Control Functionals for Monte Carlo Integration

Chris J. Oates, Mark Girolami, and Nicolas Chopin

Medha Agarwal

Department of Statistics
University of Washington

June 10, 2022

Contents

Problem Introduction

Problem Setup

- The goal is to estimate the expectation: $\mu(f) = \int f(x)\pi(x)dx$ where
  - f is the test function
  - π is probability density
- Provided IID samples $\{X_i\}_{i=1}^n$ from π, the arithmetic mean estimator $\hat{\mu}$ converges to $\mu(f)$ at $O_P(n^{-1/2})$ using CLT.
- The error variance of $\hat{\mu}$ is $\sigma^2(f)/n$ where $\sigma^2(f)$ is the variance of f under π.
- When working with complex models $\sqrt{n}$ convergence is problematic and leads to estimators with high variance for finite samples.

Variance reduction techniques

- The aim is to reduce the variance of the estimator of μ, infact have it converge at a rate faster than $\sqrt{n}$, we call this super-$\sqrt{n}$-convergence.
- We wish to construct a variance reduction technique that agrees with the following four *desiderata*:
  1. Unbiased estimation
  2. Compatibility with unnormalized π
  3. Super $\sqrt{n}$-convergence
  4. Post-hoc schemes
- No variance reduction technique fulfils all four criteria.
- Methods that offer super $\sqrt{n}$-convergence are either biased or not amenable to unnormalized target.
- The control functionals method proposed by Oates et al. (2017) satisfy all!

Variance reduction techniques

Estimation Method	Unbiased	Unnormalized π	Super $\sqrt{n}$	Post hoc
MC (MCMC) + arithmetic mean	✓(×)	×(✓)	×	×
MC + importance sampling	✓(×)	×(✓)	×	×
MC + antithetic variables	✓	×	×	×
Quasi MC	×	×	✓	×
Randomized quasi MC	✓	×	✓	×
MC (/MCMC) + Rao-Blackwellization	✓(×)	×(✓)	×	✓
MC (/MCMC) + control variates	✓(×)	×(✓)	×	✓
MC (/MCMC) + Riemann sums	×	×(✓)	✓	✓
Bayesian MC	×	×	✓	✓
MC (/MCMC) + control functionals	✓(×)	×(✓)	✓	✓

Figure: Estimator properties based on the four desiderata. The properties change when handling unnormalized π which have been shown in parenthesis.

Control variates

- In control variates (CV), one seeks a function h, called control variate, such that $\mu(h)$ is known and h is heavily correlated with f.
- We seek basis statistics $\{s_i\}_{i=1}^m$ such that $\mu(s_i) = 0$ and $h = \sum_{i=1}^m a_i s_i$ for $\{a_i\}_{i=1}^m \subset \mathbb{R}$.
- A surrogate function is constructed $\tilde{f} = f - h$ for suitably chosen $\{a_i\}_{i=1}^m \subset \mathbb{R}$. Note that $\mathbb{E}_\pi[\tilde{f}] = \mathbb{E}_\pi[f]$ and $\mathrm{Var}(\tilde{f}) = \mathrm{Var}(f - h)$. Variance reduction is obtained if f and h are correlated.
- This is a misspecified regression problem as f does not necessarily belong to the span of $\{s_i\}$ basis functions.
- Control functionals are intuitively a non-parametric extension of control variates because the h does not belong to a finite parametric class, instead a **reproducing kernel Hilbert space (RKHS)**.

Zero-variance principle and ZVCV

- Assaraf and Caffarel (1999) propose a zero-variance principle that seeks
  1. Hermitian operator H such that $\int H(x,y)\sqrt{\pi(y)}dy = 0$, and
  2. an integrable function ψ such that the tuple (H, ψ) satisfy the **fundamental equation of zero variance**:

$$\nabla_{\mathbf{x}} \cdot \{\pi(\mathbf{x})\phi(\mathbf{x})\} = \{f(\mathbf{x}) - \mu(f)\}\pi(\mathbf{x})\,. \tag{1}$$

- The surrogate is then constructed as $\tilde{f}(x) = f(x) + \int H(x,y)\psi(y)dy/\sqrt{\pi(x)}$.

- The choice of H and ψ such that it satisfies (1) (at least approximately) determines the variance reduction obtained in this setup.

- Mira et al. (2013) propose zero-variance control variate (ZV-CV) estimator that chooses H as the Hamiltonian operator and parameterizes by a low degree polynomial.

Methodology

Set-up and notation

- **Main idea in CF**:
  1. Construct a RKHS $\mathcal{H}_+$ whose mean w.r.t π is analytically tractable.
  2. Approximate f by finding its projection on this space.

- The goal is to construct an increasingly more accurate approximation of f as $n \to \infty$.

- Suppose X is in $\Omega \subseteq \mathbb{R}^d$. It is assumed that Ω is bounded and the boundary $\partial\Omega$ is piecewise smooth.

- $\mu(f) := \int f(\mathbf{x})\pi(\mathbf{x})d\mathbf{x}$ and $\sigma^2(f) := \int \{f(\mathbf{x}) - \mu(f)\}^2 \pi(\mathbf{x})d\mathbf{x}$.

- Denote by $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^n$ the collection of states. At each x_i, the function values $f(\mathbf{x}_i)$ and gradients $\nabla_{\mathbf{x}} \log\{\pi(\mathbf{x}_i)\}$ are assumed to be pre-computed and cached.

- States are split into two disjoint sets: $\mathcal{D}_0 := \{\mathbf{x}_i\}_{i=1}^m$ and $\mathcal{D}_1 := \{\mathbf{x}_i\}_{i=m+1}^n$ where $1 \leq m < n$ but for theoretical and empirical results in paper, $m = O(n^\gamma)$.

From control variates to control functionals

- $\mathcal{D}_1$ is assumed to be IID samples from π independent of $\mathcal{D}_0$. In most examples where samples are obtained through MCMC, $m = n$.
- An approximation $s_{f,\mathcal{D}_0}$ of f is constructed using only the states in $\mathcal{D}_0$ such that $s_{f,\mathcal{D}_0} \in \mathcal{L}^2(\pi)$ and $\mu(s_{f,\mathcal{D}_0})$ is analytically tractable.
- Using $s_{f,\mathcal{D}_0}$, we can construct the surrogate function as

$$f_{\mathcal{D}_0}(\mathbf{x}) := f(\mathbf{x}) - s_{f,\mathcal{D}_0}(\mathbf{x}) + \mu(s_{f,\mathcal{D}_0})\,.$$

- By construction $f_{\mathcal{D}_0}$ has finite second moment w.r.t. π. $\mu(f_{\mathcal{D}_0}) = \mu(f)$, and $\sigma^2(f_{\mathcal{D}_0}) = \sigma^2(f - s_{f,\mathcal{D}_0})$.
- The final estimator is of the form

$$\hat{\mu}(\mathcal{D}_0, \mathcal{D}_1; f) := \begin{cases} \dfrac{1}{n-m} \sum_{i=m+1}^{n} f_{\mathcal{D}_0}(\mathbf{x}_i) & \text{for } m < n\,, \\ \mu(s_{f,\mathcal{D}_0}) & \text{for } m = n\,. \end{cases}$$

Close relation to control variates

- Notice that for $m < n$, we have unbiasedness, i.e. $\mathbb{E}_{\mathcal{D}_1}[\hat{\mu}(\mathcal{D}_0, \mathcal{D}_1; f)] = \mu(f)$ and estimator variance is $\text{Var}_{\mathcal{D}_1}(\hat{\mu}(\mathcal{D}_0, \mathcal{D}_1; f)) = (n-m)^{-1}\sigma^2(f - s_{f,\mathcal{D}_0})$.
- When $s_{f,\mathcal{D}_0}$ is constrained to belong to a finite dimensional space, then this is the formulation for control variates. For e.g. it is low-degree polynomials in ZV-CV.
- The idea is to use an infinite dimensional space to construct an increasingly accurate approximation $s_{f,\mathcal{D}_0}$ as $m \to \infty$.
- **Important result:** Assuming that $m = O(n^\gamma)$ and the expected functional approximation error vanishes, i.e. $\mathbb{E}_{\mathcal{D}_0}[\sigma^2(f - s_{f,\mathcal{D}_0})] = O(m^{-\delta})$, then $\mathbb{E}_{\mathcal{D}_0}[\mathbb{E}_{\mathcal{D}_1}[\{\hat{\mu}(\mathcal{D}_0, \mathcal{D}_1; f) - \mu(f)\}^2]] = O(n^{-1-\gamma\delta})$.
  $\therefore$ Take $\gamma = 1 \implies m = O(n)$ to obtain best convergence rate.
- Now we construct a RKHS for $s_{f,\mathcal{D}_0}$ using Stein's operator such that mean of its each element is known.

Control variates based on Stein's identity

Assumption

The density π belongs to $C^1(\Omega, \mathbb{R})$.

- This means that the vector $u(\mathbf{x}) := \nabla_{\mathbf{x}} \log(\pi(\mathbf{x}))$ is well defined.
- The Stein operator is defined as

$$\mathbb{S}_\pi : C^1(\Omega, \mathbb{R}^d) \to C^0(\Omega, \mathbb{R})$$
$$\phi(.) \mapsto \mathbb{S}_\pi(\phi)(.) := \nabla_{\mathbf{x}} \cdot \phi(.) + \phi(.) \cdot u(\mathbf{x})$$

  and the surrogate function is defined as $s_{f,\mathcal{D}_0} = c + \psi(x)$ where c is a constant and $\psi = \mathbb{S}_\pi(\phi)$ is recognized as the control functional (CF).

- Under a surface integral assumption, using this operator ensures that $\mu(\psi) = 0$ and so $\mu(s_{f,\mathcal{D}_0}) = c$. This is good news because it helps us construct a mean zero space and we try to find the projection of f in this space.
- This paper takes the innovative step of setting ϕ within a RKHS to enable fully non-parametric approximation. Now we construct it!

Hilbert space of control functionals

- Recall that the control functional is $\psi(\mathbf{x}) = \nabla_{\mathbf{x}} \cdot \phi(\mathbf{x}) + \phi(\mathbf{x}) \cdot u(\mathbf{x})$.
- Specifying ψ is equivalent to specifying ϕ. Restrict each component function $\phi_i : \Omega \to \mathbb{R}$ to a Hilbert space $\mathcal{H} \subset \mathcal{L}^2(\pi) \cap C^1(\Omega, \mathbb{R})$ with inner product $\langle ., . \rangle_{\mathcal{H}}$.
- Moreover, $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS) with kernel k.
- All experiments in the paper are done using the kernel $k(\mathbf{x}, \mathbf{x}') = (1 + \alpha_1 \|\mathbf{x}\|_2^2)^{-1} (1 + \alpha_1 \|\mathbf{x}'\|_2^2)^{-1} \exp\{(-2\alpha_2)^{-1} \|\mathbf{x} - \mathbf{x}'\|_2^2\}$
- The vector valued function $\phi : \Omega \to \mathbb{R}^d$ is defined as the Cartesian product space $\mathcal{H}^d := \mathcal{H} \times \cdots \times \mathcal{H}$ which in itself is a RKHS.

Assumption

k belongs to $C^2(\Omega \times \Omega, \mathbb{R})$.

The class of control functionals

- Then ψ belongs to $\mathcal{H}_0$, the RKHS with kernel

$$k_0(\mathbf{x}, \mathbf{x}') := \nabla_{\mathbf{x}} \cdot \nabla_{\mathbf{x}'} k(\mathbf{x}, \mathbf{x}') + u(\mathbf{x}) \cdot \nabla_{\mathbf{x}'} k(\mathbf{x}, \mathbf{x}') + u(\mathbf{x}') \cdot \nabla_{\mathbf{x}} k(\mathbf{x}, \mathbf{x}') + u(\mathbf{x}) \cdot u(\mathbf{x}') k(\mathbf{x}, \mathbf{x}')$$

- Under assumptions above and a surface integral condition, the gradient based kernel k_0 satisfies $\int k_0(\mathbf{x}, \mathbf{x}')\pi(\mathbf{x}')d\mathbf{x}' = 0$ which implies that $\mathcal{H}_0$ consists only of legitimate CF, i.e. $\psi \in \mathcal{H}_0$ such that $\mu(\psi) = 0$.

- Let $\mathcal{C}$ denote the RKHS for constant functions with kernel $k_{\mathcal{C}}(\mathbf{x}, \mathbf{x}') = 1$ for all $\mathbf{x}, \mathbf{x}' \in \Omega$.

- We can show that $\mathcal{H}_+ := \mathcal{C} + \mathcal{H}_0$ is also a RKHS with kernel $k_+(\mathbf{x}, \mathbf{x}') := k_{\mathcal{C}}(\mathbf{x}, \mathbf{x}') + k_0(\mathbf{x}, \mathbf{x}')$. Well-posedness is assumed, i.e. $f \in \mathcal{H}_+$.

- The authors use the regularized least square approximation to find the projection of f in $\mathcal{H}_+$

$$s_{f,\mathcal{D}_0} := \arg\min_{g \in \mathcal{H}_+} \left[\frac{1}{m} \sum_{j=1}^{m} \left(f(\mathbf{x}_j) - g(\mathbf{x}_j) \right)^2 + \lambda \|g\|^2_{\mathcal{H}_+} \right] \quad \lambda > 0.$$

CFs based on regularized least squares

- Using the representor theorem, there exists a unique solution of the form $\hat{\psi} = \sum_{i=1}^{m} \beta_i k_0(\mathbf{x}_i, \mathbf{x})$ that solves the optimization problem.

- RLS estimate leads to a convenient closed form expression for the CF estimator under the smoothness assumptions discussed before.

- Denote $\mathbf{f}_0 := (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_m))^T$, $\mathbf{1}_p = (1, \ldots, 1)^T$ for a vector of p ones, and $\mathbf{f}_1 = (f(\mathbf{x}_{m+1}), \ldots, f(\mathbf{x}_n))^T$, and $(\mathbf{K}_0)_{i,j} = k_0(\mathbf{x}_i, \mathbf{x}_j)$. Then, the solution is

$$\hat{\mu}(\mathcal{D}_0, \mathcal{D}_1; f) = \frac{1}{n-m} \mathbf{1}^T (\mathbf{f}_1 - \hat{\mathbf{f}}_1) + \frac{1^T (K_0 + \lambda I)^{-1} \mathbf{f}_0}{1 + \mathbf{1}^T (K_0 + \lambda I)^{-1} \mathbf{1}}$$

- Here $\hat{\mathbf{f}}_1$ is the vector of fitted values for $\mathbf{f}_1$ with $(K_{1,0})_{i,j} = k_0(x_{m+i}, x_j)$

$$\hat{\mathbf{f}}_1 := K_{1,0}(K_0 + \lambda m I)^{-1} \mathbf{f}_0 + \{\mathbf{1} - K_{1,0}(K_0 + \lambda m I)^{-1} \mathbf{1}\} \frac{\mathbf{1}^T (K_0 + \lambda I)^{-1} \mathbf{f}_0}{1 + \mathbf{1}^T (K_0 + \lambda I)^{-1} \mathbf{1}}$$

Few remarks

Recall that the CF estimator for $\mu(f)$ is given by

$$\hat{\mu}(\mathcal{D}_0, \mathcal{D}_1; f) = \underbrace{\frac{1}{n-m}\mathbf{1}^T(\mathbf{f}_1 - \hat{\mathbf{f}}_1)}_{\text{(i)}} + \underbrace{\frac{1^T(K_0 + \lambda I)^{-1}\mathbf{f}_0}{1 + \mathbf{1}^T(K_0 + \lambda I)^{-1}\mathbf{1}}}_{\text{(ii)}} \tag{2}$$

- Term (ii) is equivalent to Bayesian Monte Carlo based on $\mathcal{H}_+$, i.e. term (ii) is posterior mean for $\mu(f)$ based on Gaussian process with prior $f \sim GP(0, k_+)$ and data $\mathcal{D}_0$ (Rasmussen and Ghahramani, 2003).

- The samples $\mathcal{D}_1$ only enter $\hat{\mu}(\mathcal{D}_0, \mathcal{D}_1; f)$ through term (i) which vanishes in probability to 0 as $m \to \infty$. Thus randomness due to $\mathcal{D}_1$ vanishes, this explains

- Even though CF estimator holds better theoretical properties like unbiasedness, practically the estimator performs the best when $m = n$, i.e. $\hat{\mu}(\mathcal{D}_0, \mathcal{D}_1; f)$ is BMC estimator. Also, the latter is amenable to MCMC sampling procedures.

Results

Theoretical results - Bias

- Firstly, by our construction of Hilbert space $\mathcal{H}_+$ and RLS optimal $s_{f,\mathcal{D}_0} \in \mathcal{H}_+$, we know $\mu(s_{f,\mathcal{D}_0})$ and our surrogate function is $f_{\mathcal{D}_0} = f - s_{f,\mathcal{D}_0} + \mu(s_{f,\mathcal{D}_0})$.

- Recall that

$$\hat{\mu}(\mathcal{D}_0, \mathcal{D}_1; f) := \begin{cases} \frac{1}{n-m} \sum_{i=m+1}^{n} f_{\mathcal{D}_0}(\mathbf{x}_i) & \text{for } m < n\,, \\ \mu(s_{f,\mathcal{D}_0}) & \text{for } m = n\,. \end{cases}$$

- This implies that for $m < n$, $\hat{\mu}(\mathcal{D}_0, \mathcal{D}_1; f)$ is an unbiased estimator of $\mu(f)$.

- However, for $m = n$, $\hat{\mu}(\mathcal{D}_0, \mathcal{D}_1; f) = \mu(s_{f,\mathcal{D}_0}) = \dfrac{\mathbf{1}^T(\mathbf{K}_0 + \lambda I)^{-1}\mathbf{f}_0}{(1 + \mathbf{1}^T(\mathbf{K}_0 + \lambda I)^{-1}\mathbf{1})}$. This is a weighted sum of $\mathbf{f}_0$ with weights not adding up to 1.

- This implies that the CF estimator for $m = n$ is biased. Infact it can be shown that $\hat{\mu}(\mathcal{D}_0, \mathcal{D}_1; f)$ is negatively biased.

Theoretical results - Error Variance

Under the conditions

1. All assumptions made so far are satisfied and $\sup_{\mathbf{x}\in\Omega} k_+(\mathbf{x},\mathbf{x}') < \infty$ - verified for the authors' choice of k_0,
2. the RLS estimate of $s_{f,\mathcal{D}_0} \in \mathcal{H}_+$ is obtained with $\lambda = O(m^{-1/2})$,
3. $\mathcal{D}_0$ are IID samples from π, and
4. the estimator $\hat{\mu}(\mathcal{D}_0, \mathcal{D}_1; f)$ is an unbiased estimator of $\mu(f)$,

we have

$$\mathbb{E}_{\mathcal{D}_0}[\mathbb{E}_{\mathcal{D}_1}[(\hat{\mu}(\mathcal{D}_0, \mathcal{D}_1; f) - \mu(f))^2]] = O(n^{-7/6})$$

Some shortcomings

- **Curse of Dimensionality:**
  1. Oates et al. (2019) in a follow-up work show that the rate of convergence of CF is $O_P(n^{-1-2(ab)/d+\epsilon})$, where d is problem dimension, a and b are related to the smoothness levels of π and f respectively, and ϵ is an arbitrarily small constant.
  2. Computational cost associated with solving the linear system to get $\hat{\mu}(\mathcal{D}_0, \mathcal{D}_1; f)$ is $O(n^3)$.
  3. In plain Monte Carlo, error bound $O(c^{-1/2})$ is achieved for cost c. In CF, for cost c, the error bound achieved is $O(c^{-1/6-(a\wedge b)/3d+\epsilon})$. So, it is only advisable to do CF if $a \wedge b > d$.
- **Multimodal target:**
  1. In case of even smooth one-dimensional multimodal distributions, e.g. Gaussian mixture, CF and simplified CF perform worse than simple Monte Carlo.

Gaussian mixture illustration

Consider the Gaussian mixture density

$$\pi(x) = p\mathcal{N}(-\mu, 1) + (1-p)\mathcal{N}(\mu, 1)$$

where p is the mixing proportion and $\pm\mu$ is mean of mixing Gaussian distributions. We are interested in estimating two different integrands:

1. $f(x) = x^2$ is a polynomial wherein we expect ZV-CV to perform well, and
2. $f(x) = \exp(x)$ is a very smooth function.

Here we take $n = 200$ and repeat the estimation process for 100 replications for four different combinations of π and f.

Problem		Estimator		
$f(x)$	π	Arithmetic Means	ZV-CV	CF (simplified)
$\exp(x)$	$\mathcal{N}(0,1)$	1.6486 $\pm$0.0944	1.6277 $\pm$0.0049	1.6447 $\pm 10^{-4}$
x^2	$0.5\mathcal{N}(-1,1) + 0.5\mathcal{N}(1,1)$	1.957 $\pm$0.1448	2.0082 $\pm$0.0022	1.9875 $\pm 10^{-4}$
x^2	$0.5\mathcal{N}(-2,1) + 0.5\mathcal{N}(2,1)$	4.9121 $\pm$0.4016	4.8835 $\pm$0.0305	2.6159 $\pm$6.9807
$\exp(x)$	$0.5\mathcal{N}(-2,1) + 0.5\mathcal{N}(2,1)$	6.1447 $\pm$6.2988	5.9863 $\pm$2.4247	4.829 $\pm$20.6664

Experiments

Illustrative example, $d = 1$

- $f(X) = \sin(\pi/d) \sum_{i=1}^{d} X_i$ where X is a standard Gaussian random variable.
- Simple case of $d = 1$ and $n = 50$. $\mathcal{D}_0$ and $\mathcal{D}_1$ each get half of the samples.

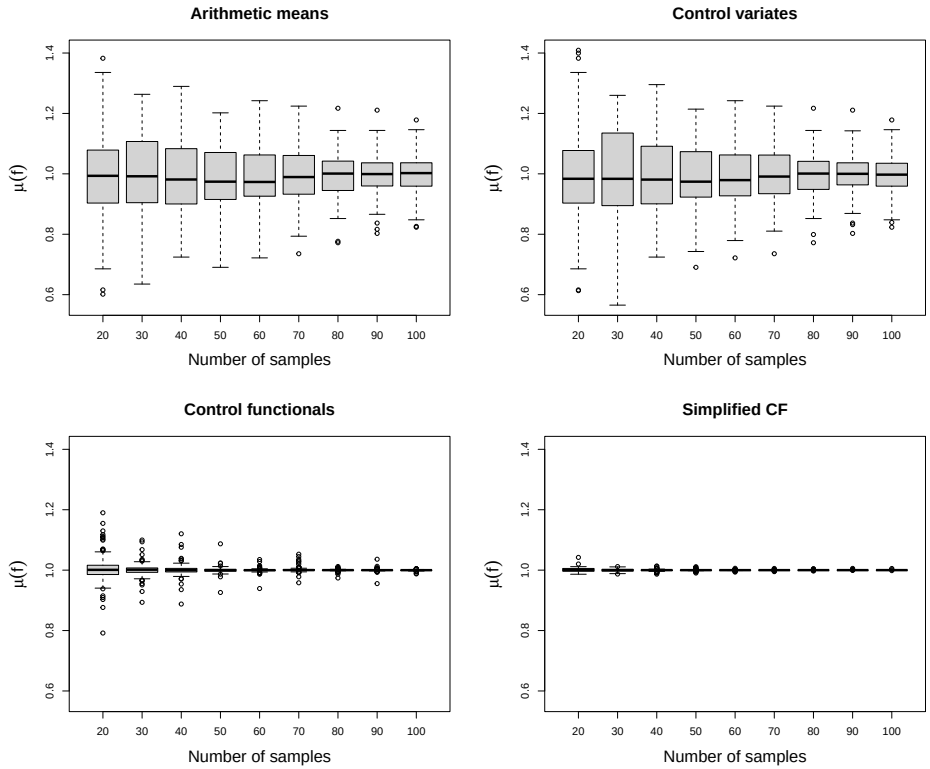

Figure: Sampling distribution of (a) arithmetic means, (b) ZV-CV, (c) CF, (d) simplified CF for 100 replications for low dimensional $d = 1$ case.

Illustrative example, $d = 10$

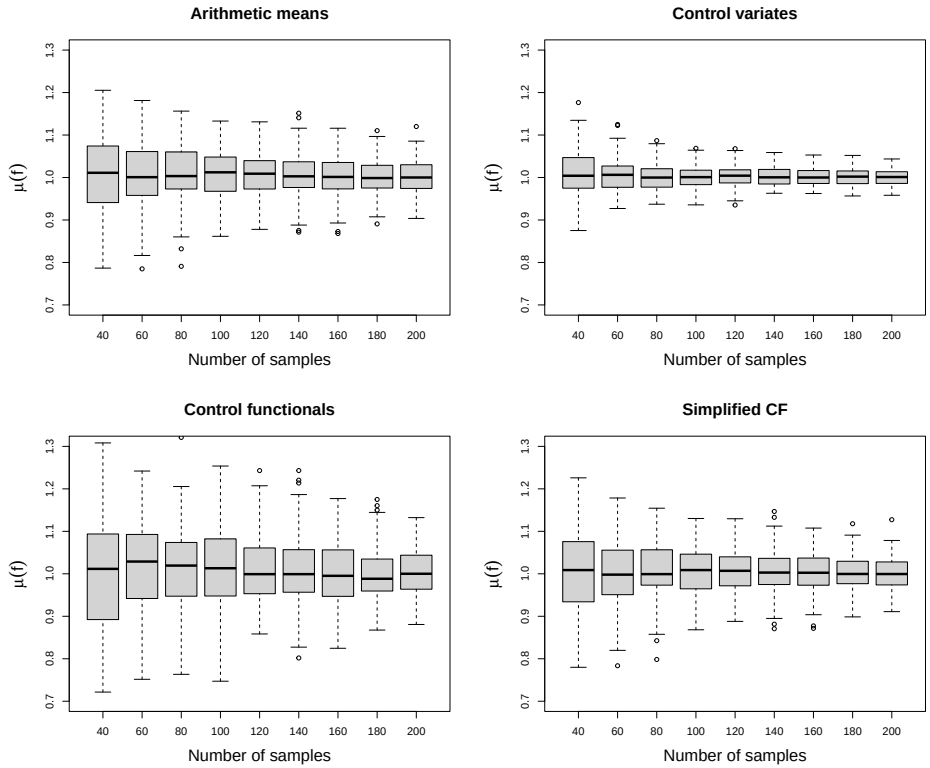

Figure: Sampling distribution of (a) arithmetic means, (b) ZV-CV, (c) CF, (d) simplified CF for 100 replications for high dimensional $d = 10$ case.

SARCOS Arm Example: Marginalising hyper-parameters in GP model

- $y_i \in \mathbb{R}$ is the measured response variable at state $z_i \in \mathbb{R}^p$, such that

$$Y_i = g(z_i) + \epsilon_i \text{ where } \epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

  $\sigma > 0$ is known. $\mathbf{y} = (y_1, \ldots, y_N)$ and $\mathbf{z} = (z_1, \ldots, z_N)$.

- GP prior $g \sim \mathcal{GP}(0, c(z, z'; \theta))$ where $c(z, z'; \theta) = \theta_1 \exp\{-[1/(2\theta_2)^2]\|z - z'\|_2^2\}$. Here $\theta = (\theta_1, \theta_2)$ are independent hyperparameters jointly distributed as π. $\theta_1 \sim \Gamma(\alpha, \beta)$ and $\theta_2 \sim \Gamma(\gamma, \delta)$.

- We are interested in predicting the value of the response Y_* corresponding to an unseen state vector z_*. The estimator is Bayesian posterior mean

$$\hat{Y_*} = \mathbb{E}[Y_*|z_*] = \int E[Y_*|y, \theta]\pi(\theta)d\theta$$

- Monte Carlo estimates are obtained by sampling $\{\theta_i\}_{i=1}^n$ from prior $\pi(\theta)$ and the integrand of interest is

$$f(\theta) = \mathbb{E}[Y_*|\mathbf{y}, \theta] = C_{*,N}(C_N + \sigma^2 I_{N\times N})^{-1}\mathbf{y}$$

  where $(C_N)_{i,j} = c(\mathbf{z}_i, \mathbf{z}_j; \theta)$ and $(C_{*,N})_{1,j} = c(\mathbf{z}_*, \mathbf{z}_j; \theta)$.

SARCOS Arm Example: Marginalising hyper-parameters in GP model

- We use the hierarchical GP model described above to estimate the inverse dynamics of a 7 degrees of freedom Sarcos anthropomorphic robot arm (Seeger, 2004).
- Here the covariates $\mathbf{z}$ are 21 dimensional (7 positions, 7 velocities, and 7 accelerations) and the response $\mathbf{y}$ is 7 dimensional (7 joint torques).
- The training data has $N = 444484$ and testing data has 4449 points.
- Here $\sigma = 0.1, \alpha = \gamma = 25$ and $\beta = \delta = 0.04$, so that each hyperparameter θ_i has prior mean 1 and standard deviation 0.2.
- For CF, recall we use the kernel $k(\mathbf{x}, \mathbf{x}') = (1 + \alpha_1 \|\mathbf{x}\|_2^2)^{-1}(1 + \alpha_1 \|\mathbf{x}'\|_2^2)^{-1} \exp\{(-2\alpha_2)^{-1} \|\mathbf{x} - \mathbf{x}'\|_2^2\}$.
- The kernel parameters are $\alpha_1 = 0.1$ and $\alpha_2 = 1$.

SARCOS Arm Example

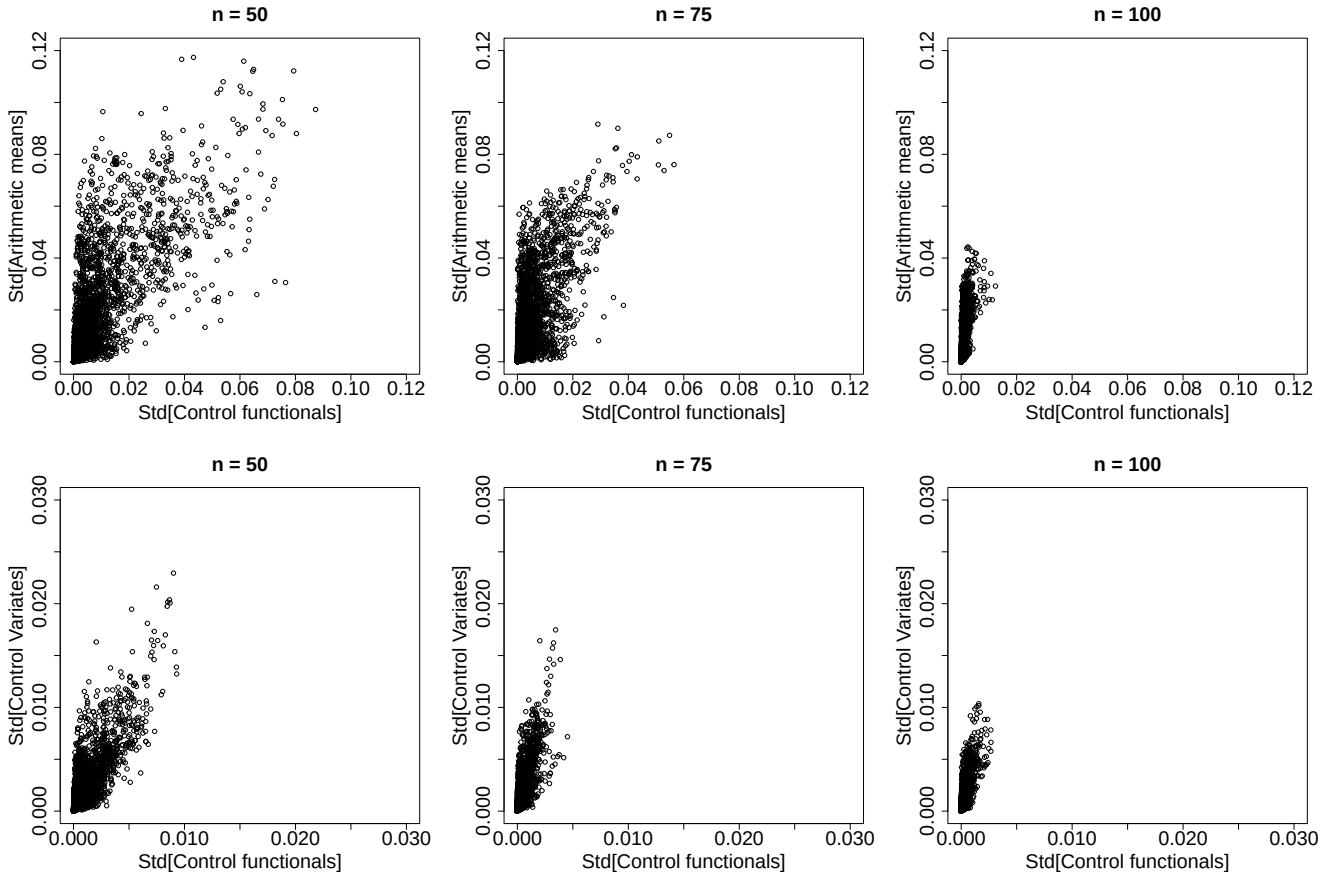

Figure: Sampling standard deviation of Monte Carlo estimators for the posterior predictive mean $\mathbb{E}[Y_*|\mathbf{y}]$ computed over 10 independent realisations. Each point, representing one (out of 4449) Monte Carlo integration problem.

Thermodynamic integration

- Suppose we have data $\mathbf{y}$ from some model. For a particular model m characterized by θ, the model evidence is

$$p(\mathbf{y}) = \int p(\mathbf{y}|\theta)p(\theta)d\theta.$$

- Power posterior is defined as $p(\theta|\mathbf{y}, t) = p(\mathbf{y}|\theta)^t p(\theta)/Z_t(\theta)$ where t is inverse temperature parameter and $Z_t(\theta) = \int p(\mathbf{y}|\theta)^t p(\theta)d\theta$ is the normalizing constant.
- Notice that $Z_1(\theta) = p(\mathbf{y})$ is the model evidence (desired quantity).
- **Important thermodynamic identity:** $\log\{p(\mathbf{y})\} = \int_0^1 \mathbb{E}_{\theta|\mathbf{y},t}[\log(p(\mathbf{y}|\theta))]dt$.
- The above integral is approximated using 4 quadrature method

$$\log(p(\mathbf{y})) \approx \sum_{t=0}^{m} (t_{i+1} - t_i)\frac{\hat{\mu}_{i+1} + \hat{\mu}_i}{2} - (t_{i+1} - t_i)^2 \frac{\hat{\nu}_{i+1} - \hat{\nu}_i}{12},$$

where $\hat{\mu}_i$ and $\hat{\nu}_i$ are Monte Carlo estimates of the posterior mean and variance respectively of $\log p(\mathbf{y}|\theta)$ when θ arises from $p(\theta|\mathbf{y}, t_i)$.

Non-linear ODE models

- We consider nonlinear dynamical systems of the form

$$\frac{d\mathbf{x}}{ds} = f(\mathbf{x}, s, \theta), \qquad \mathbf{x}(0) = \mathbf{x}_0\,, \tag{3}$$

- Only a subset of the variables are observed under noise, so that $\mathbf{x} = [\mathbf{x}_a, \mathbf{x}_b]$ and $\mathbf{y}$ is a $d \times n$ matrix of observations of the variables $\mathbf{x}_a$ at times $s_1 < s_2 < \cdots < s_n$.

- We consider a Gaussian observation process with likelihood

$$p(\mathbf{y}|\theta, \mathbf{x}_0, \sigma) = \prod_{i=1}^{n} N(\mathbf{y}(s_j)|\mathbf{x}_a(s_j; \theta, \mathbf{x}_0), \sigma^2 I)\,,$$

  where $\mathbf{x}_a(s_j; \theta, \mathbf{x}_0)$ is the solution of the system in (3) and $\sigma > 0$ is known.

- Both solutions to ODE and gradient of power density are calculated using numerical methods - Runge Kutta method and finite differences respectively.

van der Pol oscillator

- van der Pol oscillator is a non-conservative oscillator with non-linear damping. Here a position $x(s) \in \mathbb{R}$ evolves in time s according to the second order differential equation

$$\frac{d^2x}{ds^2} - \theta(1 - x^2)\frac{dx}{ds} + x = 0$$

where $\theta \in \mathbb{R}$ is an unknown parameter indicating the non-linearity and the strength of the damping.

- Letting $x_1 := x$ and $x_2 := dx/dt$ we can formulate the oscillator as the first-order system

$$f(\mathbf{x}, s, \theta) = \theta(1 - x_1^2)x_2 - x_1$$

where only the first component x_1 is observed.

- This system was solved numerically using $\theta = 1, x_0 = [0, 2]$. Observations were made once every time unit, up to 10 units, and Gaussian measurement noise of standard deviation $\sigma = 0.1$ was added.

- A log-normal prior was placed on θ such that $\log(\theta) \sim N(0, 0.25)$.

Normalizing constant for non-linear ODEs

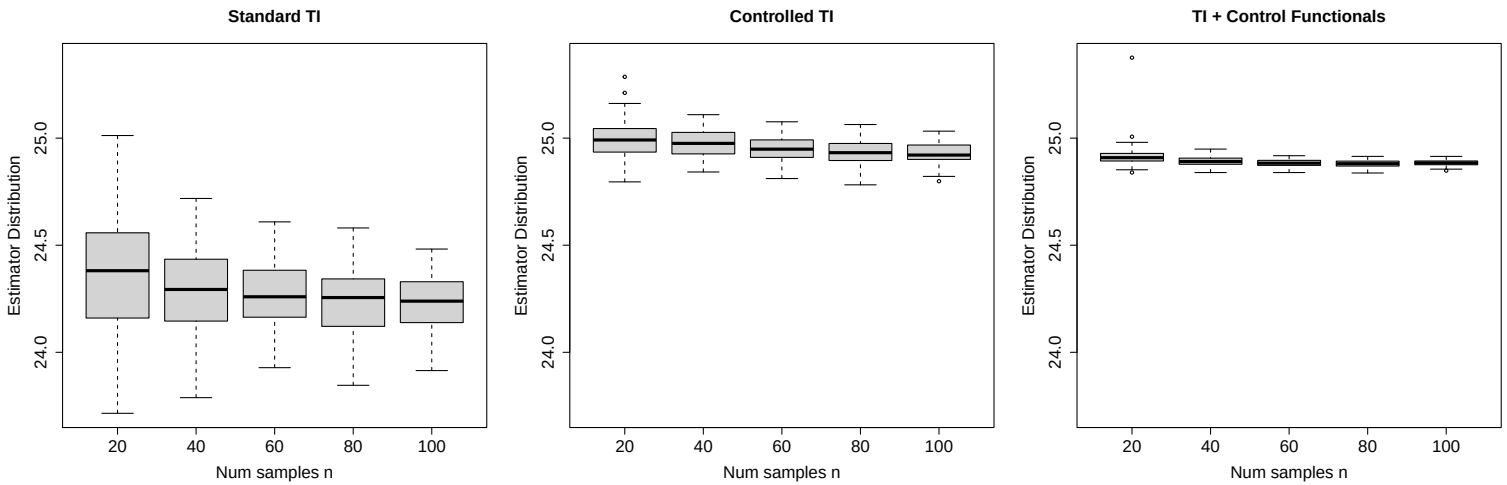

Figure: Sampling distribution of different Monte Carlo estimators over 100 replications.

References

Assaraf, R. and Caffarel, M. (1999). Zero-variance principle for monte carlo algorithms. *Physical review letters*, 83(23):4682.

Mira, A., Solgi, R., and Imparato, D. (2013). Zero variance markov chain monte carlo for bayesian estimators. *Statistics and Computing*, 23(5):653–662.

Oates, C. and Girolami, M. (2016). Control functionals for quasi-monte carlo integration. In *Artificial Intelligence and Statistics*, pages 56–65. PMLR.

Oates, C. J., Cockayne, J., Briol, F.-X., and Girolami, M. (2019). Convergence rates for a class of estimators based on stein' s method. *Bernoulli*, 25(2):1141–1159.

Oates, C. J., Girolami, M., and Chopin, N. (2017). Control functionals for monte carlo integration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):695–718.

Oates, C. J., Papamarkou, T., and Girolami, M. (2016). The controlled thermodynamic integral for bayesian model evidence evaluation. *Journal of the American Statistical Association*, 111(514):634–645.

Rasmussen, C. E. and Ghahramani, Z. (2003). Bayesian monte carlo. *Advances in neural information processing systems*, pages 505–512.

Seeger, M. (2004). Gaussian processes for machine learning. *International journal of neural systems*, 14(02):69–106.

Thank you!

Questions?

Appendix

Appendix: Some questions with answers in other literature

- What is the convergence rate when $m = n$ and therefore, the estimator is biased?
- In practice most samples are acquired using MCMC, so how is the convergence rate affected when $\mathcal{D}_0$ are not IID?
- What is the ideal splitting of samples between $\mathcal{D}_0$ and $\mathcal{D}_1$?

Appendix: Answers

- Answer: 1
  1. The $m = n$ case is called 'simplified' CF. No theoretical analysis for its convergence rate but the empirical results show it has lesser variance than CF.
  2. One explanation is that $\hat{\mu}(\mathcal{D}_0, \mathcal{D}_1; f)$ for $m = n$ comprises only of term (ii) from (2), so any variability arising from term (i) is naturally eliminated.
- Answer: 2
  1. In a follow-up paper, Oates et al. (2016) show that if the samples are generated from a uniformly ergodic Markov chain, then the rate of convergence remains same.
  2. Using MCMC can cause repeated samples which can make the matrix $\mathbf{K}_0$ singular and render a trivial solution $s_{f,\mathcal{D}_0} \equiv 0$. 'Filtering' recommended for $\mathcal{D}_0$.
- Answer: 3
  1. If a and b relate to the smoothness of π and f respectively, d is the problem dimension, then the Oates et al. (2016) suggest the optimal split
  $$\frac{m}{n} = \frac{\nu}{1+\nu} \quad \text{where } \nu = 2\frac{a \wedge b}{d}$$
  2. Therefore, if π and f are very smooth, i.e. $a \wedge b \gg d$, then all the samples should be assigned to $\mathcal{D}_0$ and the method becomes equivalent to numerical quadrature, whereas if $a \wedge b \ll d$, it is advisable to do plain Monte Carlo.

Appendix: Follow-up work

- **Convergence rates:** Oates et al. (2019) prove that the CF estimator incurs an $O_P(n^{-1-2(a\wedge b)/d+\epsilon})$ error variance where a and b are related to the smoothness of π and f respectively, d is the dimension of domain, and ϵ is an arbitrarily small constant.

- **Study in QMC setting:**
  1. QMC algorithms can theoretically achieve a $O_P(n^{-\alpha/d})$ rate for integrands with derivatives of order α.
  2. However, α-QMC is often not employed in practice, instead a sub-optimal α_l-QMC is used for $\alpha_L < \alpha$.
  3. In Oates and Girolami (2016), it is shown that CF+QMC accelerates α_L-QMC by a factor of $O_P(n^{-(\alpha-\alpha_L)/d})$.