*, pp. 3–14. Springer, 2001.

Durrett, R. *Probability: theory and examples*. Cambridge university press, 2010.

Enderling, H. and Chaplain, M. A. Mathematical modeling of tumor growth and treatment. *Current Pharmaceutical Design*, 20(30):4934–4940, 2014. ISSN 1381-6128/1873-4286. doi: 10.2174/13816128196661311251 50434. URL `http://www.eurekaselect.com/node/118301/article|`.

Fort, G., Gobet, E., and Moulines, E. MCMC design-based non-parametric regression for rare-event. application to nested risk computations. *Monte Carlo Methods Appl*, 2017.

Giles, M. B. Multilevel Monte Carlo path simulation. *Operations Research*, 56(3):607–617, 2008.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. *MCMC in practice*. CRC press, 1995.

Goda, T. Computing the variance of a conditional expectation via non-nested Monte Carlo. *Operations Research Letters*, 2016.

Goodman, N., Mansinghka, V., Roy, D. M., Bonawitz, K., and Tenenbaum, J. B. Church: a language for generative models. *UAI*, 2008.

Gordy, M. B. and Juneja, S. Nested simulation in portfolio risk measurement. *Management Science*, 2010.

Heinrich, S. Multilevel Monte Carlo methods. *LSSC*, 1: 58–67, 2001.

Hoffman, M. and Blei, D. Stochastic structured variational inference. In *AISTATS*, 2015.

Hong, L. J. and Juneja, S. Estimating the mean of a nonlinear function of conditional expectation. In *Winter Simulation Conference*, 2009.

Jacob, P. E., Thiery, A. H., et al. On nonnegative unbiased estimators. *The Annals of Statistics*, 43(2):769–784, 2015.

Kingma, D. P. and Welling, M. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Le, T. A., Baydin, A. G., and Wood, F. Nested compiled inference for hierarchical reinforcement learning. In *NIPS Workshop on Bayesian Deep Learning*, 2016.

Le, T. A., Igl, M., Rainforth, T., Jin, T., and Wood, F. Auto-encoding sequential Monte Carlo. In *ICLR*, 2018.

Lemaire, V., Pagès, G., et al. Multilevel Richardson–Romberg extrapolation. *Bernoulli*, 23(4A):2643–2692, 2017.

Liang, F. A double Metropolis–Hastings sampler for spatial models with intractable normalizing constants. *Journal of Statistical Computation and Simulation*, 80(9):1007–1022, 2010.

Longstaff, F. A. and Schwartz, E. S. Valuing American options by simulation: a simple least-squares approach. *Review of Financial studies*, 2001.

Lyne, A.-M., Girolami, M., Atchade, Y., Strathmann, H., Simpson, D., et al. On Russian roulette estimates for Bayesian inference with doubly-intractable likelihoods. *Statistical science*, 30(4):443–467, 2015.

Maddison, C. J., Lawson, D., Tucker, G., Heess, N., Norouzi, M., Mnih, A., Doucet, A., and Teh, Y. W. Filtering variational objectives. *arXiv preprint arXiv:1705.09279*, 2017.

Mantadelis, T. and Janssens, G. Nesting probabilistic inference. *arXiv preprint arXiv:1112.3785*, 2011.

Medina-Aguayo, F. J., Lee, A., and Roberts, G. O. Stability of noisy Metropolis–Hastings. *Statistics and Computing*, 26(6):1187–1211, 2016.

Murray, I., Ghahramani, Z., and MacKay, D. J. MCMC for doubly-intractable distributions. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, pp. 359–366. AUAI Press, 2006.

Myung, J. I., Cavagnaro, D. R., and Pitt, M. A. A tutorial on adaptive design optimization. *Journal of mathematical psychology*, 57(3):53–67, 2013.

Naesseth, C. A., Lindsten, F., and Schön, T. Nested sequential Monte Carlo methods. In *ICML*, 2015.

Naesseth, C. A., Linderman, S. W., Ranganath, R., and Blei, D. M. Variational sequential Monte Carlo. *arXiv preprint arXiv:1705.11140*, 2017.

O'Hagan, A. Bayes–Hermite quadrature. *Journal of statistical planning and inference*, 1991.

Ouyang, L., Tessler, M. H., Ly, D., and Goodman, N. Practical optimal experiment design with probabilistic programs. *arXiv preprint arXiv:1608.05046*, 2016.

Pages, G. Multi-step Richardson–Romberg extrapolation: remarks on variance control and complexity. *Monte Carlo Methods and Applications*, 13(1):37, 2007.

Rainforth, T. *Automating Inference, Learning, and Design using Probabilistic Programming*. PhD thesis, 2017.

Rainforth, T. Nesting probabilistic programs. In *UAI*, 2018.

Rainforth, T., Le, T. A., van de Meent, J.-W., Osborne, M. A., and Wood, F. Bayesian optimization for probabilistic programs. In *NIPS*, 2016.

Rainforth, T., Kosiorek, A. R., Le, T. A., Maddison, C. J., Igl, M., Wood, F., and Teh, Y. W. Tighter variational bounds are not necessarily better. In *ICML*, 2018.

Rudolf, D. and Schweizer, N. Perturbation theory for Markov chains via Wasserstein distance. *arXiv preprint arXiv:1503.04123*, 2015.

Sebastiani, P. and Wynn, H. P. Maximum entropy sampling and optimal Bayesian experimental design. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(1):145–157, 2000.

Stuhlmüller, A. and Goodman, N. D. A dynamic programming algorithm for inference in recursive probabilistic programs. In *Second Statistical Relational AI workshop at UAI 2012 (StaRAI-12)*, 2012.

Stuhlmüller, A. and Goodman, N. D. Reasoning about reasoning by nested conditioning: Modeling theory of mind with probabilistic programs. *Cognitive Systems Research*, 28:80–99, 2014.

Vincent, B. T. Hierarchical Bayesian estimation and hypothesis testing for delay discounting tasks. *Behavior research methods*, 48(4):1608–1620, 2016.

Vincent, B. T. and Rainforth, T. The DARC toolbox: automated, flexible, and efficient delayed and risky choice experiments using Bayesian adaptive design. *PsyArXiv*, 2017.

Wood, F., van de Meent, J. W., and Mansinghka, V. A new approach to probabilistic programming inference. In *AISTATS*, pp. 2–46, 2014.