Control Functionals for Monte Carlo Integration

Mark Girolami[2], joint work with Chris. J. Oates[1], Nicolas Chopin[3]

[1] University of Newcastle.

[2] Imperial College London, UK.

[2] CREST-LS and ENSAE, Paris, France.

MLSS2019

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2017)
79, Part 3, pp. 695–718

Control functionals for Monte Carlo integration

Chris J. Oates,

University of Technology Sydney, Australia

Mark Girolami

University of Warwick, Coventry, and Alan Turing Institute, London, UK

and Nicolas Chopin

Centre de Recherche en Economie et Statistique and Ecole Nationale de la Statistique et de l'Administration Economique, Paris, France

[Received October 2014. Final revision February 2016]

Summary. A non-parametric extension of control variates is presented. These leverage gradient information on the sampling density to achieve substantial variance reduction. It is not required that the sampling density be normalized. The novel contribution of this work is based on two important insights: a trade-off between random sampling and deterministic approximation and a new gradient-based function space derived from Stein's identity. Unlike classical control variates, our estimators improve rates of convergence, often requiring orders of magnitude fewer simulations to achieve a fixed level of precision. Theoretical and empirical results are presented, the latter focusing on integration problems arising in hierarchical models and models based on non-linear ordinary differential equations.

Keywords: Control variates; Non-parametrics; Reproducing kernel; Stein's identity; Variance reduction

Averaging

Problem: Compute an expectation

$$\mu = \int_{\mathcal{X}} f \, d\pi$$

where π is the density of a random variable $X \colon \Omega \to \mathcal{X}$ defined on a probability space $(\Omega, \sigma, \lambda)$ and taking values in a measurable space $(\mathcal{X}, \mathcal{A})$ equipped with a reference measure (with respect to which π is defined), and $f \colon \mathcal{X} \to \mathcal{Y}$ is a function of interest.

This can be a highly non-trivial problem when either

1. f is expensive to evaluate, or
2. π is intractable.

In contemporary applications of Bayesian statistics we have BOTH problems!

Averaging

Problem: Compute an expectation

$$\mu = \int_{\mathcal{X}} f \, d\pi$$

where π is the density of a random variable $X \colon \Omega \to \mathcal{X}$ defined on a probability space $(\Omega, \sigma, \lambda)$ and taking values in a measurable space $(\mathcal{X}, \mathcal{A})$ equipped with a reference measure (with respect to which π is defined), and $f \colon \mathcal{X} \to \mathcal{Y}$ is a function of interest.

This can be a highly non-trivial problem when either

1. f is expensive to evaluate, or
2. π is intractable.

In contemporary applications of Bayesian statistics we have BOTH problems!

Averaging

Problem: Compute an expectation

$$\mu = \int_{\mathcal{X}} f \, d\pi$$

where π is the density of a random variable $X : \Omega \to \mathcal{X}$ defined on a probability space $(\Omega, \sigma, \lambda)$ and taking values in a measurable space $(\mathcal{X}, \mathcal{A})$ equipped with a reference measure (with respect to which π is defined), and $f : \mathcal{X} \to \mathcal{Y}$ is a function of interest.

This can be a highly non-trivial problem when either

1. f is expensive to evaluate, or
2. π is intractable.

In contemporary applications of Bayesian statistics we have BOTH problems!

Averaging

Problem: Compute an expectation

$$\mu = \int_{\mathcal{X}} f \, d\pi$$

where π is the density of a random variable $X : \Omega \to \mathcal{X}$ defined on a probability space $(\Omega, \sigma, \lambda)$ and taking values in a measurable space $(\mathcal{X}, \mathcal{A})$ equipped with a reference measure (with respect to which π is defined), and $f : \mathcal{X} \to \mathcal{Y}$ is a function of interest.

This can be a highly non-trivial problem when either

1. f is expensive to evaluate, or
2. π is intractable.

In contemporary applications of Bayesian statistics we have BOTH problems!

Averaging

Problem: Compute an expectation

$$\mu = \int_{\mathcal{X}} f \, d\pi$$

where π is the density of a random variable $X\colon \Omega \to \mathcal{X}$ defined on a probability space $(\Omega, \sigma, \lambda)$ and taking values in a measurable space $(\mathcal{X}, \mathcal{A})$ equipped with a reference measure (with respect to which π is defined), and $f\colon \mathcal{X} \to \mathcal{Y}$ is a function of interest.

This can be a highly non-trivial problem when either

1. f is expensive to evaluate, or
2. π is intractable.

In contemporary applications of Bayesian statistics we have BOTH problems!

Averaging

Problem: Compute an expectation

$$\mu = \int_{\mathcal{X}} f \, d\pi$$

where π is the density of a random variable $X : \Omega \to \mathcal{X}$ defined on a probability space $(\Omega, \sigma, \lambda)$ and taking values in a measurable space $(\mathcal{X}, \mathcal{A})$ equipped with a reference measure (with respect to which π is defined), and $f : \mathcal{X} \to \mathcal{Y}$ is a function of interest.

This can be a highly non-trivial problem when either

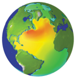
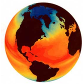
1. f is expensive to evaluate, or
2. π is intractable.

In contemporary applications of Bayesian statistics we have BOTH problems!

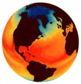
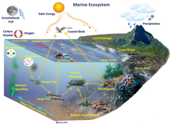
Averaging

e.g. Consider the Bayesian solution of an inverse problem:

Expensive likelihood precludes efficient sampling from $p(\boldsymbol{\theta}|\boldsymbol{y})$:

$$\log p(\boldsymbol{y}|\boldsymbol{\theta}) = -\frac{1}{2\sigma^2}\Bigg\|$$

$$-$$

$$\Bigg\|^2 + C$$

Expensive function of model parameters $\boldsymbol{\theta}$:

$$f(\boldsymbol{\theta}) = g\Bigg($$

$$\Bigg) =$$


Posterior expectations $\mu = \mathbb{E}[f(\boldsymbol{\theta})|\boldsymbol{y}]$ must be estimated from few samples:

$$\mu \approx \frac{1}{10}\sum_{n=1}^{10} f(\boldsymbol{\theta}_n), \qquad \boldsymbol{\theta}_i \sim p(\boldsymbol{\theta}|\boldsymbol{y}).$$

Unacceptably high variance.

A Survey of Approaches to Estimate $\mu(f) = \int f d\pi$

Estimation Method	Unbiased	π intractable	Sub-root-n	Post-hoc
Monte Carlo (MC)	$\checkmark$	$\times$	$\times$	$\times$
Markov Chain MC (MCMC)	$\times$	$\checkmark$	$\times$	$\times$
MC + Importance Sampling	$\checkmark(/\times)$	$\times(/\checkmark)$	$\times$	$\checkmark$
MCMC + Rao-Blackwellisation	$\times$	$\checkmark$	$\times$	$\checkmark$
MC(/MCMC) + Control Variates	$\checkmark(/\times)$	$\times(/\checkmark)$	$\times$	$\checkmark$
MC + Antithetic Variables	$\checkmark$	$\times$	$\times$	$\times$
MC(/MCMC) + Stratified Sampling	$\checkmark(/\times)$	$\times(/\checkmark)$	$\times$	$\times$
Quasi-MC (QMC)	$\times$	$\times$	$\checkmark$	$\times$
Randomised QMC (RQMC)	$\checkmark$	$\times$	$\checkmark$	$\times$
MC + Riemann Sums	$\times$	$\checkmark$	$\checkmark$	$\checkmark$
Bayesian Quadrature	$\times$	$\times$	$\checkmark$?	$\checkmark$
MC(/MCMC) Control Functionals	$\checkmark(/\times)$	$\times(/\checkmark)$	$\checkmark$	$\checkmark$

A Survey of Approaches to Estimate $\mu(f) = \int f d\pi$

Estimation Method	Unbiased	π intractable	Sub-root-n	Post-hoc
Monte Carlo (MC)	$\checkmark$	$\times$	$\times$	$\times$
Markov Chain MC (MCMC)	$\times$	$\checkmark$	$\times$	$\times$
MC + Importance Sampling	$\checkmark(/\times)$	$\times(/\checkmark)$	$\times$	$\checkmark$
MCMC + Rao-Blackwellisation	$\times$	$\checkmark$	$\times$	$\checkmark$
MC(/MCMC) + Control Variates	$\checkmark(/\times)$	$\times(/\checkmark)$	$\times$	$\checkmark$
MC + Antithetic Variables	$\checkmark$	$\times$	$\times$	$\times$
MC(/MCMC) + Stratified Sampling	$\checkmark(/\times)$	$\times(/\checkmark)$	$\times$	$\times$
Quasi-MC (QMC)	$\times$	$\times$	$\checkmark$	$\times$
Randomised QMC (RQMC)	$\checkmark$	$\times$	$\checkmark$	$\times$
MC + Riemann Sums	$\times$	$\checkmark$	$\checkmark$	$\checkmark$
Bayesian Quadrature	$\times$	$\times$	$\checkmark$?	$\checkmark$
MC(/MCMC) Control Functionals	$\checkmark(/\times)$	$\times(/\checkmark)$	$\checkmark$	$\checkmark$

Getting Warmed Up: Control Variates to Control Functionals

Control Variates

- Suppose we have some statistic $\boldsymbol{u}(\boldsymbol{X})$ such that $\mathbb{E}[\boldsymbol{u}(\boldsymbol{X})] = \mathbf{0}$. e.g. $\boldsymbol{u}(\boldsymbol{x}) = \nabla \log \pi(\boldsymbol{x})$ is the score function.
- Construct the control variate estimator:

$$\hat{\mu}_{\mathrm{CV}} = \frac{1}{n} \sum_{i=1}^{n} f(\boldsymbol{X}_i) + \boldsymbol{a}^T \boldsymbol{u}(\boldsymbol{X}_i)$$

  where $\boldsymbol{a} \in \mathbb{R}^d$ are constants. Clearly unbiased.
- When $\boldsymbol{a}$ is chosen optimally

$$\frac{\mathbb{V}[\hat{\mu}_{\mathrm{CV}}]}{\mathbb{V}[\overline{\mu}]} = 1 - \mathrm{Corr}[f(\boldsymbol{X}), \boldsymbol{a}^T \boldsymbol{u}(\boldsymbol{X})]^2$$

- For linear $f(x) = x$ and Gaussian $\pi(x)$, this estimator has zero variance.
- In general convergence remains $O(n^{-1/2})$.

Control Variates

- Suppose we have some statistic $\boldsymbol{u}(\boldsymbol{X})$ such that $\mathbb{E}[\boldsymbol{u}(\boldsymbol{X})] = \mathbf{0}$. e.g. $\boldsymbol{u}(\boldsymbol{x}) = \nabla \log \pi(\boldsymbol{x})$ is the score function.
- Construct the control variate estimator:

$$\hat{\mu}_{\mathrm{CV}} = \frac{1}{n} \sum_{i=1}^{n} f(\boldsymbol{X}_i) + \boldsymbol{a}^T \boldsymbol{u}(\boldsymbol{X}_i)$$

  where $\boldsymbol{a} \in \mathbb{R}^d$ are constants. Clearly unbiased.
- When $\boldsymbol{a}$ is chosen optimally

$$\frac{\mathbb{V}[\hat{\mu}_{\mathrm{CV}}]}{\mathbb{V}[\overline{\mu}]} = 1 - \mathrm{Corr}[f(\boldsymbol{X}), \boldsymbol{a}^T \boldsymbol{u}(\boldsymbol{X})]^2$$

- For linear $f(x) = x$ and Gaussian $\pi(x)$, this estimator has zero variance.
- In general convergence remains $O(n^{-1/2})$.

Control Variates

- Suppose we have some statistic $\boldsymbol{u}(\boldsymbol{X})$ such that $\mathbb{E}[\boldsymbol{u}(\boldsymbol{X})] = \mathbf{0}$. e.g. $\boldsymbol{u}(\boldsymbol{x}) = \nabla \log \pi(\boldsymbol{x})$ is the score function.
- Construct the control variate estimator:

$$\hat{\mu}_{\mathrm{CV}} = \frac{1}{n} \sum_{i=1}^{n} f(\boldsymbol{X}_i) + \boldsymbol{a}^T \boldsymbol{u}(\boldsymbol{X}_i)$$

  where $\boldsymbol{a} \in \mathbb{R}^d$ are constants. Clearly unbiased.
- When $\boldsymbol{a}$ is chosen optimally

$$\frac{\mathbb{V}[\hat{\mu}_{\mathrm{CV}}]}{\mathbb{V}[\overline{\mu}]} = 1 - \mathrm{Corr}[f(\boldsymbol{X}), \boldsymbol{a}^T \boldsymbol{u}(\boldsymbol{X})]^2$$

- For linear $f(x) = x$ and Gaussian $\pi(x)$, this estimator has zero variance.
- In general convergence remains $O(n^{-1/2})$.

Control Variates

- Suppose we have some statistic $\boldsymbol{u}(\boldsymbol{X})$ such that $\mathbb{E}[\boldsymbol{u}(\boldsymbol{X})] = \mathbf{0}$. e.g. $\boldsymbol{u}(\boldsymbol{x}) = \nabla \log \pi(\boldsymbol{x})$ is the score function.
- Construct the control variate estimator:

$$\hat{\mu}_{\mathrm{CV}} = \frac{1}{n} \sum_{i=1}^{n} f(\boldsymbol{X}_i) + \boldsymbol{a}^T \boldsymbol{u}(\boldsymbol{X}_i)$$

  where $\boldsymbol{a} \in \mathbb{R}^d$ are constants. Clearly unbiased.
- When $\boldsymbol{a}$ is chosen optimally

$$\frac{\mathbb{V}[\hat{\mu}_{\mathrm{CV}}]}{\mathbb{V}[\overline{\mu}]} = 1 - \mathrm{Corr}[f(\boldsymbol{X}), \boldsymbol{a}^T \boldsymbol{u}(\boldsymbol{X})]^2$$

- For linear $f(x) = x$ and Gaussian $\pi(x)$, this estimator has zero variance.
- In general convergence remains $O(n^{-1/2})$.

Control Variates

- Suppose we have some statistic $\boldsymbol{u}(\boldsymbol{X})$ such that $\mathbb{E}[\boldsymbol{u}(\boldsymbol{X})] = \mathbf{0}$. e.g. $\boldsymbol{u}(\boldsymbol{x}) = \nabla \log \pi(\boldsymbol{x})$ is the score function.
- Construct the control variate estimator:

$$\hat{\mu}_{\text{CV}} = \frac{1}{n} \sum_{i=1}^{n} f(\boldsymbol{X}_i) + \boldsymbol{a}^T \boldsymbol{u}(\boldsymbol{X}_i)$$

  where $\boldsymbol{a} \in \mathbb{R}^d$ are constants. Clearly unbiased.
- When $\boldsymbol{a}$ is chosen optimally

$$\frac{\mathbb{V}[\hat{\mu}_{\text{CV}}]}{\mathbb{V}[\overline{\mu}]} = 1 - \text{Corr}[f(\boldsymbol{X}), \boldsymbol{a}^T \boldsymbol{u}(\boldsymbol{X})]^2$$

- For linear $f(x) = x$ and Gaussian $\pi(x)$, this estimator has zero variance.
- In general convergence remains $O(n^{-1/2})$.

Control Variates

$$\hat{\mu}_{\mathrm{CV}} = \frac{1}{n}\sum_{i=1}^{n} \underbrace{f(\boldsymbol{X}_i) + \boldsymbol{a}^T \boldsymbol{u}(\boldsymbol{X}_i)}_{\tilde{f}(\boldsymbol{X}_i)}$$

We can interpret control variates as replacing f by a function $\tilde{f}$ such that

1. $\mathbb{E}[\tilde{f}(\boldsymbol{X})] = \mathbb{E}[f(\boldsymbol{X})]$, and
2. $\mathbb{V}[\tilde{f}(\boldsymbol{X})] \leq \mathbb{V}[f(\boldsymbol{X})]$.

Control Variates

$$\hat{\mu}_{\mathrm{CV}} = \frac{1}{n} \sum_{i=1}^{n} \underbrace{f(\boldsymbol{X}_i) + \boldsymbol{a}^T \boldsymbol{u}(\boldsymbol{X}_i)}_{\tilde{f}(\boldsymbol{X}_i)}$$

We can interpret control variates as replacing f by a function $\tilde{f}$ such that

1. $\mathbb{E}[\tilde{f}(\boldsymbol{X})] = \mathbb{E}[f(\boldsymbol{X})]$, and
2. $\mathbb{V}[\tilde{f}(\boldsymbol{X})] \leq \mathbb{V}[f(\boldsymbol{X})]$.

Control Functionals

A very general approach is to take

$$\tilde{f}(\boldsymbol{X}) = f(\boldsymbol{X}) - \hat{f}(\boldsymbol{X}) + \underbrace{\mathbb{E}[\hat{f}(\boldsymbol{X})]}_{\text{known}}$$

where $\hat{f}$ is a tractable approximation to f. Clearly unbiased.

Then

$$\mathbb{V}[\hat{\mu}_{\mathrm{CF}}] = \mathbb{V}\left(\frac{1}{n}\sum_{i=1}^{n}\tilde{f}(\boldsymbol{X}_i)\right) = \frac{1}{n}\mathbb{V}[f(\boldsymbol{X}) - \hat{f}(\boldsymbol{X})].$$

KEY IDEA: "Let the approximation $\hat{f} \approx f$ get better with n." i.e. $\mathbb{V}[f(\boldsymbol{X}) - \hat{f}(\boldsymbol{X})] \to 0$ as $n \to \infty$.

Called "control functionals" (CFs).

Control Functionals

A very general approach is to take

$$\tilde{f}(\boldsymbol{X}) = f(\boldsymbol{X}) - \hat{f}(\boldsymbol{X}) + \underbrace{\mathbb{E}[\hat{f}(\boldsymbol{X})]}_{\text{known}}$$

where $\hat{f}$ is a tractable approximation to f. Clearly unbiased.

Then

$$\mathbb{V}[\hat{\mu}_{\mathrm{CF}}] = \mathbb{V}\left(\frac{1}{n}\sum_{i=1}^{n}\tilde{f}(\boldsymbol{X}_i)\right) = \frac{1}{n}\mathbb{V}[f(\boldsymbol{X}) - \hat{f}(\boldsymbol{X})].$$

KEY IDEA: "Let the approximation $\hat{f} \approx f$ get better with n." i.e. $\mathbb{V}[f(\boldsymbol{X}) - \hat{f}(\boldsymbol{X})] \to 0$ as $n \to \infty$.

Called "control functionals" (CFs).

Control Functionals

A very general approach is to take

$$\tilde{f}(\boldsymbol{X}) = f(\boldsymbol{X}) - \hat{f}(\boldsymbol{X}) + \underbrace{\mathbb{E}[\hat{f}(\boldsymbol{X})]}_{\text{known}}$$

where $\hat{f}$ is a tractable approximation to f. Clearly unbiased.

Then

$$\mathbb{V}[\hat{\mu}_{\mathrm{CF}}] = \mathbb{V}\left(\frac{1}{n}\sum_{i=1}^{n}\tilde{f}(\boldsymbol{X}_i)\right) = \frac{1}{n}\mathbb{V}[f(\boldsymbol{X}) - \hat{f}(\boldsymbol{X})].$$

KEY IDEA: "Let the approximation $\hat{f} \approx f$ get better with n." i.e. $\mathbb{V}[f(\boldsymbol{X}) - \hat{f}(\boldsymbol{X})] \to 0$ as $n \to \infty$.

Called "control functionals" (CFs).

Control Functionals

A very general approach is to take

$$\tilde{f}(\boldsymbol{X}) = f(\boldsymbol{X}) - \hat{f}(\boldsymbol{X}) + \underbrace{\mathbb{E}[\hat{f}(\boldsymbol{X})]}_{\text{known}}$$

where $\hat{f}$ is a tractable approximation to f. Clearly unbiased.

Then

$$\mathbb{V}[\hat{\mu}_{\text{CF}}] = \mathbb{V}\left(\frac{1}{n}\sum_{i=1}^{n}\tilde{f}(\boldsymbol{X}_i)\right) = \frac{1}{n}\mathbb{V}[f(\boldsymbol{X}) - \hat{f}(\boldsymbol{X})].$$

KEY IDEA: "Let the approximation $\hat{f} \approx f$ get better with n." i.e. $\mathbb{V}[f(\boldsymbol{X}) - \hat{f}(\boldsymbol{X})] \to 0$ as $n \to \infty$.

Called "control functionals" (CFs).

Control Functionals

A very general approach is to take

$$\tilde{f}(\boldsymbol{X}) = f(\boldsymbol{X}) - \hat{f}(\boldsymbol{X}) + \underbrace{\mathbb{E}[\hat{f}(\boldsymbol{X})]}_{\text{known}}$$

where $\hat{f}$ is a tractable approximation to f. Clearly unbiased.

Then

$$\mathbb{V}[\hat{\mu}_{\text{CF}}] = \mathbb{V}\left(\frac{1}{n}\sum_{i=1}^{n}\tilde{f}(\boldsymbol{X}_i)\right) = \frac{1}{n}\mathbb{V}[f(\boldsymbol{X}) - \hat{f}(\boldsymbol{X})].$$

KEY IDEA: "Let the approximation $\hat{f} \approx f$ get better with n." i.e. $\mathbb{V}[f(\boldsymbol{X}) - \hat{f}(\boldsymbol{X})] \to 0$ as $n \to \infty$.

Called "control functionals" (CFs).

Automatic Control Functionals

We want to consider f unknown and π intractable, so we need to elicit $\hat{f}$ in an automatic way.

- Consider the class of functions induced by the Stein operator $\hat{f}_\phi : \mathcal{X} \to \mathbb{R}$ of the form

$$\hat{f}_\phi(\boldsymbol{x}) \quad := \quad c + \nabla_{\boldsymbol{x}} \cdot \phi(\boldsymbol{x}) + \phi(\boldsymbol{x}) \cdot \boldsymbol{u}(\boldsymbol{x})$$

  where $\phi : \mathcal{X} \to \mathbb{R}$ is a differentiable function to be specified and $\boldsymbol{u}(\boldsymbol{x}) = \nabla \log \pi(\boldsymbol{x})$ is the score function. (This is a **big** class.)

- From integration by parts we have that

$$[\pi(\boldsymbol{x})\phi(\boldsymbol{x})]_{\boldsymbol{x} \in \partial\mathcal{X}} = \boldsymbol{0} \quad \Rightarrow \quad \mathbb{E}[\hat{f}_\phi(\boldsymbol{X})] = c$$

  so that the approximation $\hat{f}_\phi(\boldsymbol{X})$ is tractable (under suitable boundary conditions).

Automatic Control Functionals

We want to consider f unknown and π intractable, so we need to elicit $\hat{f}$ in an automatic way.

- Consider the class of functions induced by the Stein operator $\hat{f}_\phi : \mathcal{X} \to \mathbb{R}$ of the form

$$\hat{f}_\phi(\boldsymbol{x}) \quad := \quad c + \nabla_{\boldsymbol{x}} \cdot \phi(\boldsymbol{x}) + \phi(\boldsymbol{x}) \cdot \boldsymbol{u}(\boldsymbol{x})$$

  where $\phi : \mathcal{X} \to \mathbb{R}$ is a differentiable function to be specified and $\boldsymbol{u}(\boldsymbol{x}) = \nabla \log \pi(\boldsymbol{x})$ is the score function. (This is a **big** class.)

- From integration by parts we have that

$$[\pi(\boldsymbol{x})\phi(\boldsymbol{x})]_{\boldsymbol{x} \in \partial \mathcal{X}} = \boldsymbol{0} \quad \Rightarrow \quad \mathbb{E}[\hat{f}_\phi(\boldsymbol{X})] = c$$

  so that the approximation $\hat{f}_\phi(\boldsymbol{X})$ is tractable (under suitable boundary conditions).

Automatic Control Functionals

We want to consider f unknown and π intractable, so we need to elicit $\hat{f}$ in an automatic way.

- Consider the class of functions induced by the Stein operator $\hat{f}_{\phi} : \mathcal{X} \to \mathbb{R}$ of the form

$$\hat{f}_{\phi}(\boldsymbol{x}) \quad := \quad c + \nabla_{\boldsymbol{x}} \cdot \phi(\boldsymbol{x}) + \phi(\boldsymbol{x}) \cdot \boldsymbol{u}(\boldsymbol{x})$$

  where $\phi : \mathcal{X} \to \mathbb{R}$ is a differentiable function to be specified and $\boldsymbol{u}(\boldsymbol{x}) = \nabla \log \pi(\boldsymbol{x})$ is the score function. (This is a **big** class.)

- From integration by parts we have that

$$[\pi(\boldsymbol{x})\phi(\boldsymbol{x})]_{\boldsymbol{x} \in \partial\mathcal{X}} = \boldsymbol{0} \quad \Rightarrow \quad \mathbb{E}[\hat{f}_{\phi}(\boldsymbol{X})] = c$$

  so that the approximation $\hat{f}_{\phi}(\boldsymbol{X})$ is tractable (under suitable boundary conditions).

Automatic Control Functionals

- We should try to specify c and ϕ in such a way that $\hat{f}_\phi$ will minimise the variance

$$\sigma_\phi^2 := \mathbb{V}[f(\boldsymbol{X}) - \hat{f}_\phi(\boldsymbol{X})] = \int (f - \hat{f}_\phi)^2 \pi(d\boldsymbol{x}).$$

- This corresponds exactly to fitting a functional regression of the form

$$f(\boldsymbol{x}) = \mu + \nabla_{\boldsymbol{x}} \cdot \phi(\boldsymbol{x}) + \phi(\boldsymbol{x}) \cdot \boldsymbol{u}(\boldsymbol{x}) + \epsilon_\phi(\boldsymbol{x})$$

where the aim is to minimise the expected mean square error $\int \epsilon_\phi^2 \pi(d\boldsymbol{x}) = \sigma_\phi^2$.

- Functional regression, inverse problems, etc.

Automatic Control Functionals

- We should try to specify c and ϕ in such a way that $\hat{f}_\phi$ will minimise the variance

$$\sigma_\phi^2 := \mathbb{V}[f(\boldsymbol{X}) - \hat{f}_\phi(\boldsymbol{X})] = \int (f - \hat{f}_\phi)^2 \pi(d\boldsymbol{x}).$$

- This corresponds exactly to fitting a functional regression of the form

$$f(\boldsymbol{x}) = \mu + \nabla_{\boldsymbol{x}} \cdot \phi(\boldsymbol{x}) + \phi(\boldsymbol{x}) \cdot \boldsymbol{u}(\boldsymbol{x}) + \epsilon_\phi(\boldsymbol{x})$$

  where the aim is to minimise the expected mean square error $\int \epsilon_\phi^2 \pi(d\boldsymbol{x}) = \sigma_\phi^2$.

- Functional regression, inverse problems, etc.

Automatic Control Functionals

- We should try to specify c and ϕ in such a way that $\hat{f}_\phi$ will minimise the variance

$$\sigma_\phi^2 := \mathbb{V}[f(\boldsymbol{X}) - \hat{f}_\phi(\boldsymbol{X})] = \int (f - \hat{f}_\phi)^2 \pi(d\boldsymbol{x}).$$

- This corresponds exactly to fitting a functional regression of the form

$$f(\boldsymbol{x}) = \mu + \nabla_{\boldsymbol{x}} \cdot \phi(\boldsymbol{x}) + \phi(\boldsymbol{x}) \cdot \boldsymbol{u}(\boldsymbol{x}) + \epsilon_\phi(\boldsymbol{x})$$

  where the aim is to minimise the expected mean square error $\int \epsilon_\phi^2 \pi(d\boldsymbol{x}) = \sigma_\phi^2$.

- Functional regression, inverse problems, etc.

Summary of Control Functionals

1. Obtain samples $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ from π.
2. Split these into two sets, $\mathcal{D}_0$ of size m and $\mathcal{D}_1$ of size $n - m$.
3. Use $\mathcal{D}_0$ to construct an approximation $\hat{f}_\phi$ to f, via estimating ϕ.
4. Use $\mathcal{D}_1$ to evaluate an average

$$\hat{\mu}_\phi(\mathcal{D}_1) = \underbrace{\mathbb{E}[\hat{f}_\phi(\boldsymbol{X})]}_{\text{known}} + \frac{1}{n-m} \sum_{i=m+1}^{n} [f(\boldsymbol{X}) - \hat{f}_\phi(\boldsymbol{X}_i)].$$

Summary of Control Functionals

1. Obtain samples $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ from π.
2. Split these into two sets, $\mathcal{D}_0$ of size m and $\mathcal{D}_1$ of size $n - m$.
3. Use $\mathcal{D}_0$ to construct an approximation $\hat{f}_\phi$ to f, via estimating ϕ.
4. Use $\mathcal{D}_1$ to evaluate an average

$$\hat{\mu}_\phi(\mathcal{D}_1) = \underbrace{\mathbb{E}[\hat{f}_\phi(\boldsymbol{X})]}_{\text{known}} + \frac{1}{n-m} \sum_{i=m+1}^{n} [f(\boldsymbol{X}) - \hat{f}_\phi(\boldsymbol{X}_i)].$$

Summary of Control Functionals

1. Obtain samples $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ from π.
2. Split these into two sets, $\mathcal{D}_0$ of size m and $\mathcal{D}_1$ of size $n - m$.
3. Use $\mathcal{D}_0$ to construct an approximation $\hat{f}_\phi$ to f, via estimating ϕ.
4. Use $\mathcal{D}_1$ to evaluate an average

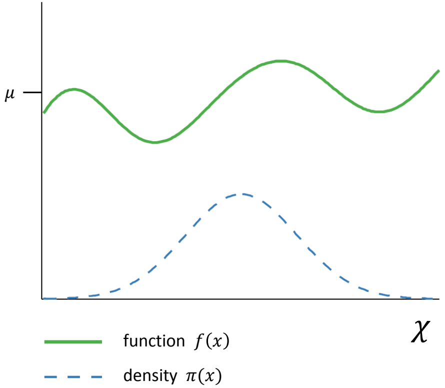
$$\hat{\mu}_\phi(\mathcal{D}_1) = \underbrace{\mathbb{E}[\hat{f}_\phi(\boldsymbol{X})]}_{\text{known}} + \frac{1}{n-m} \sum_{i=m+1}^{n} [f(\boldsymbol{X}) - \hat{f}_\phi(\boldsymbol{X}_i)].$$

Summary of Control Functionals

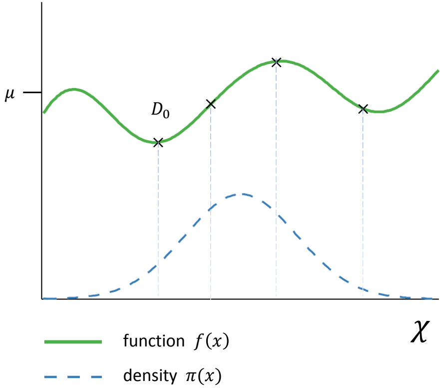
1. Obtain samples $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ from π.
2. Split these into two sets, $\mathcal{D}_0$ of size m and $\mathcal{D}_1$ of size $n - m$.
3. Use $\mathcal{D}_0$ to construct an approximation $\hat{f}_\phi$ to f, via estimating ϕ.
4. Use $\mathcal{D}_1$ to evaluate an average

$$\hat{\mu}_\phi(\mathcal{D}_1) = \underbrace{\mathbb{E}[\hat{f}_\phi(\boldsymbol{X})]}_{\text{known}} + \frac{1}{n-m} \sum_{i=m+1}^{n} [f(\boldsymbol{X}) - \hat{f}_\phi(\boldsymbol{X}_i)].$$

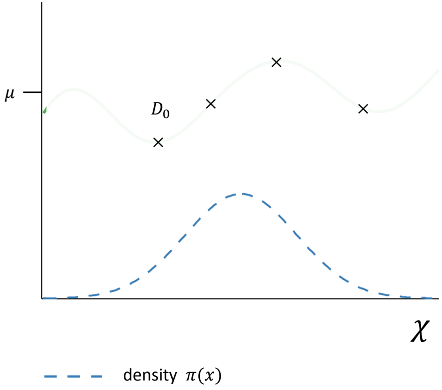
Summary of Control Functionals



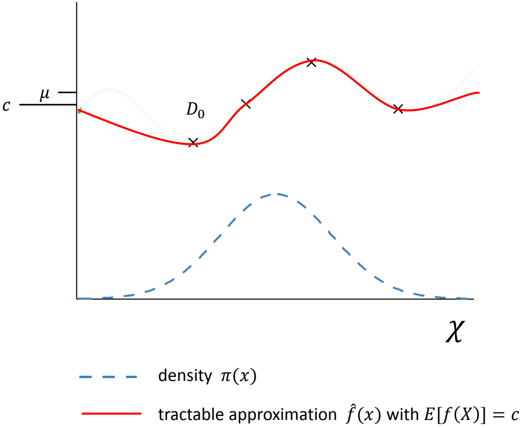
Summary of Control Functionals



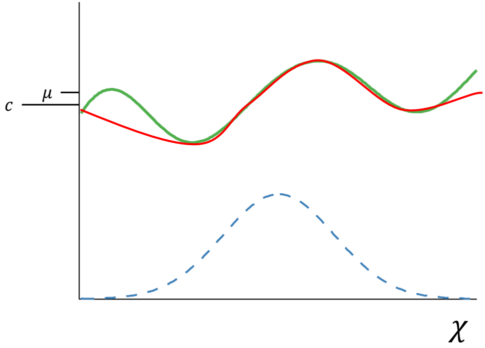
Summary of Control Functionals



Summary of Control Functionals



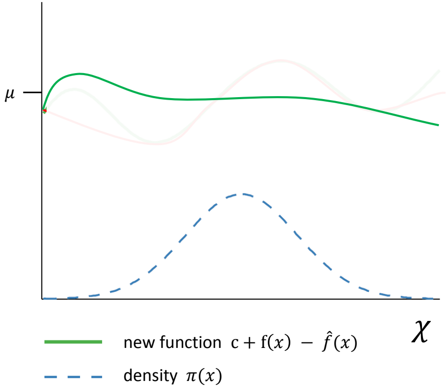
Summary of Control Functionals



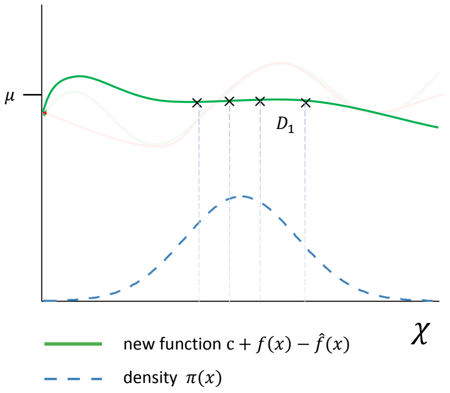
density $\pi(x)$

tractable approximation $\hat{f}(x)$ with $E[f(X)] = c$

Summary of Control Functionals

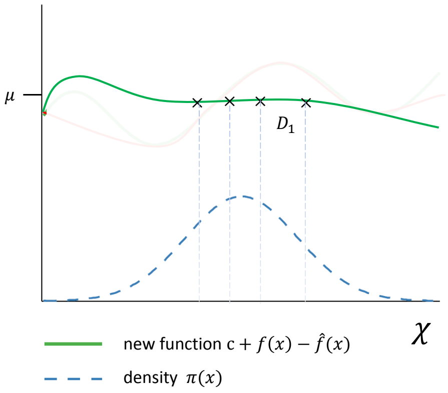


Summary of Control Functionals



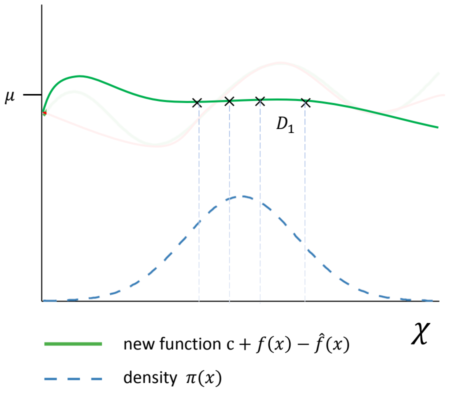
- The "design points" $\mathcal{D}_0$ need not arise from π.
- Unbiased estimation only requires that $\mathcal{D}_1$ arise from π.
- Convergence at rate $o(n^{-1/2})$.

Summary of Control Functionals



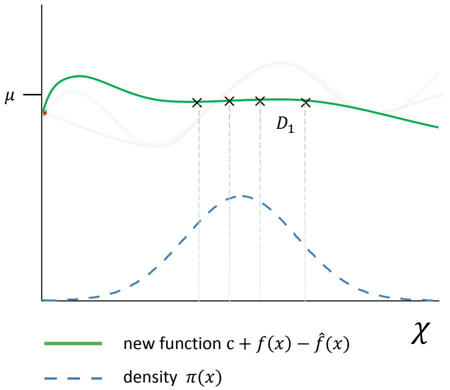
- The "design points" $\mathcal{D}_0$ need not arise from π.
- Unbiased estimation only requires that $\mathcal{D}_1$ arise from π.
- Convergence at rate $o(n^{-1/2})$.

Summary of Control Functionals



- The "design points" $\mathcal{D}_0$ need not arise from π.
- Unbiased estimation only requires that $\mathcal{D}_1$ arise from π.
- Convergence at rate $o(n^{-1/2})$.

Summary of Control Functionals



- The "design points" $\mathcal{D}_0$ need not arise from π.
- Unbiased estimation only requires that $\mathcal{D}_1$ arise from π.
- Convergence at rate $o(n^{-1/2})$.

Theory for Control Functionals

Theorem (Consistency of Control Functionals)

Suppose $\{\boldsymbol{x}_i\}_{i=1}^n$ arise from a Markov chain that targets a density $\pi(\boldsymbol{x})$.

- *Assume $\mathcal{X}$ is bounded.*
- *Assume $\pi(x)$ is bounded away from 0 on $\mathcal{X}$.*
- *Assume $\pi \in C^{2a+1}(\mathcal{X})$ & $k \in C^{2b+2}(\mathcal{X} \times \mathcal{X})$.*
- *Assume k satisfies "certain boundary conditions".*
- *Assume the Markov chain is uniformly ergodic.*

Then, for $f \in \mathcal{H}_k$, there exists $h > 0$ such that

$$1_{h_n<h}\left(\Pi[f] - \hat{\Pi}[f]\right)^2 \quad = \quad \mathcal{O}_P\left(n^{-1-\frac{2(a \wedge b)}{d}+\epsilon}\right),$$

where $\epsilon > 0$ hides logarithmic factors.

Summary: Overall Estimator

The model-averaged estimator has this form:

$$\hat{\mu}_{\mathcal{D}_0}(\mathcal{D}_1) = \underbrace{\frac{1}{n-m}\mathbf{1}^T(\boldsymbol{f}_1 - \hat{\boldsymbol{f}}_1)}_{(*)} + \frac{\mathbf{1}^T\boldsymbol{K}_0^{-1}\boldsymbol{f}_0}{\mathbf{1}^T\boldsymbol{K}_0^{-1}\mathbf{1}}$$

- $\mathcal{D}_1$ enter only through the term $(*)$.
- Posterior consistency of GPs implies that this term vanishes as $m \to \infty$.

Summary: Overall Estimator

The model-averaged estimator has this form:

$$\hat{\mu}_{\mathcal{D}_0}(\mathcal{D}_1) = \underbrace{\frac{1}{n-m}\mathbf{1}^T(\boldsymbol{f}_1 - \hat{\boldsymbol{f}}_1)}_{(*)} + \frac{\mathbf{1}^T\boldsymbol{K}_0^{-1}\boldsymbol{f}_0}{\mathbf{1}^T\boldsymbol{K}_0^{-1}\mathbf{1}}$$

- $\mathcal{D}_1$ enter only through the term $(*)$.
- Posterior consistency of GPs implies that this term vanishes as $m \to \infty$.

Summary: Overall Estimator

The model-averaged estimator has this form:

$$\hat{\mu}_{\mathcal{D}_0}(\mathcal{D}_1) = \underbrace{\frac{1}{n-m}\mathbf{1}^T(\boldsymbol{f}_1 - \hat{\boldsymbol{f}}_1)}_{(*)} + \frac{\mathbf{1}^T \boldsymbol{K}_0^{-1} \boldsymbol{f}_0}{\mathbf{1}^T \boldsymbol{K}_0^{-1} \mathbf{1}}$$

- $\mathcal{D}_1$ enter only through the term $(*)$.
- Posterior consistency of GPs implies that this term vanishes as $m \to \infty$.

Summary: Overall Estimator

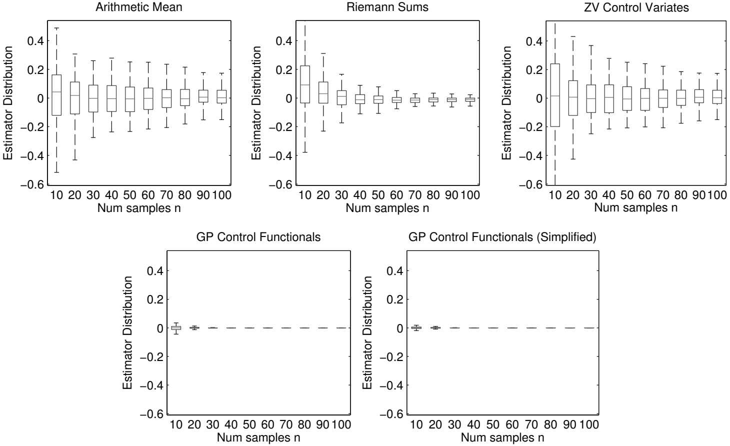
The (simplified) GPCF estimator, in code:

```
% Given samples x(i), scores u(i) and function evaluations f(i)

for i = 1:n
    for j = 1:n
        K(i,j) = d1d2k0(x(i),x(j)) + u(i)*d2k0(x(i),x(j)) ...
                 + u(j)*d1k0(x(i),x(j)) + u(i)*u(j)*k0(x(i),x(j));
    end
end

mu = (ones(1,n) * K \ f) / (ones(1,n) * K \ ones(n,1));
```

Averaging: A Survey of Approaches

Method	AP	f unkno.	π intract.	Unbiased	Convergence
Standard Monte Carlo	$\times$	✓	$\times$	✓	$O(n^{-1/2})$
Markov Chain MC	$\times$	✓	✓	$\times$	$O(n^{-1/2})$
Importance Sampling	$\times$	$\times$	$\times$	✓	$O(n^{-1/2})$
Rao-Blackwellised MCMC	✓	✓	✓	$\times$	$O(n^{-1/2})$
MC + Control Variates	$\times$	✓	$\times$	✓	$O(n^{-1/2})$
MC + Antithetic Variables	$\times$	✓	$\times$	✓	$O(n^{-1/2})$
Stratified Sampling	$\times$	✓	$\times$	✓	$O(n^{-1/2})$
Quasi-MC (QMC)	$\times$	✓	$\times$	$\times$	$O(n^{-1+\epsilon})$
Randomised QMC	$\times$	✓	$\times$	✓	$O(n^{-3/2+\epsilon})$
Riemann Sums	✓	✓	✓	$\times$	$O(n^{-1})$
Bayesian Quadrature	✓	✓	$\times$	$\times$	$o(n^{-1/2})$
Bayesian MC	✓	✓	✓	$\times$	$o(n^{-1/2})$
Control Functionals	✓	✓	✓	✓	$o(n^{-1/2})$

Illustration: $f(x) = \sin(\pi x)$, $X \sim N(0, 1)$



Application : Bayesian Model Selection

Given a competing set $\{M_1, \ldots, M_k\}$ of statistical models, parametrised by possibly different parameter sets $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k\}$.

Apply the Bayesian mantra $p(M_i|\mathbf{y}) \propto p(\mathbf{y}|M_i)p(M_i)$.

Problem: We generally do not know the marginal likelihood

$$p(\mathbf{y}|M_i) = \int p(\mathbf{y}|\boldsymbol{\theta}_i, M_i)p(\boldsymbol{\theta}_i|M_i)d\boldsymbol{\theta}_i$$

in closed form.

In applications, the "obvious approaches" typically do not work well at all.

Application : Bayesian Model Selection

Given a competing set $\{M_1, \ldots, M_k\}$ of statistical models, parametrised by possibly different parameter sets $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k\}$.

Apply the Bayesian mantra $p(M_i|\mathbf{y}) \propto p(\mathbf{y}|M_i)p(M_i)$.

Problem: We generally do not know the marginal likelihood

$$p(\mathbf{y}|M_i) = \int p(\mathbf{y}|\boldsymbol{\theta}_i, M_i)p(\boldsymbol{\theta}_i|M_i)d\boldsymbol{\theta}_i$$

in closed form.

In applications, the "obvious approaches" typically do not work well at all.

Application : Bayesian Model Selection

Given a competing set $\{M_1, \dots, M_k\}$ of statistical models, parametrised by possibly different parameter sets $\{\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_k\}$.

Apply the Bayesian mantra $p(M_i|\boldsymbol{y}) \propto p(\boldsymbol{y}|M_i)p(M_i)$.

Problem: We generally do not know the marginal likelihood

$$p(\boldsymbol{y}|M_i) = \int p(\boldsymbol{y}|\boldsymbol{\theta}_i, M_i)p(\boldsymbol{\theta}_i|M_i)d\boldsymbol{\theta}_i$$

in closed form.

In applications, the "obvious approaches" typically do not work well at all.

Application : Bayesian Model Selection

Given a competing set $\{M_1, \ldots, M_k\}$ of statistical models, parametrised by possibly different parameter sets $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k\}$.

Apply the Bayesian mantra $p(M_i|\mathbf{y}) \propto p(\mathbf{y}|M_i)p(M_i)$.

Problem: We generally do not know the marginal likelihood

$$p(\mathbf{y}|M_i) = \int p(\mathbf{y}|\boldsymbol{\theta}_i, M_i)p(\boldsymbol{\theta}_i|M_i)d\boldsymbol{\theta}_i$$

in closed form.

In applications, the "obvious approaches" typically do not work well at all.

Application : Bayesian Model Selection

One effective solution is called Thermodynamic Integration.

- The "power posterior" for parameters $\boldsymbol{\theta}$ given data $\boldsymbol{y}$ is defined as $p(\boldsymbol{\theta}|\boldsymbol{y}, t) \propto p(\boldsymbol{y}|\boldsymbol{\theta})^t p(\boldsymbol{\theta})$.
- The standard thermodynamic identity is

$$\log p(\boldsymbol{y}) = \int_0^1 \mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{y},t}[\log p(\boldsymbol{y}|\boldsymbol{\theta})]dt.$$

- In TI, this integral is evaluated numerically over a discrete temperature ladder $0 = t_0 < t_1 < \cdots < t_m = 1$. e.g.

$$\widehat{\log p(\boldsymbol{y})} := \sum_{i=0}^{m-1} \frac{(t_{i+1} - t_i)}{2} \{\widehat{\mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{y},t_i}}[\log p(\boldsymbol{y}|\boldsymbol{\theta})] + \widehat{\mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{y},t_{i+1}}}[\log p(\boldsymbol{y}|\boldsymbol{\theta})]\}.$$

- i.e. lots of averages!

Application : Bayesian Model Selection

One effective solution is called Thermodynamic Integration.

- The "power posterior" for parameters $\boldsymbol{\theta}$ given data $\boldsymbol{y}$ is defined as $p(\boldsymbol{\theta}|\boldsymbol{y}, t) \propto p(\boldsymbol{y}|\boldsymbol{\theta})^t p(\boldsymbol{\theta})$.
- The standard thermodynamic identity is

$$\log p(\boldsymbol{y}) = \int_0^1 \mathbb{E}_{\theta|y,t}[\log p(\boldsymbol{y}|\boldsymbol{\theta})]dt.$$

- In TI, this integral is evaluated numerically over a discrete temperature ladder $0 = t_0 < t_1 < \cdots < t_m = 1$. e.g.

$$\widehat{\log p(\boldsymbol{y})} := \sum_{i=0}^{m-1} \frac{(t_{i+1} - t_i)}{2} \{\widehat{\mathbb{E}_{\theta|y,t_i}}[\log p(\boldsymbol{y}|\boldsymbol{\theta})] + \widehat{\mathbb{E}_{\theta|y,t_{i+1}}}[\log p(\boldsymbol{y}|\boldsymbol{\theta})]\}.$$

- i.e. lots of averages!

Application : Bayesian Model Selection

One effective solution is called Thermodynamic Integration.

- The "power posterior" for parameters $\boldsymbol{\theta}$ given data $\boldsymbol{y}$ is defined as $p(\boldsymbol{\theta}|\boldsymbol{y}, t) \propto p(\boldsymbol{y}|\boldsymbol{\theta})^t p(\boldsymbol{\theta})$.
- The standard thermodynamic identity is

$$\log p(\boldsymbol{y}) = \int_0^1 \mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{y},t}[\log p(\boldsymbol{y}|\boldsymbol{\theta})]dt.$$

- In TI, this integral is evaluated numerically over a discrete temperature ladder $0 = t_0 < t_1 < \cdots < t_m = 1$. e.g.

$$\widehat{\log p(\boldsymbol{y})} := \sum_{i=0}^{m-1} \frac{(t_{i+1} - t_i)}{2} \{\widehat{\mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{y},t_i}}[\log p(\boldsymbol{y}|\boldsymbol{\theta})] + \widehat{\mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{y},t_{i+1}}}[\log p(\boldsymbol{y}|\boldsymbol{\theta})]\}.$$

- i.e. lots of averages!

Application : Bayesian Model Selection

One effective solution is called Thermodynamic Integration.

- The "power posterior" for parameters $\boldsymbol{\theta}$ given data $\boldsymbol{y}$ is defined as $p(\boldsymbol{\theta}|\boldsymbol{y}, t) \propto p(\boldsymbol{y}|\boldsymbol{\theta})^t p(\boldsymbol{\theta})$.
- The standard thermodynamic identity is

$$\log p(\boldsymbol{y}) = \int_0^1 \mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{y},t}[\log p(\boldsymbol{y}|\boldsymbol{\theta})]dt.$$

- In TI, this integral is evaluated numerically over a discrete temperature ladder $0 = t_0 < t_1 < \cdots < t_m = 1$. e.g.

$$\widehat{\log p(\boldsymbol{y})} := \sum_{i=0}^{m-1} \frac{(t_{i+1} - t_i)}{2} \{\widehat{\mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{y},t_i}}[\log p(\boldsymbol{y}|\boldsymbol{\theta})] + \widehat{\mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{y},t_{i+1}}}[\log p(\boldsymbol{y}|\boldsymbol{\theta})]\}.$$

- i.e. lots of averages!

Application : Bayesian Model Selection

One effective solution is called Thermodynamic Integration.

- The "power posterior" for parameters $\boldsymbol{\theta}$ given data $\boldsymbol{y}$ is defined as $p(\boldsymbol{\theta}|\boldsymbol{y}, t) \propto p(\boldsymbol{y}|\boldsymbol{\theta})^t p(\boldsymbol{\theta})$.
- The standard thermodynamic identity is

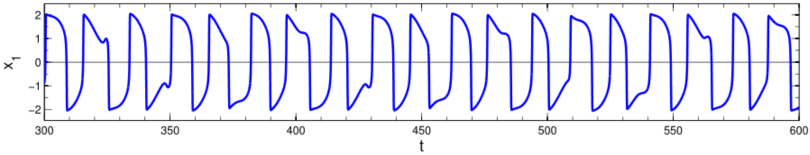
$$\log p(\boldsymbol{y}) = \int_0^1 \mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{y},t}[\log p(\boldsymbol{y}|\boldsymbol{\theta})]dt.$$

- In TI, this integral is evaluated numerically over a discrete temperature ladder $0 = t_0 < t_1 < \cdots < t_m = 1$. e.g.

$$\widehat{\log p(\boldsymbol{y})} := \sum_{i=0}^{m-1} \frac{(t_{i+1} - t_i)}{2} \{\widehat{\mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{y},t_i}}[\log p(\boldsymbol{y}|\boldsymbol{\theta})] + \widehat{\mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{y},t_{i+1}}}[\log p(\boldsymbol{y}|\boldsymbol{\theta})]\}.$$

- i.e. lots of averages!

Application : Bayesian Model Selection



Van der Pol oscillator:

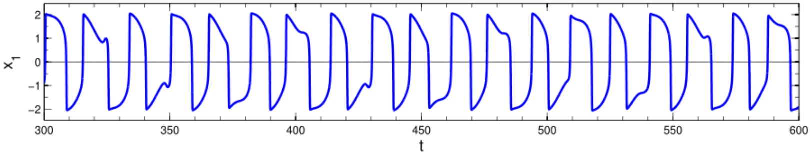
$$\frac{d^2x}{dt^2} - \theta(1 - x^2)\frac{dx}{dt} + x = 0$$

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

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

Sampling is computationally costly.

Application : Bayesian Model Selection



Van der Pol oscillator:

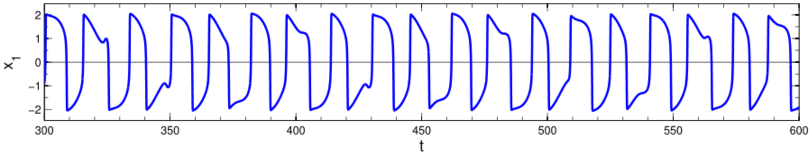
$$\frac{d^2x}{dt^2} - \theta(1 - x^2)\frac{dx}{dt} + x = 0$$

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

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

Sampling is computationally costly.

Application : Bayesian Model Selection



Van der Pol oscillator:

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

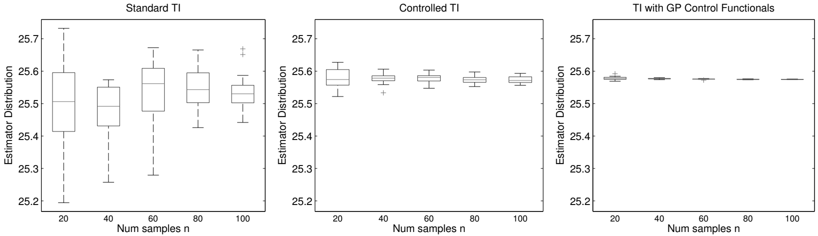
where $\theta \in \mathbb{R}$ is an unknown parameter indicating the non-linearity and the strength of the damping.

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

Sampling is computationally costly.

Application : Bayesian Model Selection

Compare against control variates (CJO *et al.*, JASA 2015):



Control functionals are much more effective!

Conclusions and Discussion

- Averaging is fundamental, but can often be difficult.
- Naive estimators converge slowly, at $O(n^{-1/2})$.
- When f or π are known, we can get better rates using QMC or Bayesian Quadrature.
- **When f and π are both intractable, we can now use control functionals.**
- Extensions of control functionals to QMC in AISTATS 2016

Conclusions and Discussion

- Averaging is fundamental, but can often be difficult.
- Naive estimators converge slowly, at $O(n^{-1/2})$.
- When f or π are known, we can get better rates using QMC or Bayesian Quadrature.
- **When f and π are both intractable, we can now use control functionals.**
- Extensions of control functionals to QMC in AISTATS 2016

Conclusions and Discussion

- Averaging is fundamental, but can often be difficult.
- Naive estimators converge slowly, at $O(n^{-1/2})$.
- When f or π are known, we can get better rates using QMC or Bayesian Quadrature.
- **When f and π are both intractable, we can now use control functionals.**
- Extensions of control functionals to QMC in AISTATS 2016

Conclusions and Discussion

- Averaging is fundamental, but can often be difficult.
- Naive estimators converge slowly, at $O(n^{-1/2})$.
- When f or π are known, we can get better rates using QMC or Bayesian Quadrature.
- **When f and π are both intractable, we can now use control functionals.**
- Extensions of control functionals to QMC in AISTATS 2016

Conclusions and Discussion

- Averaging is fundamental, but can often be difficult.
- Naive estimators converge slowly, at $O(n^{-1/2})$.
- When f or π are known, we can get better rates using QMC or Bayesian Quadrature.
- **When f and π are both intractable, we can now use control functionals.**
- Extensions of control functionals to QMC in AISTATS 2016

Bibliography

Thank you for your attention!

Mira, A., Solgi, R. and Imparato, D. (2013) Zero Variance Markov Chain Monte Carlo for Bayesian Estimators. *Stat. Comput.*, **23**, 653-662.

O'Hagan, A. (1987). Monte Carlo is fundamentally unsound. *J. R. Statist. Soc. D* **36**(2-3), 247-249.

CJO, Papamarkou, T., Girolami, M. (2015) The Controlled Thermodynamic Integral for Bayesian Model Evidence Evaluation. *J. Am. Stat. Assoc.*, to appear.

CJO, Girolami, M., Chopin, N. (2015) Control Functionals for Monte Carlo Integration. *J. R.Statist. Soc. B*, to appear, arXiv 1410.2392.

CJO, Girolami, M. (2015) Variance Reduction for QMC in Reproducing Kernel Hilbert Spaces. To Appear oral *AISTATS 2016*,arXiv 1501.03379.

Briol, F-X., Oates, C. J., Girolami, M. & Osborne, M. A. (2016). Frank-Wolfe Bayesian Quadrature: Probabilistic Integration with Theoretical Guarantees. Advances In Neural Information Processing Systems (NIPS), pages 1162-1170.

Philippe, A. (1997) Processing simulation output by Riemann sums. *J. Statist. Comput. Simul.*, **59**, 295-314.