On the Pitfalls of Nested Monte Carlo

Tom Rainforth, Rob Cornish, Hongseok Yang, and Frank Wood

{twgr,rcornish,fwood}@robots.ox.ac.uk, hongseok.yang@cs.ox.ac.uk

Overview

- Classical convergence proofs are insufficient for Nested Monte Carlo
- Despite this, nested inference is still used naïvely in a number of settings - e.g. probabilistic programming, experimental design, reinforcement learning
- We prove convergence, derive a convergence rate and provide empirical data that suggests it is observed in practise
- We prove that nested inference schemes are inherently biased
- Our results warn of the dangers of naïve nesting of inference schemes

Take Home

- Convergence is possible but requires additional assumptions to standard MC
  - Number of samples used in **each** call of the inner estimator must increase with number used in the outer
- Convergence rate is very slow - square of the number of samples of MC

Problem Formulation

Standard Monte Carlo:

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

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

We consider the case where λ is itself intractable:

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

We formally define nested Monte Carlo (NMC) as:

$$I \approx I_{N,M} = \frac{1}{N}\sum_{n=1}^{N} f(y_n,(\hat{\gamma}_M)_n) \quad \text{where} \quad y_n \sim p(y) \quad \text{and} \tag{4a}$$

$$(\hat{\gamma}_M)_n = \frac{1}{M}\sum_{m=1}^{M}\phi(y_n,z_{n,m}) \quad \text{where} \quad z_{n,m} \sim p(z|y_n). \tag{4b}$$

Reformulating to a Single Expectation

If f is linear in its 2^{nd} argument: $f(y,\alpha v+\beta w) = \alpha f(y,v)+\beta f(y,w)$, we can rearrange the problem to a single expectation

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

$\Rightarrow$ MC convergence rate for pseudo-marginal methods, PMCMC, ABC, etc

Motivating Examples

- Bayesian experimental design:

$$\text{IG}(x) = \mathbb{E}_{(y,z')\sim p(y,z'|x)}\left[\log p\left(y|z',x\right) - \log \mathbb{E}_{z\sim p(z|x)}\left[p\left(y|z,x\right)\right]\right]$$

- Nested queries in a probabilistic programming system

```
(defn outer-E [x M N]
  (-> (doquery :smc
        outer-query [x M])
      (take N)
      log-marginal))
```

```
(defn inner-E [x y M]
 (-> (doquery :smc
       inner-query [x y])
     (take M)
     log-marginal))
```

```
(defquery outer-query [x M]
  (let [z' (sample param-prior)
        experiment (setup-exp z' x)
        y (sample experiment)
        log-lik (observe* experiment y)
        log-marg (inner-E x y M)]
    (- loglik logmarg)))
```

```
(defquery inner-query [x y]
 (let [z (sample param-prior)]
 (observe (setup-exp z x) y)))
```

References

[1] 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.

[2] I. Murray, Z. Ghahramani, and D. J. MacKay. MCMC for doubly-intractable distributions. In *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence*, pages 359–366. AUAI Press, 2006.

[3] F. Wood, J.-W. van de Meent, and V. Mansinghka. A new approach to probabilistic programming inference. In *AISTATS*, pages 1024–1032, 2014.

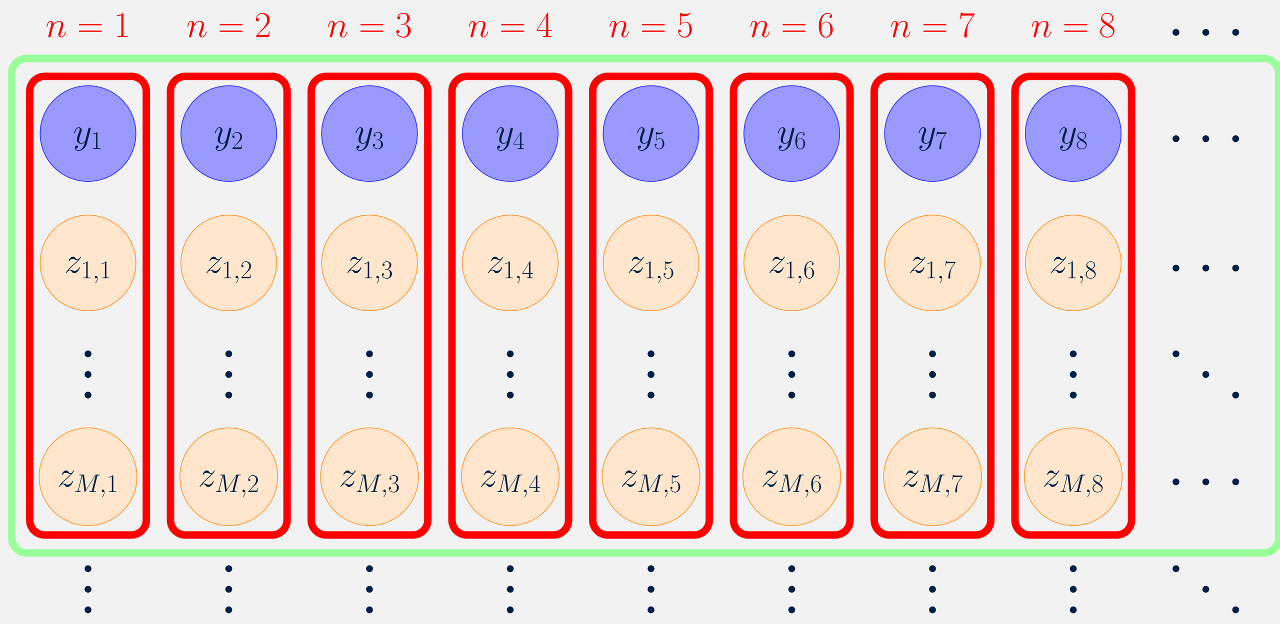
Acknowledgements

- BP, DARPA PPAML, NVIDIA

Almost Sure Convergence

Theorem 1. *Under mild assumptions on f, there exists a $\tau:\mathbb{N}\to\mathbb{N}$ such that $I_{\tau(M),M} \overset{a.s.}{\to} I$ as $M\to\infty$.*

Proof. Choose M large enough that $\left|I-\mathbb{E}\left[f(y_n,(\hat{\gamma}_M)_n)\right]\right| < \varepsilon$. For a fixed M, we have standard MC estimation on an expanded space $y, z_1, \ldots, z_M$, so we can choose $N=\tau(M)$ such that $\left|I_{\tau(M),M}-\mathbb{E}\left[f(y_n,(\hat{\gamma}_M)_n)\right]\right| < \frac{1}{M}$. We can thus make the total error arbitrary small almost surely as $M\to\infty$. □



Convergence Rate

Theorem 2. *If f is Lipschitz continuous, the mean squared error of $I_{N,M}$ converges at rate $O\left(1/N+1/M\right)$.*

Proof. By Minkowski $\|I-I_{N,M}\|_2^2 \le U^2+V^2+2UV \le 2\left(U^2+V^2\right)$ where

$$U = \left\|I-\frac{1}{N}\sum_{n=1}^{N}f(y_n,\gamma(y_n))\right\|_2 \qquad V = \left\|\frac{1}{N}\sum_{n=1}^{N}f(y_n,\gamma(y_n))-f(y_n,\gamma(y_n))\right\|_2$$

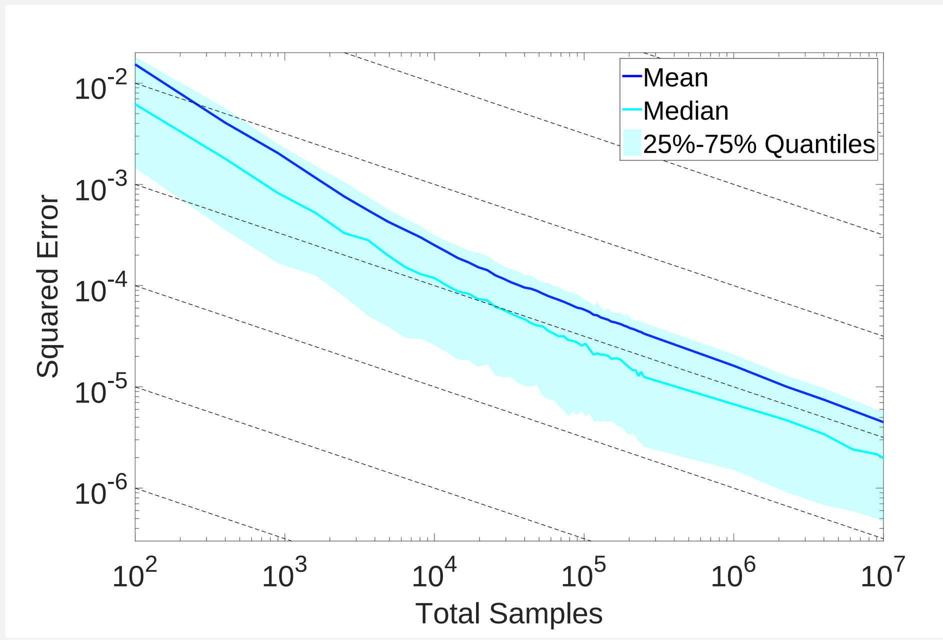
$U = O\left(1/\sqrt{N}\right)$, and using the assumption that f is Lipschitz continuous

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

where K is a fixed constant and $\|(\hat{\gamma}_M)_n-\gamma(y_n)\|_2 = O\left(1/\sqrt{M}\right)$. □

Emprical Results - Seem to Observe Rate in Practice

$$y \sim \text{Uniform}(-1,1)$$
$$z \sim \mathcal{N}(0,1)$$
$$\phi(z,y) = \sqrt{\frac{2}{\pi}}\exp\left(-2(y-z)^2\right)$$
$$f(y,\gamma(y)) = \log(\gamma(y)).$$



The Inherent Bias of Nested Inference

Theorem 3. *There does not exist a pair $(\mathcal{I},\mathcal{J})$ such that*

1. *the inner estimator $\mathcal{I}$ provides estimates $\hat{\gamma}_y \in \Phi$ at a given $y\in\mathcal{Y}$;*
2. *the outer estimator $\mathcal{J}$ maps a set of samples $\hat{\zeta} = \{(y_1,\hat{\gamma}_{y_1}),\ldots,(y_n,\hat{\gamma}_{y_n})\}$, to an unbiased estimate $\psi(\hat{\zeta},f)$ of $I(f)$, i.e. $\mathbb{E}[\psi(\hat{\zeta},f)] = I(f)$;*
3. $(\mathbb{E}_{y\sim p(y)}[\mathbb{E}[f(y,\hat{\gamma}_y)|y]]-\mathbb{E}[\psi(\hat{\zeta},f)]) \ge 0$ *for all integrable f.*

This result remains when ≥ 0 in the third condition is replaced by ≤ 0.

Proof. Construct a pair f_1 and f_2 where the above cannot hold for both. For example, $f_1(y,w) = (\gamma(y)-w)^2$ and $f_2(y,w) = -f_1(y,w)$ lead to

$$\mathbb{E}_{y\sim p(y)}\left[\mathbb{E}\left[f_1(y,\hat{\gamma}_y)|y\right]\right] > 0 > \mathbb{E}_{y\sim p(y)}\left[\mathbb{E}\left[f_2(y,\hat{\gamma}_y)|y\right]\right].$$

□