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 and 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. This can occur, for example, when nesting probabilistic programs (Mantadelis and Janssens, 2011; Le et al., 2016). Some problems are even multiply-intractable, such that they require multiple levels of nesting to encode (Stuhlmüller and Goodman, 2014). 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 and 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 and 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 and 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>.

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 and Schwartz, 2001; Belomestny et al., 2010; Gordy and 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 and 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 and 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 $\phi \colon \mathcal{Y} \times \mathcal{Z} \to \Phi$. 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
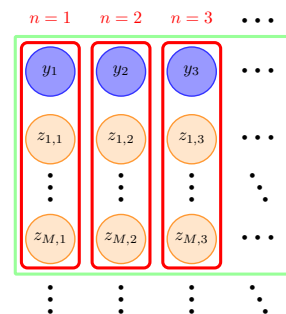

Figure 1. Informal convergence representation

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 $|I - \mathbb{E}\left[f(y_n, (\hat{\gamma}_M)_n)\right]| < \varepsilon$ (we will characterize the exact requirements for this later). For this fixed M, we have a standard MC estimator on an extended space $y, z_1, \ldots, z_M$ 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 $I_{N,M} \to \mathbb{E}\left[f(y_n, (\hat{\gamma}_M)_n)\right]$ 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 and 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 and 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 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 $\mathbb{E}\left[(\epsilon_M)_1\right] \to 0$ as $M \to \infty$ (noting that $(\epsilon_M)_n$ are i.i.d. and so $\mathbb{E}\left[(\epsilon_M)_1\right] = \mathbb{E}\left[(\epsilon_M)_n\right], \forall n \in \mathbb{N}$). 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 $\mathbb{E}\left[(\epsilon_M)_1\right] \to 0$ 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 $\omega \notin A_\delta$,
$$I_{\tau_\delta(M),M}(\omega) \overset{a.s.}{\to} I \quad as \quad M \to \infty.$$

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 the 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 $\mathbb{E}\left[(\epsilon_M)_1\right] \to 0$.

3.2 Repeated Nesting and Exact Bounds

We next consider the case of multiple levels of nesting. As previously explained, 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
$$\gamma_D\left(y^{(0:D-1)}\right) = \mathbb{E}\left[f_D\left(y^{(0:D)}\right)\Big|y^{(0:D-1)}\right] \quad \text{and}$$
$$\gamma_k(y^{(0:k-1)}) = \mathbb{E}\left[f_k\left(y^{(0:k)}, \gamma_{k+1}\left(y^{(0:k)}\right)\right)\Big|y^{(0:k-1)}\right],$$
for $0 \le k < D$, 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 $k \in 0, \ldots, D - 1$ and if

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

then

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

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

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

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

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

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

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

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

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

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

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

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

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, which occurs 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 ℓth such draw for the ℓth 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

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

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

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

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

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-
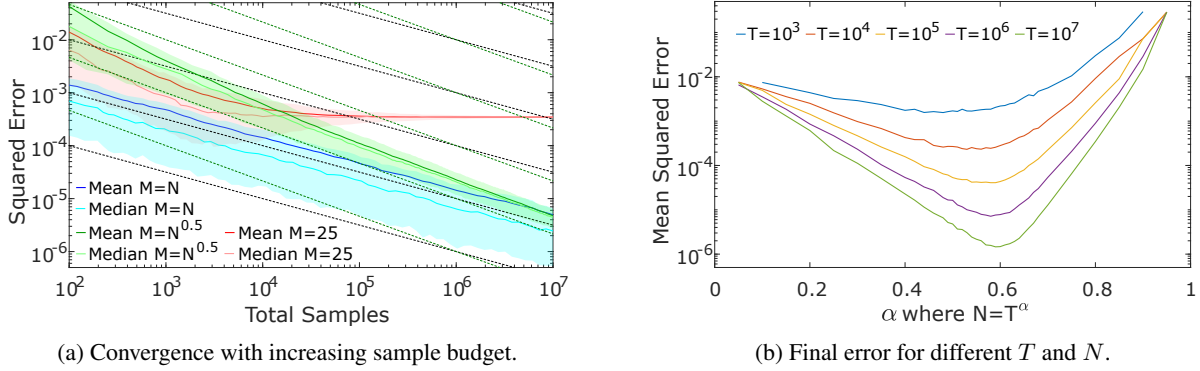
(a) Convergence with increasing sample budget.

(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.

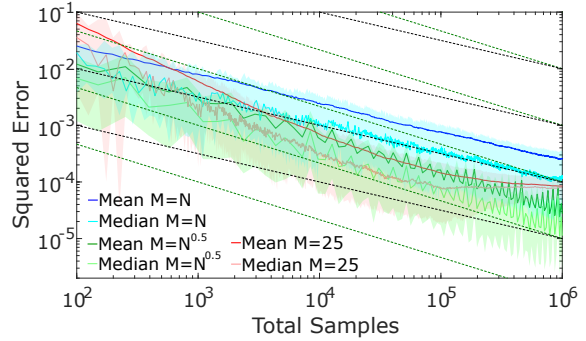To verify the convergence rate, we repeated the analysis

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.

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.

5.3 Repeated Nesting

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

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

$$f_0\left(y^{(0)}, \gamma_1\left(y^{(0)}\right)\right) = \log \gamma_1\left(y^{(0)}\right) \quad (20a)$$
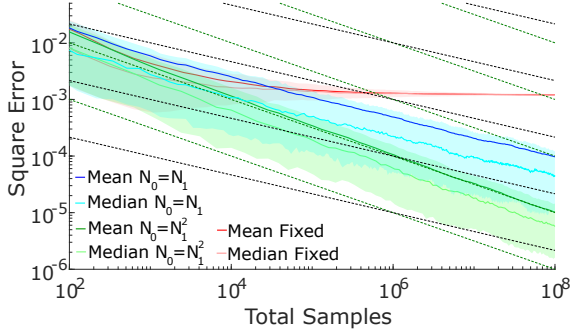
Figure 4. Empirical convergence of NMC to (20) for an increasing total sample budget $T = N_0 N_1 N_2$. Setup and conventions as per Figure 2a. Shown in red is the convergence with a fixed $N_2 = 5$ and $N_0 = N_1^2$, which we see gives a biased solution. Shown in blue is the convergence when setting $N_0 = N_1 = N_2$, 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 $O(T^{-1/2})$.

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

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

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 $T = N_0 N_1 N_2$. 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
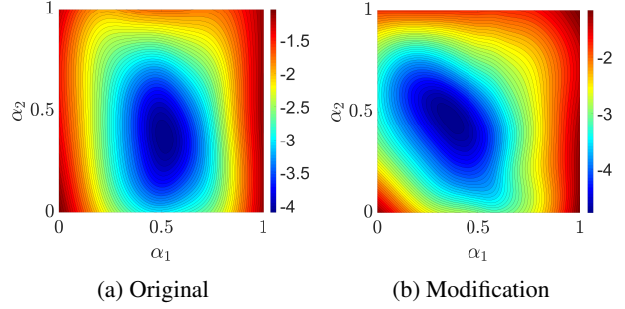

(a) Original   (b) Modification

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.

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.

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 and Verdinelli, 1995; Sebastiani and 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

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

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

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

where $\theta_{n,m} \sim p(\theta)$ for each $(m, n) \in \{0, \ldots, M\} \times \{1, \ldots, N\}$, and $y_n \sim p(y|\theta = \theta_{n,0}, d)$ for each $n \in \{1, \ldots, N\}$. 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 $y_1, \ldots, y_c$, we use the ideas introduced in Section 4.2 to
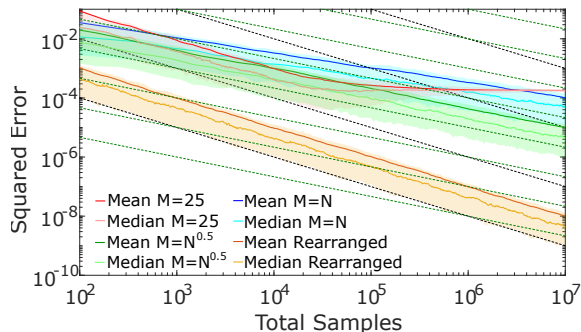
Figure 6. Convergence of NMC (i.e. (22)) and our reformulated estimator (23) for the BED problem. Experimental setup and conventions are as per Figure 2a, with a ground truth estimate made using a single run of the reformulated estimator with 10^{10} samples. We see that the theoretical convergence rates are observed, with the advantages of the reformulated estimator particularly pronounced.

derive the following improved estimator

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

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

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 and Welling, 2013) use lower bounds on the model evidence as objectives for train deep generative models and employ estimators of the form

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

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 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 $\mathbb{E}\left[(I_{N,M} - I)^2\right] \leq \frac{C_0^2 \varsigma_1^4}{4M^2}\left(1 + \frac{1}{N}\right) + \frac{K_0^2 \varsigma_1^2}{NM} + \frac{C_0 K_0 \varsigma_1^3}{NM^{3/2}} + O(\frac{1}{M^3})$. 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 and 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.

Appendix A   Proof of Theorem 1 - Simplified Convergence Rate

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 $O(1/N + 1/M)$.*

Proof. Though the Theorem follows directly from Theorem 3, we also provide the following proof for this simplified case to provide a more accessible intuition behind the result. Note that the approach taken is distinct from the proof of Theorem 3.

Using Minkowski's inequality, we can bound the mean squared error of $I_{N,M}$ by

$$\mathbb{E}[(I - I_{N,M})^2] = \|I - I_{N,M}\|_2^2 \leq U^2 + V^2 + 2UV \leq 2\left(U^2 + V^2\right) \tag{25}$$

$$\text{where} \quad U = \left\|I - \frac{1}{N}\sum_{n=1}^{N} f(y_n, \gamma(y_n))\right\|_2 \quad \text{and} \quad V = \left\|\frac{1}{N}\sum_{n=1}^{N} f(y_n, \gamma(y_n)) - I_{N,M}\right\|_2.$$

We see immediately that $U = O\left(1/\sqrt{N}\right)$, since $\frac{1}{N}\sum_{n=1}^{N} f(y_n, \gamma(y_n))$ is a MC estimator for I, noting our assumption that $f(y_n, \gamma(y_n)) \in L^2$. For the second term,

$$\begin{aligned} V &= \left\|\frac{1}{N}\sum_{n=1}^{N} f(y_n, (\hat{\gamma}_M)_n) - f(y_n, \gamma(y_n))\right\|_2 \\ &\leq \frac{1}{N}\sum_{n=1}^{N} \|f(y_n, (\hat{\gamma}_M)_n) - f(y_n, \gamma(y_n))\|_2 \leq \frac{1}{N}\sum_{n=1}^{N} K\|(\hat{\gamma}_M)_n - \gamma(y_n)\|_2 \end{aligned}$$

where K is a fixed constant, again by Minkowski and using the assumption that f is Lipschitz. We can rewrite

$$\|(\hat{\gamma}_M)_n - \gamma(y_n)\|_2^2 = \mathbb{E}\left[\mathbb{E}\left[((\hat{\gamma}_M)_n - \gamma(y_n))^2 \big| y_n\right]\right].$$

by the tower property of conditional expectation, and note that

$$\mathbb{E}\left[((\hat{\gamma}_M)_n - \gamma(y_n))^2 \big| y_n\right] = \mathrm{Var}\left(\frac{1}{M}\sum_{m=1}^{M}\phi(y_n, z_{n,m})\bigg| y_n\right) = \frac{1}{M}\mathrm{Var}\left(\phi(y_n, z_{n,1}) | y_n\right)$$

since each $z_{n,m}$ is i.i.d. and conditionally independent given y_n. As such

$$\|(\hat{\gamma}_M)_n - \gamma(y_n)\|_2^2 = \frac{1}{M}\mathbb{E}\left[\mathrm{Var}\left(\phi(y_n, z_{n,1}) | y_n\right)\right] = O(1/M),$$

noting that $\mathbb{E}\left[\mathrm{Var}\left(\phi(y_n, z_{n,1}) | y_n\right)\right]$ is a finite constant by our assumption that $\phi(y_n, z_{n,m}) \in L^2$. Consequently,

$$V \leq \frac{NK}{N}O\left(1/\sqrt{M}\right) = O\left(1/\sqrt{M}\right).$$

Substituting these bounds for U and V in (25) gives

$$\|I - I_{N,M}\|_2^2 \leq 2\left(O\left(1/\sqrt{N}\right)^2 + O\left(1/\sqrt{M}\right)^2\right) = O(1/N + 1/M)$$

as desired. □

Appendix B   The Inevitable Bias of Nested Estimation

In this section we demonstrate formally that NMC schemes must produce biased estimates of $I(f)$ for certain functions f. In fact, our result applies more generally: we show that this holds for any MC scheme that makes use of imperfect estimates $\hat{\zeta}_n$ of $\gamma(y_n)$, either via a NMC procedure (e.g. $\hat{\zeta}_n = (\hat{\gamma}_M)_n$), or when these inner estimates are generated by some other methods such as a variational approximation (Blei et al., 2016) or Bayesian quadrature (O'Hagan, 1991).

Theorem 6. *Suppose that the random variables $\hat{\zeta}_n$ satisfy $\mathbb{P}(\hat{\zeta}_n \neq \gamma(y_n)) > 0$. Then we can choose f such that if $y_n \sim p(y)$, $\mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N} f(y_n, \hat{\zeta}_n)\right] \neq I(f)$ for any N (including the limit $N \to \infty$).*

Proof. Take $f(y, w) = (\gamma(y) - w)^2$. Then $f(y, \gamma(y)) = 0$, so that $I(f) = 0$. On the other hand, $f(y_n, \hat{\zeta}_n) \geq 0$ since f is non-negative. Moreover, $f(y_n, \hat{\zeta}_n) > 0$ on the event $\{\hat{\zeta}_n \neq \gamma(y_n)\}$. Since we assumed this event has nonzero probability, it follows that $\mathbb{E}\left[f(y_n, \hat{\zeta}_n)\right] > 0$ and hence

$$\mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N} f(y_n, \hat{\zeta}_n)\right] = \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\left[f(y_n, \hat{\zeta}_n)\right] > 0 = I(f)$$

which gives the required result. □

It also follows from Jensen's inequality that *any* strictly convex or concave f entails a biased estimator when $\hat{\zeta}_n$ is unbiased but has non-zero variance given y_n, e.g. when $\hat{\zeta}_n$ is a MC estimate. More formally we have

Theorem 7. *Suppose that $y_n \sim p(y)$ and that each $\hat{\zeta}_n$ satisfies $\mathbb{E}\left[\hat{\zeta}_n \middle| y_n\right] = \gamma(y_n)$. Define $\mathcal{A} \subseteq \mathcal{Y}$ as $\mathcal{A} = \left\{y \in \mathcal{Y} \middle| \operatorname{Var}\left(\hat{\zeta}_n \middle| y_n = y\right) > 0\right\}$ and assume that $\mathbb{P}(y_n \in \mathcal{A}) > 0$. Then for any f that is strictly convex in its second argument,*

$$\mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N} f(y_n, \hat{\zeta}_n)\right] > I(f).$$

Similarly for any f that is strictly concave in its second argument,

$$\mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N} f(y_n, \hat{\zeta}_n)\right] < I(f).$$

Proof. We prove our claim for the case that f is strictly convex; our proof for the other concave case is symmetrical. We have

$$\mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N} f(y_n, \hat{\zeta}_n)\right] = \mathbb{E}\left[f(y_1, \hat{\zeta}_1)\right] = \mathbb{E}\left[\mathbb{E}\left[f(y_1, \hat{\zeta}_1)\middle| y_1\right]\right] \geq \mathbb{E}\left[f\left(y_1, \mathbb{E}\left[\hat{\zeta}_1\middle| y_1\right]\right)\right] = I(f)$$

where the $\geq$ is a result of Jensen's inequality on the inner expectation. Since f is strictly convex and therefore non-linear, equality holds if and only if $\hat{\zeta}_1$ is almost surely constant given y_1. This is violated whenever $y_1 \in \mathcal{A}$, which by assumption has a non-zero probability of occurring. Consequently, the inequality must be strict, giving the desired result. □

In addition to some special cases discussed in the Section 4, it may still be possible to develop unbiased estimation schemes for certain non-linear f using Russian roulette sampling (Lyne et al., 2015) or other debiasing techniques. However, these induce their own complications: for some problems the resultant estimates have infinite variance (Lyne et al., 2015) and as shown by (Jacob et al., 2015), there are no general purpose "f-factories" that produce both non-negative and unbiased estimates for non-constant, positive output functions $f : \mathbb{R} \to \mathbb{R}^+$, given unbiased estimates for the inputs.

Appendix C   Proof of Theorem 2 - "Almost almost sure" convergence

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 $\mathbb{E}\left[(\epsilon_M)_1\right] \to 0$ 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 $\omega \notin A_\delta$,

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

Proof. For all N, M, we have by the triangle inequality that

$$|I_{N,M} - I| \leq V_{N,M} + U_N, \quad \text{where}$$

$$V_{N,M} = \left|\frac{1}{N}\sum_{n=1}^{N} f(y_n, \gamma(y_n)) - I_{N,M}\right| \quad \text{and} \quad U_N = \left|I - \frac{1}{N}\sum_{n=1}^{N} f(y_n, \gamma(y_n))\right|.$$

A second application of the triangle inequality then allows us to write

$$V_{N,M} \leq \frac{1}{N}\sum_{n=1}^{N} (\epsilon_M)_n$$

where we recall that $(\epsilon_M)_n = |f(y_n, \gamma(y_n)) - f(y_n, \hat{\gamma}_n)|$. Now, for all fixed M, each $(\epsilon_M)_n$ is i.i.d. Furthermore, since $\mathbb{E}\left[(\epsilon_M)_1\right] \to 0$ as $M \to \infty$ by our assumption and $(\epsilon_M)_n$ is nonnegative, there exists some $L \in \mathbb{N}$ such that

$\mathbb{E}\left[|(\epsilon_M)_n|\right] < \infty$ for all $M \geq L$. Consequently, the strong law of large numbers means that as $N \to \infty$ then for all $M \geq L$

$$\frac{1}{N}\sum_{n=1}^{N}(\epsilon_M)_n \overset{a.s.}{\to} \mathbb{E}\left[(\epsilon_M)_1\right]. \tag{26}$$

For any fixed $\delta > 0$ then by repeatedly applying Egorov's theorem to each $M \geq L$, we can find a sequence of events

$$B_L, B_{L+1}, B_{L+2}, \ldots$$

such that for every $M \geq L$,

$$\mathbb{P}(B_M) < \frac{\delta}{4} \cdot \frac{1}{2^{M-L}}$$

and outside of B_M, the sequence $\frac{1}{N}\sum_{n=1}^{N}(\epsilon_M)_n$ converges *uniformly* to $\mathbb{E}\left[(\epsilon_M)_1\right]$. This uniform convergence (as opposed to just the piecewise convergence implied by (26)) now guarantees that we can define some function $\tau_\delta^1 : \mathbb{N} \to \mathbb{N}$ such that

$$\left|\frac{1}{M'}\sum_{n=1}^{M'}(\epsilon_M)_n(\omega) - \mathbb{E}\left[(\epsilon_M)_1\right]\right| < \frac{1}{M} \tag{27}$$

for all $M \geq L$, $M' \geq \tau_\delta^1(M)$, and $\omega \notin B_M$, remembering that ω is a point in our sample space. We further have that (27) holds for all $M \geq M_0$, $M' \geq \tau_\delta^1(M)$, and $\omega \notin B_\delta := \bigcup_{M \geq L} B_M$. Consequently, for all such M, M' and ω,

$$V_{M',M}(\omega) \leq \frac{1}{M'}\sum_{n=1}^{M'}(\epsilon_M)_n(\omega) < \frac{1}{M} + \mathbb{E}\left[(\epsilon_M)_1\right], \tag{28}$$

while we also have

$$\mathbb{P}(B_\delta) \leq \sum_{M \geq L}\mathbb{P}(B_M) < \sum_{M \geq L}\frac{\delta}{4}\cdot\frac{1}{2^{M-L}} = \frac{\delta}{2}. \tag{29}$$

To complete the proof, we must remove the dependence of U_N on N as well. This is straightforward once we observe that $U_N \overset{a.s.}{\to} 0$ as $N \to \infty$ by the strong law of large numbers. So, by Egorov's theorem again, there exists an event C_δ such that

$$\mathbb{P}(C_\delta) < \frac{\delta}{2} \tag{30}$$

and outside of C_δ, the sequence U_N converges uniformly to 0. This uniform convergence, in turn, ensures the existence of a function $\tau_\delta^2 : \mathbb{N} \to \mathbb{N}$ such that

$$U_{M'}(\omega) < \frac{1}{M} \tag{31}$$

for all $M \in \mathbb{N}$, $M' \geq \tau_\delta^2(M)$, and $\omega \notin C_\delta$.

We can now define $\tau_\delta(M) = \max(\tau_\delta^1(M), \tau_\delta^2(M))$, and $A_\delta = B_\delta \cup C_\delta$. By inequalities in (29) and (30),

$$\mathbb{P}(A_\delta) \leq \mathbb{P}(B_\delta) + \mathbb{P}(C_\delta) < \delta.$$

Also, by the inequalities in (28) and (31),

$$\left|I - I_{\tau_\delta(M),M}(\omega)\right| \leq V_{\tau_\delta(M),M}(\omega) + U_{\tau_\delta(M)}(\omega) \leq \frac{1}{M} + \frac{1}{M} + \mathbb{E}\left[(\epsilon_M)_1\right]$$

for all $M \geq L$ and $\omega \notin A_\delta$. Since $\mathbb{E}\left[(\epsilon_M)_1\right] \to 0$, we have here that $I_{\tau_\delta(M),M}(\omega) \to I$ as desired. $\qquad\square$

Appendix D   Proof of Theorem 3 - Convergence for Repeated Nesting

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 $k \in 0, \ldots, D-1$ and if

$$\varsigma_k^2 := \mathbb{E}\left[\left(f_k\left(y^{(0:k)}, \gamma_{k+1}\left(y^{(0:k)}\right)\right) - \gamma_k\left(y^{(0:k-1)}\right)\right)^2\right]$$

$$< \infty \quad \forall k \in 0, \ldots, D$$

then

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

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

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

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

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

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

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

Proof. As this is a long and involved proof, we start by defining a number of useful terms that will be used throughout. Unless otherwise stated, these definitions hold for all $k \in \{0, \ldots, D\}$. Note that most of these terms implicitly depend on the number of samples $N_0, N_1, \ldots, N_D$. However, s_k, $\zeta_{d,k}$, and ς_k do not and are thus constants for a particular problem. $E_k\left(y^{(0:k-1)}\right)$ is the MSE of the estimator at depth k given $y^{(0:k-1)}$

$$E_k\left(y^{(0:k-1)}\right) := \mathbb{E}\left[\left.\left(I_k\left(y^{(0:k-1)}\right) - \gamma_k\left(y^{(0:k-1)}\right)\right)^2\right|y^{(0:k-1)}\right] \tag{32}$$

$\bar{f}_k\left(y^{(0:k-1)}\right)$ is the expected value of the estimate at depth k, or equivalently the expected function output using the estimate of the layer below

$$\bar{f}_k\left(y^{(0:k-1)}\right) := \mathbb{E}\left[\left.I_k\left(y^{(0:k-1)}\right)\right|y^{(0:k-1)}\right] \quad \forall k \in \{1, \ldots, D-1\}$$
$$= \mathbb{E}\left[\left.f_k\left(y^{(0:k)}, I_{k+1}\left(y^{(0:k)}\right)\right)\right|y^{(0:k-1)}\right] \tag{33}$$

$v_k^2\left(y^{(0:k-1)}\right)$ is the variance of the estimator at depth k

$$v_k^2\left(y^{(0:k-1)}\right) := \text{Var}\left[\left.I_k\left(y^{(0:k-1)}\right)\right|y^{(0:k-1)}\right]$$
$$= \mathbb{E}\left[\left.\left(I_k\left(y^{(0:k-1)}\right) - \bar{f}_k\left(y^{(0:k-1)}\right)\right)^2\right|y^{(0:k-1)}\right] \tag{34}$$

$\beta_k\left(y^{(0:k-1)}\right)$ is the bias of the estimator at depth k

$$\beta_k\left(y^{(0:k-1)}\right) := \mathbb{E}\left[\left.I_k\left(y^{(0:k-1)}\right) - \gamma_k\left(y^{(0:k-1)}\right)\right|y^{(0:k-1)}\right]$$
$$= \bar{f}_k\left(y^{(0:k-1)}\right) - \gamma_k\left(y^{(0:k-1)}\right)$$
$$= \mathbb{E}\left[\left.f_k\left(y^{(0:k)}, I_{k+1}\left(y^{(0:k)}\right)\right) - f_k\left(y^{(0:k)}, \gamma_{k+1}\left(y^{(0:k)}\right)\right)\right|y^{(0:k-1)}\right] \tag{35}$$

$s_k^2\left(y^{(0:k-1)}\right)$ is the variance at depth k of the true function output

$$s_k^2\left(y^{(0:k-1)}\right) := \mathbb{E}\left[\left(f_k\left(y^{(0:k)},\gamma_{k+1}\left(y^{(0:k)}\right)\right) - \gamma_k\left(y^{(0:k-1)}\right)\right)^2\middle|y^{(0:k-1)}\right] \tag{36}$$

$$s_D^2\left(y^{(0:D-1)}\right) := \mathbb{E}\left[\left(f_D\left(y^{(0:D)}\right) - \gamma_D\left(y^{(0:D)}\right)\right)^2\middle|y^{(0:D-1)}\right] \tag{37}$$

$\zeta_{d,k}^2\left(y^{(0:k-1)}\right)$ is expectation of $s_d^2\left(y^{(0:d-1)}\right)$ over $y^{(k:d-1)}$

$$\begin{aligned}\zeta_{d,k}^2\left(y^{(0:k-1)}\right) &:= \mathbb{E}\left[s_d^2\left(y^{(0:d-1)}\right)\middle|y^{(0:k-1)}\right]\\ &= \mathbb{E}\left[\left(f_d\left(y^{(0:d)},\gamma_{d+1}\left(y^{(0:d)}\right)\right) - \gamma_d\left(y^{(0:d-1)}\right)\right)^2\middle|y^{(0:k-1)}\right]\end{aligned} \tag{38}$$

ς_k^2 is expectation of $s_k^2\left(y^{(0:k-1)}\right)$ over all $y^{(0:k-1)}$

$$\varsigma_k^2 := \zeta_{k,0}^2 = \mathbb{E}\left[\left(f_k\left(y^{(0:k)},\gamma_{k+1}\left(y^{(0:k)}\right)\right) - \gamma_k\left(y^{(0:k-1)}\right)\right)^2\right] \tag{39}$$

$A_k\left(y^{(0:k-1)}\right)$ is the MSE in the function output from using the estimate of the next layer, rather than the true value $\gamma_{k+1}\left(y^{(0:k)}\right)$, we fix $A_D := 0$

$$A_k\left(y^{(0:k-1)}\right) := \mathbb{E}\left[\left(f_k\left(y^{(0:k)},I_{k+1}\left(y^{(0:k)}\right)\right) - f_k\left(y^{(0:k)},\gamma_{k+1}\left(y^{(0:k)}\right)\right)\right)^2\middle|y^{(0:k-1)}\right] \tag{40}$$

$\sigma_k^2\left(y^{(0:k-1)}\right)$ is the variance in the function output from using the estimate of the next layer, we fix $\sigma_D^2\left(y^{0:D-1}\right) := s_D^2\left(y^{0:D-1}\right)$

$$\begin{aligned}\sigma_k^2\left(y^{(0:k-1)}\right) &:= \mathrm{Var}\left[f_k\left(y^{(0:k)},I_{k+1}\left(y^{(0:k)}\right)\right)\middle|y^{(0:k-1)}\right]\\ &= \mathbb{E}\left[\left(f_k\left(y^{(0:k)},I_{k+1}\left(y^{(0:k)}\right)\right) - \bar{f}_k\left(y^{(0:k-1)}\right)\right)^2\middle|y^{(0:k-1)}\right]\end{aligned} \tag{41}$$

$\omega_k\left(y^{(0:k-1)}\right)$ is the expectation over $y^{(k)}$ of the MSE for the next layer, we fix $\omega_D\left(y^{(0:D-1)}\right) := 0$

$$\begin{aligned}\omega_k\left(y^{(0:k-1)}\right) &:= \mathbb{E}\left[E_{k+1}\left(y^{(0:k)}\right)\middle|y^{(0:k-1)}\right]\\ &= \mathbb{E}\left[\left(I_{k+1}\left(y^{(0:k)}\right) - \gamma_{k+1}\left(y^{(0:k)}\right)\right)^2\middle|y^{(0:k-1)}\right]\end{aligned} \tag{42}$$

$\lambda_k\left(y^{(0:k-1)}\right)$ is the expectation over $y^{(k)}$ of the magnitude of the bias for the next layer, we fix $\lambda_D\left(y^{(0:D-1)}\right) := 0$ and note that $\lambda_{D-1}\left(y^{(0:D-2)}\right) := 0$ also as the innermost layer is an unbiased

$$\begin{aligned}\lambda_k\left(y^{(0:k-1)}\right) &:= \mathbb{E}\left[\left|\beta_{k+1}\left(y^{(0:k)}\right)\right|\,\middle|\,y^{(0:k-1)}\right]\\ &= \mathbb{E}\left[\left|\mathbb{E}\left[\left(I_{k+1}\left(y^{(0:k)}\right) - \gamma_{k+1}\left(y^{(0:k)}\right)\right)\middle|y^{(0:k)}\right]\right|\,\middle|\,y^{(0:k-1)}\right]\end{aligned} \tag{43}$$

Lipschitz Continuous Case

Given these definitions, we start by breaking the error down into a variance and bias term. Using the standard bias-variance decomposition we have

$$\begin{aligned}E_k\left(y^{(0:k-1)}\right) &= \mathbb{E}\left[\left(I_k\left(y^{(0:k-1)}\right) - \gamma_k\left(y^{(0:k-1)}\right)\right)^2\middle|y^{(0:k-1)}\right]\\ &= v_k^2\left(y^{(0:k-1)}\right) + \left(\beta_k\left(y^{(0:k-1)}\right)\right)^2\end{aligned} \tag{44}$$

It is immediately clear from its definition in (35) that the bias term $\left(\beta_k\left(y^{(0:k-1)}\right)\right)^2$ is independent of N_0. On the other hand, we will show later that the dominant components of the variance term for large $N_{0:D}$ depend only on N_0. We can thus think of increasing N_0 as reducing the variance of the estimator and increasing $N_{1:D}$ as reducing the bias.

We first consider the variance term

$$v_k^2\left(y^{(0:k-1)}\right) = \mathbb{E}\left[\left(\left(\frac{1}{N_k}\sum_{n=1}^{N_k}f_k\left(y_n^{(0:k)}, I_{k+1}\left(y_n^{(0:k)}\right)\right) - \bar{f}_k\left(y^{(0:k-1)}\right)\right)^2\middle|y^{(0:k-1)}\right]\right.$$

$$= \frac{1}{N_k}\mathbb{E}\left[\left(f_k\left(y^{(0:k)}, I_{k+1}\left(y^{(0:k)}\right)\right) - \bar{f}_k\left(y^{(0:k-1)}\right)\right)^2\middle|y^{(0:k-1)}\right]$$

with the equality following because the $y_n^{(0:k)}$ being drawn i.i.d. and the expectation of each $f_k\left(y^{(0:k)}, I_{k+1}\left(y^{(0:k)}\right)\right)$ equaling $\bar{f}_k\left(y^{(0:k-1)}\right)$ means that all the cross terms are zero. By the definition of σ_k^2 we now have

$$v_k^2\left(y^{(0:k-1)}\right) = \frac{\sigma_k^2\left(y^{(0:k-1)}\right)}{N_k}. \tag{45}$$

By using Minkowski's inequality and the definition of A_k it also follows that

$$\sigma_k\left(y^{(0:k-1)}\right) \le \left(A_k\left(y^{(0:k-1)}\right)\right)^{\frac{1}{2}} + \left(\mathbb{E}\left[\left(f_k\left(y^{(0:k)}, \gamma_{k+1}\left(y^{(0:k)}\right)\right) - \bar{f}_k\left(y^{(0:k-1)}\right)\right)^2\middle|y^{(0:k-1)}\right]\right)^{\frac{1}{2}}. \tag{46}$$

Using a bias-variance decomposition on the second term above and noting that $s_k^2\left(y^{(0:k-1)}\right)$ and $\bar{f}_k\left(y^{(0:k-1)}\right) - \beta_k\left(y^{(0:k-1)}\right)$ are respectively the variance and expectation of $f_k\left(y^{(0:k)}, \gamma_{k+1}\left(y^{(0:k)}\right)\right)$, we can rearrange the right-hand size of (46) to give

$$\sigma_k\left(y^{(0:k-1)}\right) \le \left(A_k\left(y^{(0:k-1)}\right)\right)^{\frac{1}{2}} + \left(s_k^2\left(y^{(0:k-1)}\right) + \left(\beta_k\left(y^{(0:k-1)}\right)\right)^2\right)^{\frac{1}{2}}. \tag{47}$$

Here s_k^2 is independent of the number of samples used at any level of the estimate, while A_k and β_k^2 are independent of $N_d\ \forall d \le k$. Now by Jensen's inequality, we have that

$$\left(\beta_k\left(y^{(0:k-1)}\right)\right)^2 \le A_k\left(y^{(0:k-1)}\right) \tag{48}$$

noting that the only difference in the definition of $\left(\beta_k\left(y^{(0:k-1)}\right)\right)^2$ and $A_k\left(y^{(0:k-1)}\right)$ is whether the squaring occurs inside or outside the expectation. Therefore, presuming that A_k does not increase with $N_d\ \forall d > k$, neither will $\sigma_k^2\left(y^{(0:k-1)}\right)$, and so the variance term will converge to zero with rate $O(1/N_k)$. Further, if $A_k \to 0$ as $N_{k+1}, \ldots, N_D \to \infty$, then for a large number of inner samples $\sigma_k^2 \to s_k^2$ and thus we will have $v_k^2\left(y^{(0:k-1)}\right) \le \frac{s_k^2}{N_k} + O\left(\epsilon\right)$ where $O\left(\epsilon\right)$ represents higher order terms that are dominated in the limit $N_k, \ldots, N_D \to \infty$. Provided this holds, we will also, therefore, have that

$$E_k\left(y^{(0:k-1)}\right) = \frac{\sigma_k^2\left(y^{(0:k-1)}\right)}{N_k} + \beta_k^2\left(y^{(0:k-1)}\right) = \frac{s_k^2\left(y^{(0:k-1)}\right)}{N_k} + \beta_k^2\left(y^{(0:k-1)}\right) + O(\epsilon). \tag{49}$$

We now show that Lipschitz continuity is sufficient for $A_k \to 0$ and derive a concrete bound on the variance by bounding A_k. By definition of Lipschitz continuity, we have that

$$\left(A_k\left(y^{(0:k-1)}\right)\right)^{\frac{1}{2}} \le \left(\mathbb{E}\left[K_k^2\left(I_{k+1}\left(y^{(0:k)}\right) - \gamma_{k+1}\left(y^{(0:k)}\right)\right)^2\middle|y^{(0:k-1)}\right]\right)^{\frac{1}{2}}$$

$$= K_k\left(\omega_k\left(y^{(0:k-1)}\right)\right)^{\frac{1}{2}} \tag{50}$$

where we remember that $\omega_k\left(y^{(0:k-1)}\right) = \mathbb{E}\left[E_{k+1}\left(y^{(0:k)}\right)\middle|y^{(0:k-1)}\right]$ is the expected MSE of the next level estimator. Once we also have an expression for the bias, we will thus be able to use this bound on A_k along with (44), (45), and (47) to inductively derive a bound on the error.

For the case where we only assume Lipschitz continuity then we will simply use the bound on the bias given by (48) leading to

$$E_k\left(y^{(0:k-1)}\right) \le \frac{\sigma_k^2\left(y^{(0:k-1)}\right)}{N_k} + A_k\left(y^{(0:k-1)}\right) \tag{51}$$

$$\le \frac{s_k^2\left(y^{(0:k-1)}\right) + 2A_k\left(y^{(0:k-1)}\right) + 2\left(A_k\left(y^{(0:k-1)}\right)\right)^{\frac{1}{2}}\left(s_k^2\left(y^{(0:k-1)}\right) + A_k\left(y^{(0:k-1)}\right)\right)^{\frac{1}{2}}}{N_k} + A_k\left(y^{(0:k-1)}\right)$$

$$= \frac{s_k^2\left(y^{(0:k-1)}\right) + 2K_k^2\omega_k\left(y^{(0:k-1)}\right)}{N_k} + K_k^2\omega_k\left(y^{(0:k-1)}\right)$$

$$+ \frac{2K_k\left(\omega_k\left(y^{(0:k-1)}\right)\right)^{\frac{1}{2}}\left(s_k^2\left(y^{(0:k-1)}\right) + K_k^2\omega_k\left(y^{(0:k-1)}\right)\right)^{\frac{1}{2}}}{N_k}$$

$$\leq \frac{s_k^2\left(y^{(0:k-1)}\right) + 4K_k^2\omega_k\left(y^{(0:k-1)}\right) + 2K_k\left(\omega_k\left(y^{(0:k-1)}\right)\right)^{\frac{1}{2}} s_k\left(y^{(0:k-1)}\right)}{N_k} + K_k^2\omega_k\left(y^{(0:k-1)}\right) \tag{52}$$

which fully defines a bound on conditional the variance of one layer given the mean squared error of the layer below. In particular as $\omega_D\left(y^{(0:D-1)}\right) = 0$ we now have

$$E_D\left(y^{(0:D-1)}\right) \leq \frac{s_D^2\left(y^{(0:D-1)}\right)}{N_D} = \frac{\mathbb{E}\left[\left(f_D\left(y^{(0:D)}\right) - \gamma_D\left(y^{(0:D)}\right)\right)^2 \middle| y^{(0:D-1)}\right]}{N_D}$$

which is the standard error for Monte Carlo convergence. We further have

$$\omega_{D-1}\left(y^{(0:D-2)}\right) = \mathbb{E}\left[E_D\left(y^{(0:D-1)}\right)\middle| y^{(0:D-2)}\right] = \frac{\zeta_{D,D-1}^2\left(y^{(0:D-2)}\right)}{N_D}.$$

and thus

$$\begin{aligned}E_{D-1}\left(y^{(0:D-2)}\right) \leq &\frac{s_{D-1}^2\left(y^{(0:D-2)}\right)}{N_{D-1}} + \frac{4K_{D-1}^2\zeta_{D,D-1}^2\left(y^{(0:D-2)}\right)}{N_D N_{D-1}} \\ &+ \frac{2K_{D-1}s_{D-1}\left(y^{(0:D-2)}\right)\zeta_{D,D-1}\left(y^{(0:D-2)}\right)}{N_{D-1}\sqrt{N_D}} + \frac{K_{D-1}^2\zeta_{D,D-1}^2\left(y^{(0:D-2)}\right)}{N_D}.\end{aligned} \tag{53}$$

This leads to the following result for the single nesting case

$$E_0 \leq \frac{\varsigma_0^2}{N_0} + \frac{4K_0^2\varsigma_1^2}{N_0 N_1} + \frac{2K_0\varsigma_0\varsigma_1}{N_0\sqrt{N_1}} + \frac{K_0^2\varsigma_1^2}{N_1} \tag{54}$$

$\approx \frac{\varsigma_0^2}{N_0} + \frac{K_0^2\varsigma_1^2}{N_1} = O\left(\frac{1}{N_0} + \frac{1}{N_1}\right)$ where the approximation becomes exact as $N_0, N_1 \to \infty$. Note that there is no $O\left(\epsilon\right)$ term as this bound is exact in the finite sample case.

Things quickly get messy for double nesting and beyond so we will ignore non-dominant terms in the limit $N_0, \ldots, N_D \to \infty$ and resort to using $O(\epsilon)$ for these instead. We first note that removing dominated terms from (52) gives

$$E_k\left(y^{(0:k-1)}\right) \leq \frac{s_k^2}{N_k} + K_k^2\omega_k\left(y^{(0:k-1)}\right) + O(\epsilon) \tag{55}$$

as s_k^2 does not decrease with increasing $N_{k+1:D}$ whereas the other terms do. We therefore also have

$$\omega_k\left(y^{(0:k-1)}\right) = \mathbb{E}\left[E_{k+1}\left(y^{(0:k)}\right)\middle| y^{(0:k-1)}\right]$$

$$\leq \mathbb{E}\left[\frac{s_{k+1}^2\left(y^{(0:k)}\right)}{N_{k+1}} + K_{k+1}^2\omega_{k+1}\left(y^{(0:k)}\right)\middle| y^{(0:k-1)}\right] + O(\epsilon) \tag{56}$$

and therefore by recursively substituting (56) into itself we have

$$K_k^2\omega_k\left(y^{(0:k-1)}\right) \leq \sum_{d=k+1}^{D} \frac{\left(\prod_{\ell=k}^{d-1} K_\ell^2\right)\mathbb{E}\left[s_d^2\left(y^{(0:d-1)}\right)\middle| y^{(0:k-1)}\right]}{N_d} + O(\epsilon). \tag{57}$$

Now noting that $\zeta_{d,k}^2\left(y^{(0:k-1)}\right) = \mathbb{E}\left[s_d^2\left(y^{(0:d-1)}\right)\middle| y^{(0:k-1)}\right]$, substituting (57) back into (55) gives

$$E_k\left(y^{(0:k-1)}\right) = \frac{s_k^2\left(y^{(0:k-1)}\right)}{N_k} + \sum_{d=k+1}^{D} \frac{\left(\prod_{\ell=k}^{d-1} K_\ell^2\right)\zeta_{d,k}^2\left(y^{(0:k-1)}\right)}{N_d} + O(\epsilon). \tag{58}$$

By definition we have that $\zeta_{0,0}^2 = s_0^2 = \varsigma_0^2$ and $\zeta_{d,0}^2 = \varsigma_d^2$ and as (58) holds in the case $k = 0$, the mean squared error of the overall estimator is as follows

$$\mathbb{E}\left[\left(I_0 - \gamma_0\right)^2\right] = E_0 \leq \frac{\varsigma_0^2}{N_0} + \sum_{k=1}^{D} \frac{\left(\prod_{\ell=0}^{k-1} K_\ell^2\right)\varsigma_k^2}{N_k} + O(\epsilon) \tag{59}$$

and we have the desired result for the Lipschitz case.

Continuously Differentiable Case

We now revisit the bound for the bias in the continuously differentiable case to show that a tighter overall bound can be found. We first remember that

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

Taylor's theorem implies that for any continuously differentiable f_k we can write

$$
\begin{aligned}
f_k \left(y^{(0:k)}, I_{k+1} \left(y^{(0:k)} \right) \right) &- f_k \left(y^{(0:k)}, \gamma_{k+1} \left(y^{(0:k)} \right) \right) \\
&= \frac{\partial f_k \left(y^{(0:k)}, \gamma_{k+1}(y^{(0:k)}) \right)}{\partial \gamma_{k+1}} \left(I_{k+1} \left(y^{(0:k)} \right) - \gamma_{k+1} \left(y^{(0:k)} \right) \right) \\
&+ \frac{1}{2} \frac{\partial f_k^2 \left(y^{(0:k)}, \gamma_{k+1}(y^{(0:k)}) \right)}{\partial \gamma_{k+1}^2} \left(I_{k+1} \left(y^{(0:k)} \right) - \gamma_{k+1} \left(y^{(0:k)} \right) \right)^2 \\
&+ h_3 \left(I_{k+1} \left(y^{(0:k)} \right) \right) \left(I_{k+1} \left(y^{(0:k)} \right) - \gamma_{k+1} \left(y^{(0:k)} \right) \right)^3
\end{aligned}
\tag{60}
$$

where $\lim_{x \to \gamma_{k+1}(y^{(0:k)})} h_3(x) = 0$. Consequently, the last term is $O \left(\left(I_{k+1} \left(y^{(0:k)} \right) - \gamma_{k+1} \left(y^{(0:k)} \right) \right)^3 \right)$ and so will diminish in magnitude faster than the first two terms provided that the derivatives are bounded, which is guaranteed by our assumptions. We will thus use $O(\epsilon)$ for this term and note that it is dominated in the limit.

Now defining

$$\delta_{\ell,k} = \mathbb{E} \left[\frac{\partial f_k^\ell \left(y^{(0:k)}, \gamma_{k+1}(y^{(0:k)}) \right)}{\partial \gamma_{k+1}^\ell} \left(I_{k+1} \left(y^{(0:k)} \right) - \gamma_{k+1} \left(y^{(0:k)} \right) \right)^\ell \Big| y^{(0:k-1)} \right]$$

then we have

$$\beta_k^2 \left(y^{(0:k-1)} \right) = \delta_{1,k}^2 + \frac{\delta_{2,k}^2}{4} + \delta_{1,k} \delta_{2,k} + O(\epsilon).$$

By using the tower property we further have that

$$
\begin{aligned}
\delta_{\ell,k} &= \mathbb{E} \left[\mathbb{E} \left[\frac{\partial f_k^\ell \left(y^{(0:k)}, \gamma_{k+1}(y^{(0:k)}) \right)}{\partial \gamma_{k+1}^\ell} \left(I_{k+1} \left(y^{(0:k)} \right) - \gamma_{k+1} \left(y^{(0:k)} \right) \right)^\ell \Big| y^{(0:k)} \right] \Big| y^{(0:k-1)} \right] \\
&= \mathbb{E} \left[\frac{\partial f_k^\ell \left(y^{(0:k)}, \gamma_{k+1}(y^{(0:k)}) \right)}{\partial \gamma_{k+1}^\ell} \mathbb{E} \left[\left(I_{k+1} \left(y^{(0:k)} \right) - \gamma_{k+1} \left(y^{(0:k)} \right) \right)^\ell \Big| y^{(0:k)} \right] \Big| y^{(0:k-1)} \right] \\
&\leq \mathbb{E} \left[\left| \frac{\partial f_k^\ell \left(y^{(0:k)}, \gamma_{k+1}(y^{(0:k)}) \right)}{\partial \gamma_{k+1}^\ell} \right| \left| \mathbb{E} \left[\left(I_{k+1} \left(y^{(0:k)} \right) - \gamma_{k+1} \left(y^{(0:k)} \right) \right)^\ell \Big| y^{(0:k)} \right] \right| \Big| y^{(0:k-1)} \right] \\
&\leq \left(\sup_{y^{(0)}} \left| \frac{\partial^\ell f_k \left(y^{(0:k)}, \gamma_{k+1}(y^{(0:k)}) \right)}{\partial \gamma_{k+1}^\ell} \right| \right) \mathbb{E} \left[\left| \mathbb{E} \left[\left(I_{k+1} \left(y^{(0:k)} \right) - \gamma_{k+1} \left(y^{(0:k)} \right) \right)^\ell \Big| y^{(0:k)} \right] \right| \Big| y^{(0:k-1)} \right].
\end{aligned}
$$

Now for the particular cases of $\ell = 1$ and $\ell = 2$ then the derivative terms where defined in the theorem and the expectations correspond respectively to our definitions of λ_k and ω_k giving

$$\delta_{1,k} \leq K_k \lambda_k \left(y^{(0:k-1)} \right)$$

$$\delta_{2,k} \leq C_k \omega_k \left(y^{(0:k-1)} \right)$$

and therefore

$$
\begin{aligned}
\beta_k^2 \left(y^{(0:k-1)} \right) &\leq K_k^2 \lambda_k^2 \left(y^{(0:k-1)} \right) + \frac{C_k^2}{4} \omega_k^2 \left(y^{(0:k-1)} \right) + K_k C_k \lambda_k \left(y^{(0:k-1)} \right) \omega_k \left(y^{(0:k-1)} \right) + O(\epsilon) \\
&= \left(K_k \lambda_k \left(y^{(0:k-1)} \right) + \frac{C_k}{2} \omega_k \left(y^{(0:k-1)} \right) \right)^2 + O(\epsilon).
\end{aligned}
\tag{61}
$$

Remembering (49) we can recursively define the error bound in the same manner as the Lipschitz case. We can immediately see that, as $\beta_D = 0$ without any nesting, we recover the bound from the Lipschitz case for the inner most estimator as

expected. As the innermost estimator is unbiased we also have $\lambda_{D-1}\left(y^{(0:D-2)}\right) = 0$ and so

$$\beta_{D-1}^2\left(y^{(0:D-2)}\right) \leq \frac{C_{D-1}^2}{4}\omega_{D-1}^2\left(y^{(0:D-2)}\right) + O(\epsilon)$$

$$\leq \frac{C_{D-1}^2}{4}\left(\mathbb{E}\left[\left.\frac{s_D^2\left(y^{(0:D-1)}\right)}{N_D}\right|y^{(0:D-2)}\right]\right)^2 + O(\epsilon)$$

$$= \frac{C_{D-1}^2\,\zeta_{D,D-1}^4\left(y^{(0:D-2)}\right)}{4N_D^2} + O(\epsilon).$$

Going back to our original bound on $\sigma_{D-1}^2\left(y^{(0:D-2)}\right)$ given in (47) and substituting for $\beta_{D-1}\left(y^{(0:D-2)}\right)$ we now have

$$\sigma_{D-1}\left(y^{(0:D-2)}\right) \leq \left(A_{D-1}\left(y^{(0:D-2)}\right)\right)^{\frac{1}{2}} + \left(s_{D-1}^2\left(y^{(0:D-2)}\right) + \frac{C_{D-1}^2\,\zeta_{D,D-1}^4\left(y^{(0:D-2)}\right)}{4N_D^2} + O(\epsilon)\right)^{\frac{1}{2}}. \quad (62)$$

There does not appear to be tighter bound for $A_{D-1}\left(y^{(0:D-2)}\right)$ than in the Lipschitz continuous case and so using the same bound of $A_{D-1}\left(y^{(0:D-2)}\right) \leq K_{D-1}^2\zeta_{D,D-1}^2\left(y^{(0:D-2)}\right)/N_{D-1}$ we have

$$E_{D-1}\left(y^{(0:D-2)}\right) \leq \frac{\sigma_{D-1}^2\left(y^{(0:D-2)}\right)}{N_{D-1}} + \frac{C_{D-1}^2\,\zeta_{D,D-1}^4\left(y^{(0:D-2)}\right)}{4N_D^2} + O(\epsilon)$$

$$\leq \frac{s_{D-1}^2\left(y^{(0:D-2)}\right)}{N_{D-1}} + \frac{K_{D-1}^2\zeta_{D,D-1}^2\left(y^{(0:D-2)}\right)}{N_D N_{D-1}} + \frac{C_{D-1}^2\,\zeta_{D,D-1}^4\left(y^{(0:D-2)}\right)}{4N_D^2}\left(1 + \frac{1}{N_{D-1}}\right)$$

$$+ \frac{2K_{D-1}\zeta_{D,D-1}\left(y^{(0:D-2)}\right)}{N_{D-1}\sqrt{N_D}}\left(s_{D-1}\left(y^{(0:D-2)}\right)^2 + \frac{C_{D-1}^2\,\zeta_{D,D-1}^4\left(y^{(0:D-2)}\right)}{4N_D^2}\right)^{\frac{1}{2}} + O(\epsilon). \quad (63)$$

Therefore for the single nesting case, we now have

$$E_0 \leq \frac{\varsigma_0^2}{N_0} + \frac{K_0^2\varsigma_1^2}{N_0 N_1} + \frac{2K_0\varsigma_1}{N_0\sqrt{N_1}}\sqrt{\varsigma_0^2 + \frac{C_0^2\varsigma_1^4}{4N_1^2}} + \frac{C_0^2\varsigma_1^4}{4N_1^2}\left(1 + \frac{1}{N_0}\right) + O\left(\frac{1}{N_1^3}\right) \quad (64)$$

$\approx \frac{\varsigma_0^2}{N_0} + \frac{C_0^2\varsigma_1^4}{4N_1^2} = O\left(\frac{1}{N_0} + \frac{1}{N_1^2}\right)$ where again the approximation becomes tight when $N_0, N_1 \to \infty$. Here we have used the fact that the only $O(\epsilon)$ term comes from the Taylor expansion and is equal to $O\left(\frac{1}{N_1^3}\right)$ because we have $\delta_{1,D-1} = 0$ and therefore

$$O(\epsilon) = O\left(\delta_{2,D-1}\delta_{3,D-1} + \delta_{2,D-1}\delta_{4,D-1}\right)$$

$$= O\left(\delta_{2,D-1}\mathbb{E}\left[\left.\left(I_1\left(y^{(0)}\right) - \gamma_1\left(y^{(0)}\right)\right)^3\right|y^{(0)}\right]\right) + O\left(\delta_{2,D-1}\mathbb{E}\left[\left.\left(I_1\left(y^{(0)}\right) - \gamma_1\left(y^{(0)}\right)\right)^4\right|y^{(0)}\right]\right)$$

$$= O\left(\frac{1}{N_1}\mathbb{E}\left[\left.\left(\frac{1}{N_1}\sum_{n=1}^{N_1}f_1\left(y_n^{(0:1)}\right) - \mathbb{E}\left[\left.f_1\left(y^{(0:1)}\right)\right|y^{(0)}\right]\right)^3\right|y^{(0)}\right]\right)$$

$$+ O\left(\frac{1}{N_1}\mathbb{E}\left[\left.\left(\frac{1}{N_1}\sum_{n=1}^{N_1}f_1\left(y_n^{(0:1)}\right) - \mathbb{E}\left[\left.f_1\left(y^{(0:1)}\right)\right|y^{(0)}\right]\right)^4\right|y^{(0)}\right]\right)$$

now noting that the $y_n^{(0:1)}$ are i.i.d., and that $\mathbb{E}\left[\left.f_1\left(y_1^{(0:1)}\right) - \mathbb{E}\left[f_1\left(y^{(0:1)}\right)|y^{(0)}\right]\right|y^{(0)}\right] = 0$ such many of the cross terms when expanding the brackets are zero, we have

$$= O\left(\frac{1}{N_1^4}\sum_{n=1}^{N_1}\mathbb{E}\left[\left.\left(f_1\left(y_1^{(0:1)}\right) - \mathbb{E}\left[\left.f_1\left(y^{(0:1)}\right)\right|y^{(0)}\right]\right)^3\right|y^{(0)}\right]\right)$$

$$+ O\left(\frac{1}{N_1^5}\sum_{n=1}^{N_1}\mathbb{E}\left[\left.\left(f_1\left(y_1^{(0:1)}\right) - \mathbb{E}\left[\left.f_1\left(y^{(0:1)}\right)\right|y^{(0)}\right]\right)^4\right|y^{(0)}\right]\right)$$

$$+ O\left(\frac{3}{N_1^5}\sum_{n=1}^{N_1}\sum_{m=1,m\neq n}^{N_1}\left(\mathbb{E}\left[\left.\left(f_1\left(y_1^{(0:1)}\right) - \mathbb{E}\left[\left.f_1\left(y^{(0:1)}\right)\right|y^{(0)}\right]\right)^2\right|y^{(0)}\right]\right)^2\right)$$

$$= O\left(\frac{1}{N_1^3}\right) + O\left(\frac{1}{N_1^4}\right) + O\left(\frac{1}{N_1^3}\right) = O\left(\frac{1}{N_1^3}\right)$$

as required.

Returning to calculating the bound for the repeated nesting case then by substituting (61) into (49) we have more generally

$$E_k\left(y^{(0:k-1)}\right) \le \frac{s_k^2\left(y^{(0:k-1)}\right)}{N_k} + \left(K_k \lambda_k\left(y^{(0:k-1)}\right) + \frac{C_k}{2}\omega_k\left(y^{(0:k-1)}\right)\right)^2 + O(\epsilon). \tag{65}$$

Now remembering that $\omega_k\left(y^{(0:k-1)}\right) = \mathbb{E}\left[E_{k+1}\left(y^{(0:k)}\right)\big|y^{(0:k-1)}\right]$ from (49) we have

$$\omega_k\left(y^{(0:k-1)}\right) = \mathbb{E}\left[\frac{s_{k+1}^2\left(y^{(0:k)}\right)}{N_{k+1}} + \beta_{k+1}^2\left(y^{(0:k)}\right)\bigg|y^{(0:k-1)}\right] + O(\epsilon)$$

$$= \frac{\zeta_{k+1,k}^2}{N_{k+1}} + \mathbb{E}\left[\beta_{k+1}^2\left(y^{(0:k)}\right)\big|y^{(0:k-1)}\right] + O(\epsilon). \tag{66}$$

We also have that except at $k = D - 1$ and $k = D$ (for which both λ_k and β_{k+1} are zero), then

$$\lambda_k\left(y^{(0:k-1)}\right) = \mathbb{E}\left[\left|\beta_{k+1}\left(y^{(0:k)}\right)\right|\bigg|y^{(0:k)}\right] \gg \mathbb{E}\left[\beta_{k+1}^2\left(y^{(0:k)}\right)\big|y^{(0:k-1)}\right]$$

for sufficiently large $N_{k+1}, \ldots, N_D$. This means that when we substitute (66) into (65), the second term in (66) becomes dominated giving

$$E_k\left(y^{(0:k-1)}\right) \le \frac{s_k^2\left(y^{(0:k-1)}\right)}{N_k} + \left(K_k \lambda_k\left(y^{(0:k-1)}\right) + \frac{C_k \zeta_{k+1,k}^2}{2N_{k+1}}\right)^2 + O(\epsilon). \tag{67}$$

Now as $\beta_{k+1}^2\left(y^{(0:k)}\right) = E_{k+1}\left(y^{(0:k)}\right) - \frac{s_{k+1}^2\left(y^{(0:k)}\right)}{N_{k+1}}$ we have

$$\lambda_k\left(y^{(0:k-1)}\right) = \mathbb{E}\left[\sqrt{E_{k+1}\left(y^{(0:k)}\right) - \frac{s_{k+1}^2\left(y^{(0:k)}\right)}{N_{k+1}}}\bigg|y^{(0:k-1)}\right] + O(\epsilon)$$

and substituting in (67) gives

$$\lambda_k\left(y^{(0:k-1)}\right) \le \mathbb{E}\left[K_{k+1}\lambda_{k+1}\left(y^{(0:k)}\right) + \frac{C_{k+1}\zeta_{k+2,k+1}^2}{2N_{k+2}}\bigg|y^{(0:k-1)}\right] + O(\epsilon)$$

$$= \frac{C_{k+1}\zeta_{k+2,k}^2}{2N_{k+2}} + K_{k+1}\mathbb{E}\left[\lambda_{k+1}\left(y^{(0:k)}\right)\big|y^{(0:k-1)}\right] + O(\epsilon)$$

$$\le \frac{C_{k+1}\zeta_{k+2,k}^2}{2N_{k+2}} + \sum_{d=k+1}^{D-2}\mathbb{E}\left[\left(\prod_{\ell=k+1}^{d}K_\ell\right)\frac{C_{d+1}\zeta_{d+2,d}^2}{2N_{d+2}}\bigg|y^{(0:k-1)}\right] + O(\epsilon)$$

$$\le \frac{C_{k+1}\zeta_{k+2,k}^2}{2N_{k+2}} + \sum_{d=k+1}^{D-2}\left(\prod_{\ell=k+1}^{d}K_\ell\right)\frac{C_{d+1}\zeta_{d+2,k}^2}{2N_{d+2}} + O(\epsilon)$$

and thus

$$E_k\left(y^{(0:k-1)}\right) \le \frac{s_k^2\left(y^{(0:k-1)}\right)}{N_k} + \frac{1}{4}\left(\frac{C_k \zeta_{k+1,k}^2}{N_{k+1}} + \sum_{d=k}^{D-2}\left(\prod_{\ell=k}^{d}K_\ell\right)\frac{C_{d+1}\zeta_{d+2,k}^2}{N_{d+2}}\right)^2 + O(\epsilon).$$

and therefore

$$\mathbb{E}\left[(I_0 - \gamma_0)^2\right] = E_0 \le \frac{\varsigma_0^2}{N_0} + \frac{1}{4}\left(\frac{C_0 \varsigma_1^2}{N_1} + \sum_{k=0}^{D-2}\left(\prod_{d=0}^{k}K_d\right)\frac{C_{k+1}\varsigma_{k+2}^2}{N_{k+2}}\right)^2 + O(\epsilon)$$

as required and we are done.

$\square$

Appendix E   Proof of Theorem 4 - Convergence Rate for Finite Realisations of y

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. Denote $P_c = P(y = y_c)$ and $f_c = f(y_c, \gamma(y_c))$ noting that as the y_c are fixed values, so are P_c and f_c. Then, Minkowski's inequality allows us to bound the mean squared error as

$$\mathbb{E}\left[(I_N - I)^2\right] = \|I_N - I\|_2^2 \le \left(\sum_{c=1}^{C} W_c\right)^2 \quad \text{where} \quad W_c := \left\|(\hat{P}_N)_c\, (\hat{f}_N)_c - P_c\, f_c\right\|_2.$$

Moreover, again by Minkowski, we have $W_c \le U_c + V_c$ where

$$U_c = \left\|(\hat{P}_N)_c\, (\hat{f}_N)_c - (\hat{P}_N)_c\, f_c\right\|_2, \quad V_c = \left\|(\hat{P}_N)_c\, f_c - P_c\, f_c\right\|_2.$$

Factoring out $(\hat{P}_N)_c$ in U_c and noting that each y_n and $z_{n,c}$ are sampled independently gives

$$U_c = \sqrt{\mathbb{E}\left[(\hat{P}_N)_c^2\left((\hat{f}_N)_c - f_c\right)^2\right]} = \sqrt{\mathbb{E}\left[(\hat{P}_N)_c^2\right]}\sqrt{\mathbb{E}\left[\left((\hat{f}_N)_c - f_c\right)^2\right]}.$$

Using Minkowski's inequality, we may write the first right-hand term as

$$\sqrt{\mathbb{E}\left[(\hat{P}_N)_c^2\right]} = \left\|(\hat{P}_N)_c\right\|_2 \le \frac{1}{N}\sum_{n=1}^{N}\|\mathbb{1}(y_n = y_c)\|_2 = \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\left[\mathbb{1}(y_n = y_c)^2\right] = \frac{1}{N}\sum_{n=1}^{N}P_c = P_c.$$

For the second term, note that by Lipschitz continuity, we have for some constant $K > 0$

$$\sqrt{\mathbb{E}\left[\left((\hat{f}_N)_c - f_c\right)^2\right]} = \left\|(\hat{f}_N)_c - f_c\right\|_2 \le K\left\|\frac{1}{N}\sum_{n=1}^{N}\phi(y_c, z_{n,c}) - \gamma(y_c)\right\|_2 = K \cdot O(1/\sqrt{N}) = O(1/\sqrt{N}),$$

since $\frac{1}{N}\sum_{n=1}^{N}\phi(y_c, z_{n,c})$ is a Monte Carlo estimator for $\gamma(y_c)$. Altogether then, we have that

$$U_c = P_c \cdot O(1/\sqrt{N}) = O(1/\sqrt{N}).$$

We can also factor out f_c in V_c to obtain

$$V_c = |f_c| \cdot \left\|(\hat{P}_N)_c - P_c\right\|_2 = |f_c| \cdot O(1/\sqrt{N}) = O(1/\sqrt{N}),$$

since $(\hat{P}_N)_c$ is a Monte Carlo estimator for P_c. Now by noting that $(A+B)^2 \le 2(A^2 + B^2)$ for any $A, B \in \mathbb{R}$, an inductive argument shows that

$$\left(\sum_{\ell=1}^{L} A_\ell\right)^2 \le 2^{\lceil \log_2 L \rceil}\sum_{\ell=1}^{L} A_\ell^2$$

for all $A_1, \cdots, A_L \in \mathbb{R}$. We can now show that our asymptotic bounds for U_c and V_c entail that our overall mean squared error satisfies

$$
\begin{aligned}
\mathbb{E}\left[(I_N - I)^2\right] &\le 2^{\lceil \log_2 C \rceil}\sum_{c=1}^{C} W_c^2 \le 2^{\lceil \log_2 C \rceil}\sum_{c=1}^{C}(U_c + V_c)^2 \le 2^{\lceil \log_2 C \rceil + 1}\sum_{c=1}^{C} U_c^2 + V_c^2 \\
&= 2^{\lceil \log_2 C \rceil + 1}\sum_{c=1}^{C} O(1/N) + O(1/N) = O(1/N),
\end{aligned}
$$

as desired. $\qquad\square$

Appendix F   Proof for Theorem 5 - Products of Expectations

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. Consider fixed sizes of nested sample sets, $\{M_\ell\}_{\ell=1:L}$. For each $y \in \mathcal{Y}$ and
$$x = \{\{z'_{\ell,m}\}_{m=1:M_\ell}\}_{\ell=1:L} \in \mathcal{X} = \mathcal{Z}_1^{M_1} \otimes \cdots \otimes \mathcal{Z}_L^{M_L},$$
define
$$\eta(y, x) = f\left(y, \prod_{\ell=1}^{L} \frac{1}{M_\ell} \sum_{m_\ell=1}^{M_\ell} \psi_\ell(y, z'_{\ell,m})\right).$$

Now, $I_N = \frac{1}{N}\sum_{n=1}^{N} \eta(y_n, x_n)$ is a standard MC estimator on the space $\mathcal{Y} \otimes \mathcal{X}$. Thus, $I_N \overset{a.s.}{\to} \mathbb{E}[I_N]$ with convergence properties and rate as per standard MC. We finish the proof by showing that $\mathbb{E}[I_N] = I$ when f is linear:

$$\mathbb{E}[I_N] = \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N} f\left(y_n, \prod_{\ell=1}^{L}\frac{1}{M_\ell}\sum_{m=1}^{M_\ell}\psi_\ell(y_n, z'_{n,\ell,m})\right)\right] = \mathbb{E}\left[\mathbb{E}\left[f\left(y_1, \prod_{\ell=1}^{L}\frac{1}{M_\ell}\sum_{m=1}^{M_\ell}\psi_\ell(y_1, z'_{1,\ell,m})\right)\Bigg|y_1\right]\right],$$

now using the linearity of f

$$= \mathbb{E}\left[f\left(y_1, \mathbb{E}\left[\prod_{\ell=1}^{L}\frac{1}{M_\ell}\sum_{m=1}^{M_\ell}\psi_\ell(y_1, z'_{1,\ell,m})\Bigg|y_1\right]\right)\right],$$

and using the fact that terms for different ℓ are by construction independent

$$= \mathbb{E}\left[f\left(y_1, \prod_{\ell=1}^{L}\mathbb{E}\left[\frac{1}{M_\ell}\sum_{m=1}^{M_\ell}\psi_\ell(y_1, z'_{1,\ell,m})\Bigg|y_1\right]\right)\right] = \mathbb{E}\left[f\left(y_1, \prod_{\ell=1}^{L}\mathbb{E}\left[\psi_\ell(y_1, z'_{1,\ell,1})|y_1\right]\right)\right] = I,$$

as required. $\qquad\qquad\square$

Appendix G   Optimizing the Convergence Rates

We have shown that the mean squared error converges at a rate
$$O\left(\sum_{k=0}^{D}\frac{1}{N_k}\right) \quad \text{or} \quad O\left(\frac{1}{N_0} + \left(\sum_{k=1}^{D}\frac{1}{N_k}\right)^2\right)$$

depending on the smoothness assumptions that can be made about f. Here we show that given a sample budget for the inner most estimator $T = \prod_{k=0}^{D} N_k$, then these bounds are optimized by setting $N_0 \propto N_1 \propto \cdots \propto N_D$ and $N_0 \propto N_1^2 \propto \cdots \propto N_D^2$ respectively for the two cases and that this gives bounds of $O\left(1/T^{\frac{1}{D+1}}\right)$ and $O\left(1/T^{\frac{2}{D+2}}\right)$ respectively. For the single nested case, this gives bounds of $O(1/\sqrt{T})$ and $O(1/T^{2/3})$ respectively.

We start by explaining why T is an appropriate measure of the overall computational cost. First note that for each sample of $y^{(0:k)}$, the NMC estimator requires N_k samples of $y^{(k+1)}$. Thus there are N_0 samples of the outermost level, $N_0 \times N_1$ of the next level, and $T = \prod_{k=0}^{D} N_k$ samples of the innermost level, regardless of the setup. In other words, each individual estimate of the innermost level uses N_D samples and we generate $\prod_{k=0}^{D-1} N_k = T/N_D$ of these estimates because we need to generate one estimate for each sample of the layer above. Thus what we can vary for a fixed T is whether we use more estimates each using fewer samples, or fewer estimates each using more samples.

Now the total cost of generating I_0 scales with sum the costs of each individual layer, namely
$$\text{Cost} = \sum_{k=0}^{D} c_k \prod_{\ell=0}^{k} N_\ell$$

where c_k is the per sample cost local computations made at the kth layer (i.e. sampling $y^{(0:k)}$ and evaluating f_k for given inputs), which is independent of the N_k. For large N_D, we see that the dominant cost is that of the inner most layer, namely $c_T \prod_{\ell=0}^{D} N_\ell = c_T T$, and we asymptotically spend 100% of our time dealing with the innermost estimator. To give intuition to this, think about writing the estimator as a hierarchy of nested for loops; as the length of the loops increases we spend an increasing proportion of our time inside the innermost loop. Consequently, in the asymptotic regime, our computational cost is $O(T)$ and we can use T is an appropriate measure of the overall computational cost.

To derive the optimal rates, we first consider the single nested case where $D = 1$, $N_0 = N$, and $N_1 = M$. Consider setting

$N = \tau(M)$ then $T = \tau(M) \cdot M$ and our bounds become $O(R)$, where

$$R = 1/\tau(M) + 1/M \quad \text{and} \quad R = 1/\tau(M) + 1/M^2.$$

for the two cases respectively.

In this first case supposing $\tau(M) = O(M)$ easily gives

$$T = M\tau(M) = O\left(M^2\right)$$

and as such

$$R = O\left(\frac{1}{M}\right) = O\left(\frac{1}{\sqrt{T}}\right) \tag{68}$$

as $M \to \infty$. In contrast, consider the case that $\tau(M) \gg M$ as $M \to \infty$. We then have $\frac{1}{\sqrt{M}} \gg \frac{1}{\sqrt{\tau(M)}}$ as $M \to \infty$, so that

$$R = O\left(\frac{1}{M}\right) \gg \frac{1}{\sqrt{M}}\frac{1}{\sqrt{\tau(M)}} = \frac{1}{\sqrt{T}}.$$

Comparing with (68), we observe that, for the same total budget of samples T, this choice of τ provides a strictly weaker convergence guarantee than in the previous case. When $M \gg \tau(M)$ also then we have $\frac{1}{\sqrt{\tau(M)}} \gg \frac{1}{\sqrt{M}}$ as $M \to \infty$ and so

$$R = O\left(\frac{1}{\tau(M)}\right) \gg \frac{1}{\sqrt{M}}\frac{1}{\sqrt{\tau(M)}} = \frac{1}{\sqrt{T}}$$

which is again a weaker bound. We thus see that the $O(1/N + 1/M)$ bound is optimized when $N \propto M$, giving a convergence rate of $O(1/\sqrt{T})$.

In the second case suppose that $\tau(M) = O(M^2)$ as $M \to \infty$. This now gives

$$T = M\tau(M) = O\left(M^3\right)$$

and therefore

$$R = O\left(\frac{1}{M^2}\right) = O\left(\frac{1}{T^{2/3}}\right)$$

as $M \to \infty$. Now considering the cases $\tau(M) \gg M^2$ leads to $\frac{1}{M^{4/3}} \gg \frac{1}{\tau(M)^{2/3}}$ and thus

$$R = O\left(\frac{1}{M^2}\right) \gg \frac{1}{M^{2/3}}\frac{1}{\tau(M)^{2/3}} = \frac{1}{T^{2/3}}.$$

Similarly, if $\tau(M) \ll M^2$ then $\frac{1}{\tau(M)^{1/3}} \gg \frac{1}{M^{2/3}}$ and thus

$$R = O\left(\frac{1}{\tau(M)}\right) \gg \frac{1}{M^{2/3}}\frac{1}{\tau(M)^{2/3}} = \frac{1}{T^{2/3}}.$$

Both of these cases lead to weaker bounds and so we see that the $O(1/N + 1/M^2)$ bound is tightest when $N \propto M^2$, giving a convergence rate of $O(1/T^{2/3})$.

We now consider the repeated nesting case without continuously differentiability such that our bound is $O\left(\sum_{k=0}^{D}\frac{1}{N_k}\right)$. Here we can immediately see that $N_0 \propto N_1 \propto \cdots \propto N_D$ leads to $N_k \propto T^{\frac{1}{D+1}}$ and thus $O\left(1/T^{\frac{1}{D+1}}\right)$ convergence. If we were to set any $N_k \ll T^{\frac{1}{D+1}}$ then this term would dominate the sum and lead to a worse converge. Thus the result from the single nested case trivially extends to the multiple nested case, giving the required result.

Finally considering repeated nesting for the bound $O\left(\frac{1}{N_0} + \left(\sum_{k=1}^{D}\frac{1}{N_k}\right)^2\right)$ then we have from the previous result that $N_1 \propto N_2 \propto \cdots \propto N_D$ is required for optimality as otherwise one of the terms in the summation dominates the other terms. If we now define $M = \prod_{k=1}^{D} N_k = T/N_0$ then we get a convergence rate of $O(1/N_0 + 1/M^2)$ which is identical to the single nesting case for this tighter bound. We, therefore, have that the optimal configuration must be $N_0 \propto N_1^2 \propto \cdots \propto N_D^2$ giving a bound of $O\left(1/T^{\frac{2}{D+2}}\right)$ as it gives $N_0 \propto T^{\frac{2}{D+2}}$.

Appendix H   Additional details pertaining to cancer simulator

In this section, we elucidate some more details about the cancer simulator described in the manuscript, provide more rigorous mathematical definitions for the relevant terms using the same nomenclature, and also include more results figures.
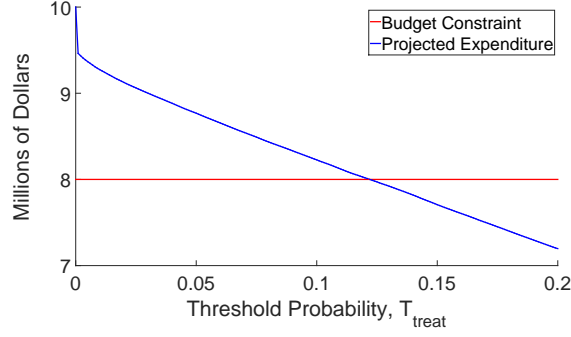
Figure 7. Projected expenditure (proportional to $I_{N,M}$) evaluated at different values of T_{treat}. The budget constraint is shown by the horizontal red line. The optimal value of T_{treat} is found by the intersection and occurs at $T_{\text{treat}} = 12.5\%$. Evaluated was carried out at 100 $T_{\text{treat}}1$. Only the bottom 20% is pictured as this is the operating range for most treatment centers.

H.1   Simulator details

We define $I(T_{\text{treat}})$ to be the expected proportion of patients who receive treatment. A particular patient is represented by $y \in \mathcal{R}^d$. Specifically, y consists of only a single real number ($d = 1$) representing the size of the tumor upon discovery. Initial tumor size is drawn from a scaled Rayleigh distribution. The outcome of the simulator is then $\phi(y, z) \in \{0, 1\}$, and is the binary outcome of whether that particular patient and sample of unobserved parameters yield an expected tumor size below the threshold, T_{opp}, after a fixed time duration, t_{max}. The simulator is a pair of coupled, parameterized differential equations for the action of an anti-tumor treatment such as chemotherapy, as described in Enderling and Chaplain (2014):

$$\frac{dc}{dt} = -\lambda c \log \left(\frac{c}{K} \right) - \xi c \tag{69}$$

$$\frac{dK}{dt} = \phi c - \psi K c^{2/3}, \tag{70}$$

where $c(t, x) \in \mathcal{R}_+$ represents tumor size, with initial size y_n. Similarly, $K(t, x) \in \mathcal{R}_+$ represents the notion of a carrying capacity, with the initial carrying capacity, $K(0, z)$, set to a known constant K_0. The magnitude of the patient response to an anti-tumor treatment (such as chemotherapy) is represented by $\xi \in [0, 1]$, drawn from a beta distribution. $\{\lambda, \psi, \phi\} \in \mathcal{R}_+^3$ represent the parameters of the simulator. We also define $x_{n,m} = \{\lambda, \psi, \phi K_0, \xi\}$ and $z_{n,m} = \{x_{n,m}, T_{opp}, t_{max}\}$, where all but ξ are set to constant values. Expanding this to condition all values on y_n is trivial given domain knowledge. Alternatively, they could also be drawn at random, but not be conditioned on y_n. Such relations are omitted here for simplicity.

We can now fully define ϕ as:

$$\phi(y_n, z_{n,m}) = \mathbb{1}(c(t_{\max}, x_{n,m}) < T_{\text{opp}}). \tag{71}$$

Taking the expectation of ϕ over M different realizations of z yields the estimate $(\hat{\gamma}_M)_n$. This value is the probability that treatment will be successful for a particular patient, marginalizing over possible unobserved dynamics. This is the point at which clinician decides whether initiate the treatment plan. This decision is represented $f(y_n, (\hat{\gamma}_M)_n) \in [0, 1]$ as:

$$f(y_n, (\hat{\gamma}_M)_n) = \mathbb{1}((\hat{\gamma}_M)_n > T_{\text{treat}}) \tag{72}$$

where T_{treat} is the minimum probability of success required for that patient to receive the treatment, and again, could be conditioned on y also. Taking the expectation of f over patients yields the expected frequency with which the treatment will be delivered, given a value of T_{treat}. The hospital wishes to estimate the value T_{treat} that maximizes the number of patients treated, while only treating those patients with the highest probability of success, and (in expectation) staying within the budgetary constraint.

The model is completed by the definition of the following distributions and parameters.

$$K_0 = 100000000, \quad \phi = 0.001, \quad \psi = 0.05, \quad \lambda = 0.5, \quad \xi \sim \text{Beta}(5, 2),$$

$$c_0 \sim 1000 * \text{Rayleigh}(10), \quad T_{\text{opp}} = 2000, \quad T_{\text{treat}} = 0.35, \quad t_{\max} = 250, \quad t_{\text{step}} = 0.01$$

H.2   Budget result

In the example outlined above, the treatment center is not actually attempting to evaluate the value of I, but to find the optimal value of T_{treat} subject to a budgetary constraint. A simplistic way of evaluating the optimal value is to perform a dense search over different values of the parameter, each time evaluating the estimated expenditure, and select the best performing value.

Figure 7 shows the variation of predicted expenditure against the threshold probability, as well as the budget constraint. The intersection of these curves is the optimal setting of T_{opp}, here evaluated to be 12.5%. From the blue line, it is clear that the relationship between expenditure and treatment probability is non-linear, especially at the extrema of the distribution, and hence the use of NMC was necessarily for evaluating the optimal value.

Appendix I   Bayesian Experimental Design

Bayesian experimental design provides a framework for designing experiments in a manner that is optimal from an information-theoretic viewpoint (Chaloner and Verdinelli, 1995; Sebastiani and Wynn, 2000). By minimizing the entropy in the posterior distribution of the parameters of interest, one can maximize the information gathered by the experiment.

Let the parameters of interest be denoted by $\theta \in \Theta$ for which we define a prior distribution $p(\theta)$. Let the probability of achieving outcome $y \in \mathcal{Y}$, given parameters θ and a design $d \in \mathcal{D}$, be defined by likelihood model $p(y|\theta, d)$. Under our model, the outcome of the experiment given a chosen d is distributed according to

$$p(y|d) = \int_{\Theta} p(y, \theta|d)d\theta = \int_{\Theta} p(y|\theta, d)p(\theta)d\theta. \tag{73}$$

where we have used the fact that $p(\theta) = p(\theta|d)$ because θ is independent of the design. Our aim is to choose the optimal design d under some criterion. We, therefore, define a utility function, $U(y, d)$, representing the utility of choosing a design d and getting a response y. Typically our aim is to maximize information gathered from the experiment, and so we set $U(y, d)$ to be the gain in Shannon information between the prior and the posterior:

$$U(y, d) = \int_{\Theta} p(\theta|y, d) \log(p(\theta|y, d))d\theta - \int_{\Theta} p(\theta) \log(p(\theta))d\theta \tag{74}$$

However, we are still uncertain about the outcome. Thus, we use the expectation of $U(y, d)$ with respect to $p(y|d)$ as our target:

$$\begin{aligned}
\bar{U}(d) &= \int_{\mathcal{Y}} U(y, d)p(y|d)dy \\
&= \int_{\mathcal{Y}} \int_{\Theta} p(y, \theta|d) \log(p(\theta|y, d))d\theta dy - \int_{\Theta} p(\theta) \log(p(\theta))d\theta \\
&= \int_{\mathcal{Y}} \int_{\Theta} p(y, \theta|d) \log\left(\frac{p(\theta|y, d)}{p(\theta)}\right)d\theta dy.
\end{aligned} \tag{75}$$

noting that this corresponds to the mutual information between the parameters θ and the observations y. The Bayesian-optimal design is then given by

$$d^* = \underset{d \in \mathcal{D}}{\operatorname{argmax}} \bar{U}(d). \tag{76}$$

Finding d^* is challenging because the posterior $p(\theta|y, d)$ is rarely known in closed form. To solve the problem, we proceed by rearranging (75) using Bayes' rule (remembering that $p(\theta) = p(\theta|d)$):

$$\begin{aligned}
\bar{U}(d) &= \int_{\mathcal{Y}} \int_{\Theta} p(y, \theta|d) \log\left(\frac{p(\theta|y, d)}{p(\theta)}\right)d\theta dy \\
&= \int_{\mathcal{Y}} \int_{\Theta} p(y, \theta|d) \log\left(\frac{p(y|\theta, d)}{p(y|d)}\right)d\theta dy \\
&= \int_{\mathcal{Y}} \int_{\Theta} p(y, \theta|d) \log(p(y|\theta, d))d\theta dy - \int_{\mathcal{Y}} p(y|d) \log(p(y|d))dy.
\end{aligned} \tag{77}$$

The first of these terms can now be evaluated using standard MC approaches as the integrand is analytic. In contrast, the second term is not directly amenable to standard MC estimation as the marginal $p(y|d)$ represents an expectation and taking its logarithm represents a non-linear functional mapping.

To derive an estimator, we will now consider these terms separately. Starting with the first term,

$$\bar{U}_1(d) = \int_{\mathcal{Y}} \int_{\Theta} p(y, \theta | d) \log(p(y|\theta, d)) d\theta dy \approx \frac{1}{N} \sum_{n=1}^{N} \log(p(y_n|\theta_n, d)) \tag{78}$$

where $\theta_n \sim p(\theta)$ and $y_n \sim p(y|\theta = \theta_n, d)$. We note that evaluating (78) involves both sampling from $p(y|\theta, d)$ and directly evaluating it point-wise. The latter of these cannot be avoided, but in the scenario where we do not have direct access to a sampler for $p(y|\theta, d)$, we can use the standard importance sampling trick, sampling instead $y_n \sim q(y|\theta = \theta_n, d)$ and weighting the samples in (78) by $w_n = \frac{p(y_n|\theta_n, d)}{q(y_n|\theta_n, d)}$.

Now considering the second term we have

$$\bar{U}_2(d) = \int_{\mathcal{Y}} p(y|d) \log(p(y|d)) dy \approx \frac{1}{N} \sum_{n=1}^{N} \log\left(\frac{1}{M} \sum_{m=1}^{M} p(y_n|\theta_{n,m}, d)\right) \tag{79}$$

where $\theta_{n,m} \sim p(\theta)$ and $y_n \sim p(y|d)$. Here we can sample the latter by first sampling an otherwise unused $\theta_{n,0} \sim p(\theta)$ and then sampling $y_n \sim p(y|\theta_{n,0}, d)$. Again we can use importance sampling if we do not have direct access to a sampler for $p(y|\theta_{n,0}, d)$.

Putting (78) and (79) together (and renaming θ_n from (78) as $\theta_{n,0}$ for notational consistency with (79)) we now have the following complete estimator given in the main paper and implicitly used by (Myung et al., 2013) amongst others

$$\bar{U}(d) \approx \frac{1}{N} \sum_{n=1}^{N} \left[\log(p(y_n|\theta_{n,0}, d)) - \log\left(\frac{1}{M} \sum_{m=1}^{M} p(y_n|\theta_{n,m}, d)\right) \right] \tag{80}$$

where $\theta_{n,m} \sim p(\theta) \ \forall m \in 0 : M, \ n \in 1 : N$ and $y_n \sim p(y|\theta = \theta_{n,0}, d) \ \forall n \in 1 : N$.

We now show that if y can only take on one of C possible values $(y_1, \ldots, y_C)$, we can achieve significant improvements in the convergence rate by using a similar to that introduced in Section 3.2 to convert to single MC estimator:

$$\bar{U}(d) = \int_{\mathcal{Y}} \int_{\Theta} p(y, \theta | d) \log(p(y|\theta, d)) d\theta dy - \int_{\mathcal{Y}} p(y|d) \log(p(y|d)) dy$$

$$= \int_{\Theta} \left[\sum_{c=1}^{C} p(\theta) p(y_c|\theta, d) \log(p(y_c|\theta, d)) \right] d\theta - \sum_{c=1}^{C} p(y_c|d) \log(p(y_c|d))$$

$$\approx \frac{1}{N} \sum_{n=1}^{N} \sum_{c=1}^{C} p(y_c|\theta_n, d) \log(p(y_c|\theta_n, d)) - \sum_{c=1}^{C} \left[\left(\frac{1}{N} \sum_{n=1}^{N} p(y_c|\theta_n, d)\right) \log\left(\frac{1}{N} \sum_{n=1}^{N} p(y_c|\theta_n, d)\right) \right] \tag{81}$$

where $\theta_n \sim p(\theta) \ \forall n \in 1, \ldots, N$. As C is a fixed constant, the MSE for first term clearly converges at the standard MC error rate of $O(1/N)$. Similarly each $\hat{P}_N(y_c|d) = \frac{1}{N} \sum_{n=1}^{N} p(y_c|\theta_n, d)$ term also converges at a rate $O(1/N)$ to $p(y_c|d)$. Now noting that $\hat{P}_N(y_c|d) \leq 1$ and that $f(x) = x \log x$ is Lipschitz continuous in the range $(0, 1]$, each $\hat{P}_N(y_c|d) \log\left(\hat{P}_N(y_c|d)\right)$ term must also converge at the MC error rate if $p(y_c|d) > 0 \ \forall c = 1, \ldots, C$. Finally if we assume that when $p(y_c|d) = 0$ then $\hat{P}_N(y_c|d) = 0$ almost surely for sufficiently large N, then the second term also converges at the MC error when $p(y_c|d) = 0$. We now have a finite sum of terms which each convergence to $\bar{U}(d)$ with MC MSE rate $O(1/N)$, and so the overall estimator (81) must also converge at this rate. This compares to $O(1/T^{2/3})$ for (80) (assuming we take $N \propto M^2$), noting that generating T samples for (80) has the same cost up to a constant factor as generating N for (81). To the best of our knowledge, this is the first introduction of this superior estimator in the literature.

We finish by showing that the theoretical advantages of this reformulation also leads to empirical gains in the estimation of $\bar{U}(d)$. For this, we consider a model used in psychology experiments for delay discounting introduced by (Vincent, 2016; Vincent and Rainforth, 2017). Our experiment comprises of asking questions of the form *"Would you prefer £A now, or £B in D days?"* and we wish to choose the question variables $d = \{A, B, D\}$ in the manner that will give the most incisive questions. The target participant is presumed to have parameters $\theta = \{k, \alpha\}$ and the following response model

$$y \sim \text{Bernoulli}\left(0.01 + 0.98 \cdot \Phi\left(\frac{1}{\alpha}\left(\frac{B}{1 + e^k D} - A\right)\right)\right) \tag{82}$$

where $y = 1$ indicates choosing the delayed response and Φ represents the cumulative normal distribution. As more questions are asked, the distribution over the parameters θ is updated, such that the most optimal question to ask at a particular time depends on the previous questions and responses. For the sake of brevity, when comparing the performance
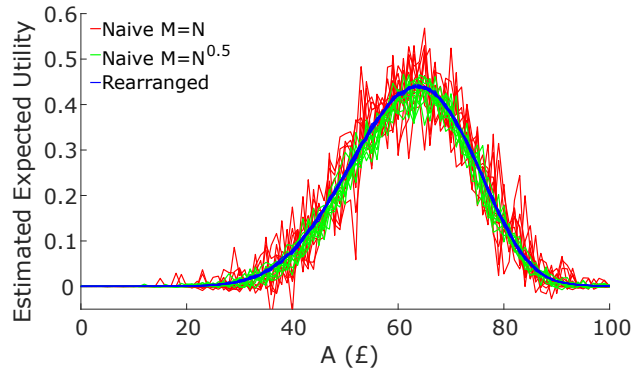
Figure 8. Estimated expected utilities $\bar{U}(d)$ for different values of one of the design parameters $A \in \{1, 2, \ldots, 100\}$ given a fixed total sample budget of $T = 10^4$. Here the lines correspond to 10 independent runs, showing that the variance of (80) is far higher than (81).

of (80) and (81) we will neglect the problem of how best to optimize the design, and consider only the problem of evaluating $\bar{U}(d)$. We will further consider the case where $B = 100$ and $D = 50$ are fixed and we are only choosing the delayed value A. We presume the following distribution on the parameters

$$k \sim \mathcal{N}(-4.5, 0.5^2)$$
$$\alpha \sim \Gamma(2, 2).$$

We first consider convergence in the estimate of $\bar{U}(d)$ for the case $A = 70$ for our suggested method (81) and the naïve solution (80), the results of which are shown in Figure 2a in the main paper. Here we see that the convergence rates of the two methods are both as expected and that our suggested method offers significant empirical performance improvements.

We next consider setting a total sample budget $T = 10^4$ and look at the variation in the estimated values of $\bar{U}(d)$ for different values of A for the two methods as shown in Figure 8. This shows that the improvement in MSE leads to clearly visible improvements in the characterization of $\bar{U}(d)$ that will translate to improvements in seeking the optimum.

Acknowledgements

References

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

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

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

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

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

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

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

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

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

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

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

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

H. Enderling and M. A. Chaplain. Mathematical modeling of tumor growth and treatment. *Current Pharmaceutical Design*, 20(30):4934–4940, 2014. ISSN 1381-6128/1873-4286.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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