A guided tour in Monte Carlo

François Portier

Télécom Paris
Institut Polytechnique de Paris

March, 14 2019

Introduction : Why bother with random sampling?

PART 1 : Adaptive importance sampling

- Independent importance sampling
- Adaptive sampling
- Main result
- Illustration

PART 2 : Control variates

- Presentation
- Main result
- Application: GLM with random effects

The underlying integration problem

Let μ be a probability measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ and $\varphi : \mathbb{R}^d \to \mathbb{R}$ be integrable.

- **Goal** : Estimate

$$\mu(\varphi) = \int \varphi \, \mathrm{d}\mu$$

- **Constraint:** only based on $\varphi(x_1), \ldots, \varphi(x_n)$, where $x_1, \ldots, x_n$ are called nodes. Here φ might be black-box function[1].

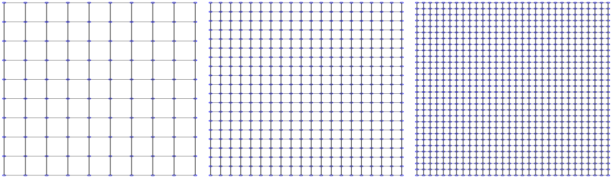
- **Central question:** number of nodes n necessary to obtain a given accuracy

[1] if φ has an explicit form, e.g., $\varphi(x) = \exp(-\|x\|^2)$, then some approximation techniques are probably more appropriate

Riemann's sums method for $\int_{[0,1]^d} \varphi(x)\,\mathrm{d}x$:

$$n^{-d} \sum_{x_i \in \mathrm{Grid}} \varphi(x_i),$$

where $\mathrm{Grid} = \{(i_1/n, \dots, i_d/n) \,:\, 1 \le i_k \le n,\ \forall k = 1, \dots, d\}$



n= 10 n= 20 n= 30

Define

$$\Phi_d = \left\{ \varphi : [0,1]^d \mapsto \mathbb{R} \,:\, |\varphi(x) - \varphi(y)| \le \max_{k=1,\dots,d} |x_k - y_k| \right\}$$

Error bound

We have

$$\sup_{\varphi \in \Phi_d} \left| n^{-d} \sum_{x \in \mathrm{Grid}} \varphi(x) - \int_{[0,1]^d} \varphi(x)\,\mathrm{d}x \right| \le n^{-1}.$$

Consider linear integration rules

$$\sum_{i=1}^{n^d} w_i \varphi(x_i).$$

The accuracy of the best algorithm over a class Φ is

$$e(n^d, \Phi) = \inf_{(w_i, x_i)_{i=1\dots n}} \sup_{\varphi \in \Phi} \left| \sum_{i=1}^{n^d} w_i \varphi(x_i) - \int_{[0,1]^d} \varphi(x)\,\mathrm{d}x \right|$$

Complexity results (Novak, 2016)

$$e(n^d, \Phi_d) = \left(\frac{d}{2d+2} \right) n^{-1}$$

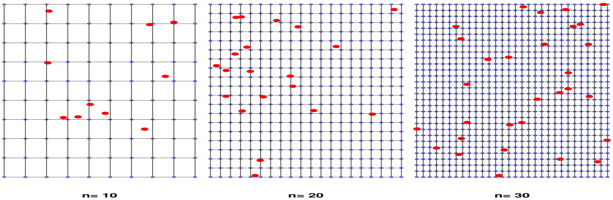
The midpoint rule is the optimal algorithm[2].

[2] If $\Phi_{k,d} = \{\varphi : [0,1]^d \to \mathbb{R}\,,\ \|D_\alpha \varphi\|_\infty \leq 1, \forall |\alpha| \leq k\}$, then $e(n^d, \Phi_{k,d}) \simeq n^{-k}$.

Monte Carlo

Let $(X_1, \ldots, X_n) \overset{iid}{\sim} \mathcal{U}[0,1]^d$, the Monte Carlo estimate of $\int_{[0,1]^d} \varphi(x)\,\mathrm{d}x$ is

$$n^{-1} \sum_{i=1}^{n} \varphi(X_i)$$



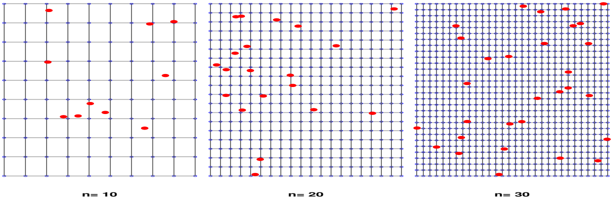
Uniform results (Talagrand, 1996; McDiarmid, 1998; Giné and Guillou, 2001)

with probability larger than $1 - \delta$,

$$\sup_{\varphi \in \Phi} \left| n^{-1} \sum_{i=1}^{n} \varphi(X_i) - \int_{[0,1]^d} \varphi(x)\,\mathrm{d}x \right| \leq 2\mathbb{E}|R_n(\Phi)| + \sqrt{\frac{2\log(2/\delta)}{n}}$$

If for instance, Φ is of VC-type, $\mathbb{E}|R_n(\Phi)| \simeq n^{-1/2}$.

Summary



	determisitic	random	Monte Carlo
$e(n, \Phi_d)$	$n^{-1/d}$	$n^{-1/d}$ $n^{-1/2}$	$n^{-1/2}$
$e(n, \Phi_d^k)$	$n^{-k/d}$	$n^{-k/d}$ $n^{-1/2}$	$n^{-1/2}$

Quasi-Monte Carlo methods provide rates in $n^{-1} \log(n)^{d-1}$ but under more complicated smoothness assumptions (Novak, 2016)

Popular methods

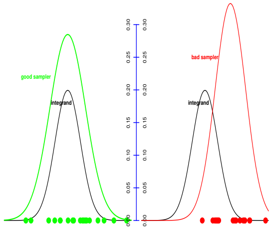
Monte Carlo

1. Draw $X_1, \dots, X_n \overset{iid}{\sim} P$
2. Compute $\frac{1}{n}\sum_{i=1}^{n} \varphi(X_i)$

Control variates

- Use the knowledge of $\mathbb{E}[h_j(X)] = 0$ for functions $h_1, \dots, h_m$

Importance sampling, stratified sampling...



Others

- Quasi-Monte Carlo
- Quadrature rules

Books : Evans and Swartz (2000), Robert and Casella (2004), Glasserman (2003), Owen (2013)

Introduction : Why bother with random sampling?

PART 1 : Adaptive importance sampling

- Independent importance sampling
- Adaptive sampling
- Main result
- Illustration

PART 2 : Control variates

- Presentation
- Main result
- Application: GLM with random effects

The importance sampling game

Let μ be a probability measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ and $\varphi : \mathbb{R}^d \to \mathbb{R}$ be integrable.

- **Goal:** Estimate

$$\mu(\varphi) = \int \varphi \, \mathrm{d}\mu = \int \varphi f \, \mathrm{d}\lambda$$

  where $\mathrm{d}\mu = f \mathrm{d}\lambda$

- Based on

$$\hat{I}_{is}^{(n)}(q) = n^{-1} \sum_{i=1}^{n} \varphi(X_i) \frac{f(X_i)}{q(X_i)}$$

  where $X_1, \ldots, X_n$ are **iid from q**, a density

Importance sampling question

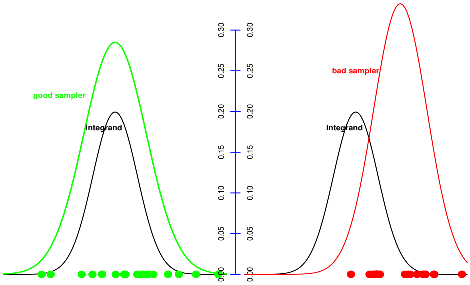
How to choose q?

Basic results

- $\hat{I}_{is}^{(n)}(q)$ is unbiased whenever $\text{supp}(q) \supseteq \text{supp}(\varphi f)$
- The variance is given by

$$\text{Var}(\hat{I}_{is}^{(n)}(q)) = n^{-1} V(\varphi f, q)$$

with $V(\varphi f, q) = \text{Var}_q(\varphi f / q)$



The accuracy heavily depends on the choice of q

Optimal sampler (Evans and Swartz, 2000)

The following holds

1. $$q^* \stackrel{\text{def}}{=} \operatorname*{arg\,min}_{q\,:\,\text{supp}(q) \supseteq \text{supp}(\varphi f)} V(\varphi f, q)$$ is unique

2. $$q^* \propto |\varphi| f$$

3. $$\text{Var}(\hat{I}_{is}^{(n)}(q^*)) = n^{-1} \left\{ \left(\int |\varphi| f \mathrm{d}\lambda \right)^2 - \left(\int \varphi f \mathrm{d}\lambda \right)^2 \right\}$$

Basic method

2-stage parametric importance sampling (Kloek and Van Dijk, 1978)

input: A family of samplers $\mathcal{Q}$ and an initial sampler q_0

- Generate $(X_1^{(1)}, \ldots, X_{n_1}^{(1)}) \overset{iid}{\sim} q_0$
- Compute

$$\hat{q}_1 \in \underset{q \in \mathcal{Q}}{\arg\min}\; n_1^{-1} \sum_{i=1}^{n_1} \frac{\varphi(X_i^{(1)})^2 f(X_i^{(1)})^2}{q(X_i^{(1)}) q_0(X_i^{(1)})}$$

- Generate $(X_1^{(2)}, \ldots, X_{n_2}^{(2)}) \overset{iid}{\sim} \hat{q}_1$ and compute $\hat{I}_{is}^{(n_2)}(\hat{q}_1)$

Adaptive sampling

Goal

- To efficiently visit the space : one must learn from the past action (similar to reinforcement learning) and update the policy at each step

Examples

- Metropolis Hastings (surveyed in Robert (2010))
  - particular MCMC, well suited for Bayesian estimation
  - polynomial complexity in the dimension $\|Q_N - Q^*\|_{tv} \le \epsilon$ whenever $N \ge O(d^2 \log(M/\epsilon))$ (Belloni and Chernozhukov, 2009); concentration inequality (Bertail and Portier, 2018)
  - Adaptive Metropolis (Haario et al., 2001)
- Adaptive/sequential sampling (surveyed in Iba (2001))
  - adaptive importance sampling (Oh and Berger, 1992; Cappé et al., 2004; Douc et al., 2007a; Cornuet et al., 2012)
  - sequential Monte Carlo (Doucet et al., 2001)

Adaptive importance sampling (Oh and Berger, 1992; Cappé et al., 2004; Richard and Zhang, 2007; Douc et al., 2007a,b)

input: A family of samplers $\mathcal{Q}$, an initial sampler $\hat{q}_0 \in \mathcal{Q}$, an allocation policy $(n_t)_{t=1,\dots,T}$

For $t = 1, \dots, T$

- Generate $X_1^{(t)}, \dots, X_{n_t}^{(t)} \overset{iid}{\sim} \hat{q}_{t-1}$ and compute $\hat{I}^{(t)} = \hat{I}_{is}^{(n_t)}(\hat{q}_{t-1})$
- Update:

$$\hat{q}_t = \arg\min_{q \in \mathcal{Q}} \hat{\ell}_{\mathcal{F}_t}(q)$$

where $\hat{\ell}_{\mathcal{F}_t}$ depends on the past particles

---

$$\hat{I}_{ais}^{(T)} = \frac{\sum_{t=1}^{T} n_t \hat{I}_{is}^{(n_t)}(\hat{q}_{t-1})}{\sum_{t=1}^{T} n_t}$$

Choice of the loss

Variance

$$\hat{\ell}_{\mathcal{F}_1}(q) = n_1^{-1} \sum_{i=1}^{n_1} \frac{\varphi(X_i^{(1)})^2 f(X_i^{(1)})^2}{q(X_i^{(1)}) q_0(X_i^{(1)})}$$

$$\ell(q) = \int \varphi^2 f^2 / q \,\mathrm{d}\lambda$$

Kullback-Leibler divergence

$$\hat{\ell}_{\mathcal{F}_1}(q) = -n_1^{-1} \sum_{i=1}^{n_1} \log(q(X_i^{(1)})) \frac{f(X_i^{(1)})}{q_0(X_i^{(1)})}$$

$$\ell(q) = -\int \log(q) f \,\mathrm{d}\lambda$$

Generalized method of moments

$$\hat{\ell}_{\mathcal{F}_1}(q) = \left\| E_q[g] - n_1^{-1} \sum_{i=1}^{n_1} g(X_i^{(1)}) \frac{f(X_i^{(1)})}{q_0(X_i^{(1)})} \right\|_2^2$$

$$\ell(q) = \left\| \int gq \,\mathrm{d}\lambda - \int gf \,\mathrm{d}\lambda \right\|_2^2$$

where $g : \mathbb{R}^d \to \mathbb{R}^q$ is some moment function.

- Previous results obtained when T is fixed and $n_T \to \infty$
- Our framework: $\sum_{t=1}^{T} n_t \to \infty$

Based on 1 simple remark

AIS averages over the terms

$$\frac{\varphi(X_j) f(X_j)}{q_{j-1}(X_j)}, \qquad \text{with } X_j \sim q_{j-1}$$

where j is the sample index and corresponds to $n_1 + \ldots + n_t + i$ for some (t, i)

Define

$$M_n = \sum_{j=1}^{n} \left(\frac{\varphi(X_j) f(X_j)}{q_{j-1}(X_j)} - \int \varphi f \, \mathrm{d}\lambda \right)$$

Property

Assume that for all $1 \leq j \leq n$, the support of q_j contains the support of φf, then the sequence $(M_n, \mathcal{F}_n)$ is a martingale. The quadratic variation of M satisfies $\langle M \rangle_n = \sum_{j=1}^{n} V(\varphi f, q_{j-1})$.

Main result

We consider

$$\text{a loss :} \qquad \ell(q) = \int m_q \,\mathrm{d}\lambda,$$

$$\text{a (parametric) set of samplers :} \qquad \mathcal{Q}$$

Theorem (Delyon and P., 2018)

Under some technical assumptions but without any restriction on $(n_t)_{t=1,\ldots,T}$, as $T \to \infty$,

$$\sqrt{\left(\sum_{t=1}^{T} n_t\right)} \left(\hat{I}_{ais}^{(T)} - \int \varphi f \,\mathrm{d}\lambda\right) \rightsquigarrow \mathcal{N}(0, v^*),$$

where

$$v^* = V(\varphi f, q^*) \quad \text{with} \quad q^* \in \operatorname*{arg\,min}_{q \in \mathcal{Q}} \ell(q)$$

Remark 1: optimality

If $\ell(q) = \int \varphi f / q d\lambda$, then v^* is the best variance that we can achieve over the class of sampler $\mathcal{Q}$

Remark 2: fast rate

Whenever $\varphi > 0$ and $\varphi f / (\int \varphi f d\lambda) \in \mathcal{Q}$,

$$\hat{I}_{ais}^{(T)} - \int \varphi f \, \mathrm{d}\lambda = o_P\left(\left(\sum_{t=1}^{T} n_t\right)^{-1/2}\right)$$

Remark 3: normalized estimates

$$\sum_i \varphi(X_i) \frac{f(X_i)}{q(X_i)} \Big/ \sum_i \frac{f(X_i)}{q(X_i)}$$

are studied as a corollary

A re-weighting to forget bad samplers

Define the weighted estimate, for any function ψ,

$$I_T^{(\alpha)}(\psi) = N_T^{-1} \sum_{t=1}^{T} \alpha_{T,t} \sum_{i=1}^{n_t} \frac{\psi(X_i^{(t)})}{q_{t-1}(X_i^{(t)})}.$$

with $\sum_{t=1}^{T} n_t \alpha_{T,t} = N_T$ (for unbiasedness)

Optimal choice (Douc et al., 2007a)

$$\alpha_{T,t}^{-1} \propto \mathrm{Var}_{q_t}(\varphi f / q_t)$$

Our proposal

$$\alpha_{T,t}^{-1} \propto \mathrm{Var}_{q_t}(f / q_t) \simeq \sum_{i=1}^{n_t} \left(\frac{f(X_i^{(t)})}{q_{t-1}(X_i^{(t)})} - 1 \right)^2$$

Illustration on a toy example

- Aim is to compute $\mu_* = \int x\phi_{\mu_*,\sigma_*}(x)dx$ where $\phi_{\mu,\sigma}$ is the pdf of $\mathcal{N}(\mu, \sigma^2 I_d)$, $\mu_* = (5, \ldots 5)^T \in \mathbb{R}^d$, $\sigma_* = 1$
- $\mathcal{Q}$ the collection of multivariate Student distributions of degree $\nu = 3$ and $\Sigma_0 = 5I_d(\nu - 2)/\nu$, parametrized by the mean
- $q \mapsto \ell(q)$ is the GMM loss
- The initial sampling policy is set as $\mu_0 = (0, \ldots 0) \in \mathbb{R}^d$
- methods in competition : AIS, wAIS and adaptive MH
- For each method that returns μ, the mean squared error (MSE) is computed as the average of $\|\mu - \mu_*\|^2$ computed over 100 replicates of μ

Illustration on a toy example

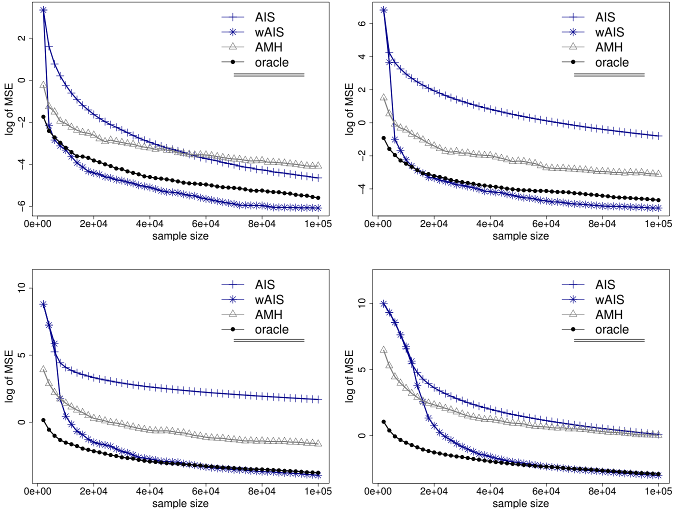


Figure: From left to right $d = 2, 4, 8, 16$. AIS and wAIS are computed with $T = 50$ with constant $n_t = 2e3$. Plotted is the logarithm of the MSE (computed for each method over 100 replicates) with respect to the number of requests to the integrand.

Illustration on a toy example

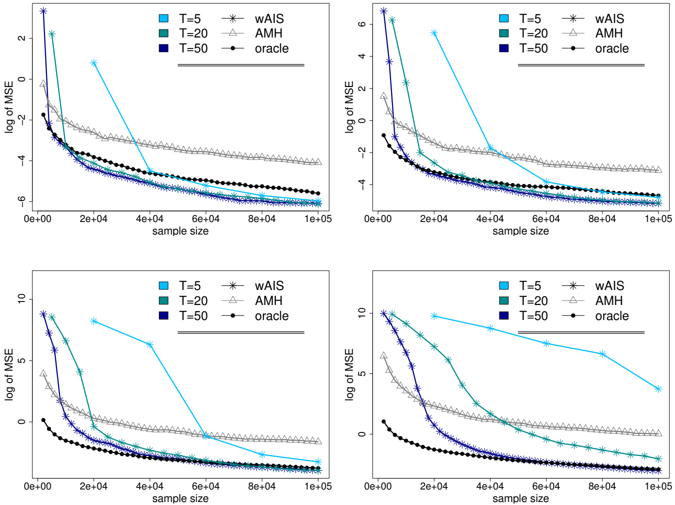


Figure: From left to right $d = 2, 4, 8, 16$. AIS and wAIS are computed with $T = 5, 20, 50$, with a constant allocation policy, resp. $n_t = 2e4, 5e3, 2e3$. Plotted is the logarithm of the MSE (computed for each method over 100 replicates) with respect to the number of requests to the integrand.

Introduction : Why bother with random sampling?

PART 1 : Adaptive importance sampling
- Independent importance sampling
- Adaptive sampling
- Main result
- Illustration

PART 2 : Control variates
- Presentation
- Main result
- Application: GLM with random effects

Let μ be a probability measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ and $\varphi : \mathbb{R}^d \to \mathbb{R}$ be integrable.

The Control variates game

- **Goal:** Estimate

$$\mu(\varphi) = \int \varphi \,\mathrm{d}\mu$$

- **Constraint:** only based on $\varphi(X_1), \ldots, \varphi(X_n)$, where $X_1, \ldots, X_n$ are **iid from μ**

- **New piece of information is available:** $h_1, \ldots, h_m$ test functions such that, for every $\ell = 1, \ldots, m$,

$$\mu(h_k) = \int h_k \,\mathrm{d}\mu \quad \text{is known}$$

Control variates issue

How to use this auxiliary information efficiently?

Control variates method heuristic

Consider the unbiased family

$$\hat{I}_{cv}(\alpha) = n^{-1} \sum_{i=1}^{n} \left\{ \varphi(X_i) - \sum_{k=1}^{m} \alpha_k (h_k(X_i) - \mu(h_k)) \right\}$$

Two steps approach

input : the sample size n, the space $\mathrm{span}(h_1, \ldots, h_m)$

- Step 1. Estimate the **optimal control variate**

$$\alpha \in \operatorname*{arg\,min}_{\alpha \in \mathbb{R}^m} \ \mathrm{var}\left(\varphi - \sum_{k=1}^{m} \alpha_k h_k \right)$$

- Step 2. Compute the modified Monte Carlo estimate

$$\hat{I}_{cv}(\hat{\alpha})$$

Theorem (Glynn and Szechtman, 2002)

Under suitable moments conditions, we have as $n \to \infty$,

$$n^{1/2}\left(\hat{I}_{cv}(\hat{\alpha}) - \int \varphi \, \mathrm{d}\mu\right) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \sigma_m^2)$$

where $\sigma_m^2 = \min_{\alpha\in\mathbb{R}^m} \mathrm{Var}(\varphi - \sum_{k=1}^m \alpha_k h_k) \quad \leq \quad \mathrm{Var}(\varphi)$ (= Monte Carlo variance)

- This applies to 6 different versions of control variates
- The one we promote and study is the OLS version:

$$(\hat{\alpha}_0, \hat{\alpha}) = \operatorname*{arg\,min}_{(\alpha_0,\alpha)\in\mathbb{R}\times\mathbb{R}^m} \sum_{i=1}^n \left(\varphi(X_i) - \alpha_0 - \sum_{k=1}^m \alpha_k h_k(X_i)\right)^2$$

- Among the six control variates, this is the only one that integrates without errors functions $\varphi \in \mathrm{span}(1, h_1, \ldots, h_m)$.
- Linear integration rule : $\hat{\alpha}_0 = \sum_{i=1}^n w_{i,n}\varphi(X_i)$

Growing number of control variates $m = m_n$

Theorem (P. and Segers, 2018)

Under suitable moments conditions, we have as $n \to \infty$, $m_n = o(n^{1/2})$,

$$\left(\frac{n^{1/2}}{\sigma_{m_n}}\right)\left(\hat{\alpha}_0 - \int \varphi \, \mathrm{d}\mu\right) \xrightarrow{\mathrm{d}} \mathcal{N}(0,1)$$

where $\sigma_m^2 = \min_{\alpha \in \mathbb{R}^m} \mathrm{Var}(\varphi - \sum_{k=1}^m \alpha_k h_k)$

Related works

- Oates et al. (2016): control variates taken in a RKHS. They provide a bound on the error when 2 independent samples are used in step 1 and 2.
- Gobet and Surana (2014): sequential approximation of the regression coefficients. Bound when 2 independent samples are used.

Example (The smoother f, the faster the rate)

Suppose that

- Let (h_j) be the Legendre polynomials
- Let f be $k+1$ times continuously differentiable

then $\sigma^2_{m_n} = O(m_n^{-2k-1})$ and

$$\hat{\alpha}_0 - \int \varphi \, \mathrm{d}\mu = O_p(m_n^{-k-1/2} n^{-1/2})$$

Applications

Importance sampling

- random variable generation (Erraqabi et al., 2016)
- Bayesian statistics, e.g., Cornuet et al. (2012)
- option pricing, e.g., Douc et al. (2007a)
- optimization (Hashimoto et al., 2018)
- reinforcement learning (Jie and Abbeel, 2010)

Control variates

- numerical integration, e.g., $\mathbb{E}[\varphi(W_1, W_2)]$ and we know $\mathbb{E}[W_1], \mathbb{E}[W_2]$
- queuing network (Lavenberg and Welch, 1981)
- option pricing (Hull and White, 1988)
- Bayesian statistics e.g., (Oates et al., 2016)
- variance reduction for stochastic gradient descent (Wang et al., 2013)
- latent variable model (P. and Segers, 2018)

Logit model with random effect

Observations $(y_{j,k}, x_{j,k}) \in \{0,1\} \times \mathbb{R}$

- classes $k = 1, \ldots, q$
- observations $j = 1, \ldots, N$ in each class

Model

Random effects $u_1, \ldots, u_q$ iid $\mathcal{N}(0,1)$ (latent) such that

$$y_{j,k} \mid u_1, \ldots, u_q \sim \text{Bernoulli}(p_{j,k})$$
$$\text{logit}(p_{j,k}) = \beta x_{j,k} + \sigma u_k$$

Likelihood proportional to:

$$\prod_{k=1}^{q} \int_{\mathbb{R}} \prod_{j=1}^{N} \left(\frac{e^{y_{j,k}(\beta x_{j,k} + \sigma u)}}{1 + e^{\beta x_{j,k} + \sigma u}} \right) e^{-u^2/2} \, \mathrm{d}u$$

More generally: generalized linear models with random effects (McCulloch and Searle, 2001)

Maximum simulated likelihood

	EM	MC		OLSMC	
n	sd	sd	rMSE	sd	rMSE
100	0.1227	0.1027	0.1027	2e-4	3e-4
500	0.0546	0.0468	0.0467	2e-5	2e-4
1000	0.0388	0.0334	0.0334	3e-6	2e-4

Methods:

- Expectation–Maximization
  - E-step: Monte Carlo
- Monte Carlo
- OLS Monte Carlo
  - change of variables to $[0, 1]$
  - polynomial basis
  - $m = \lfloor 2\sqrt{n} \rfloor$

Artificial data set (Booth and Hobert, 1999)

- $q = 10$ classes
- $N = 15$ observations per class
- $\beta = 5$, $\sigma = 1/2$
- fixed design $x_{j,k} = j/N$
- 200 replications

target: MLE (deterministic integration)

Multinomial logit model with random effects

Booth and Hobert (1999): Medical studies $i = 1, \dots, N$

- n_{i1} (n_{i2}) nb of (non-)smokers
- y_{i1} (y_{i2}) nb of patients with lung cancer among (non-)smokers

Model

Latent random $\mathcal{N}(0,1)$ effects u_i, v_{i1}, v_{i2} such that

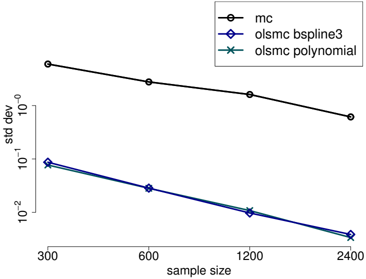
$$y_{ij} \sim \mathrm{Binom}(\pi_{ij}, n_{ij})$$
$$\mathrm{logit}(\pi_{ij}) = \beta_0 + \beta_1 1_{\{j=1\}} + \sigma_u u_i + \sigma_v v_{ij}$$

Likelihood proportional to

$$\prod_{i=1}^{N} \int_{\mathbb{R}^3} b_{i,1}(u, v_1)\, b_{i,2}(u, v_2)\, \phi_{\sigma_u}(u)\, \phi_{\sigma_v}(v_1)\, \phi_{\sigma_v}(v_2)\, \mathrm{d}(u, v_1, v_2)$$

$$\text{where} \qquad \begin{aligned} b_{i,j}(u,v) &= \pi_j(u,v)^{y_{ij}} \left\{1 - \pi_j(u,v)\right\}^{n_{ij}-y_{ij}} \\ \pi_j(u,v) &= \mathrm{logit}^{-1}(\beta_0 + \beta_1 1_{\{j=1\}} + \sigma_u u + \sigma_v v) \end{aligned}$$

Maximum simulated likelihood



$N = 20$ studies
$n_{i1} + n_{i2} = 50$ persons per study
200 replications

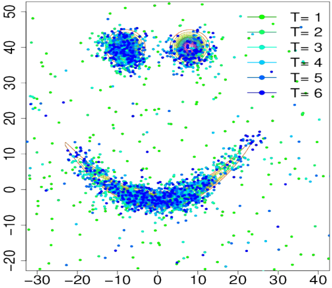
- N integrals on $[0,1]^3$
- cubic B-splines or polynomials
- tensor products
- k functions per dimension

$\Longrightarrow$ $m = (k+1)^3 - 1$ control functions

k	3	4	5	6
m	63	124	215	342
n	300	600	1200	2400

- points X_i and weights $w_{n,i}$ common for all N integrals

Work in progress: AIS with flexible nonparametric methods



References:

- Bertail, P. and Portier, F. (2019). Rademacher complexity for markov chains: Applications to kernel smoothing and metropolis-hasting. To appear in Bernoulli
- Delyon, B. and Portier, F. (2018). Asymptotic optimality of adaptive importance sampling. NIPS18, pp. 3138–3148.
- Portier, F. and Segers, J. (2018). Monte carlo integration with a growing number of control variates. arXiv preprint arXiv:1801.01797.

Bibliography I

Belloni, A. and V. Chernozhukov (2009). On the computational complexity of mcmc-based estimators in large samples. *The Annals of Statistics*, 2011–2055.

Bertail, P. and F. Portier (2018). Rademacher complexity for markov chains: Applications to kernel smoothing and metropolis-hasting. *arXiv preprint arXiv:1806.02107*.

Booth, J. G. and J. P. Hobert (1999). Maximizing generalized linear mixed model likelihoods with an automated monte carlo em algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 61*(1), 265–285.

Cappé, O., A. Guillin, J.-M. Marin, and C. P. Robert (2004). Population monte carlo. *Journal of Computational and Graphical Statistics 13*(4), 907–929.

Cornuet, J.-M., J.-M. Marin, A. Mira, and C. P. Robert (2012). Adaptive multiple importance sampling. *Scandinavian Journal of Statistics 39*(4), 798–812.

Delyon, B. and F. P. (2018). Asymptotic optimality of adaptive importance sampling. In *Advances in Neural Information Processing Systems*, pp. 3138–3148.

Douc, R., A. Guillin, J.-M. Marin, and C. P. Robert (2007a). Convergence of adaptive mixtures of importance sampling schemes. *The Annals of Statistics*, 420–448.

Douc, R., A. Guillin, J.-M. Marin, and C. P. Robert (2007b). Minimum variance importance sampling via population monte carlo. *ESAIM: Probability and Statistics 11*, 427–447.

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

Erraqabi, A., M. Valko, A. Carpentier, and O. Maillard (2016). Pliable rejection sampling. In *International Conference on Machine Learning*, pp. 2121–2129.

Evans, M. and T. Swartz (2000). *Approximating integrals via Monte Carlo and deterministic methods*. Oxford Statistical Science Series. Oxford University Press, Oxford.

Bibliography II

Giné, E. and A. Guillou (2001). On consistency of kernel density estimators for randomly censored data: rates holding uniformly over adaptive intervals. *Ann. Inst. H. Poincaré Probab. Statist. 37*(4), 503–522.

Glasserman, P. (2003). *Monte Carlo Methods in Financial Engineering*. New York: Springer.

Glynn, P. W. and R. Szechtman (2002). Some new perspectives on the method of control variates. In *Monte Carlo and quasi-Monte Carlo methods, 2000 (Hong Kong)*, pp. 27–49. Springer, Berlin.

Gobet, E. and K. Surana (2014). A new sequential algorithm for l2-approximation and application to monte-carlo integration.

Haario, H., E. Saksman, and J. Tamminen (2001). An adaptive metropolis algorithm. *Bernoulli 7*(2), 223–242.

Hashimoto, T. B., S. Yadlowsky, and J. C. Duchi (2018). Derivative free optimization via repeated classification. *arXiv preprint arXiv:1804.03761*.

Hull, J. and A. White (1988). The use of the control variate technique in option pricing. *Journal of Financial and Quantitative analysis 23*(03), 237–251.

Iba, Y. (2001). Population monte carlo algorithms. *Transactions of the Japanese Society for Artificial Intelligence 16*(2), 279–286.

Jie, T. and P. Abbeel (2010). On a connection between importance sampling and the likelihood ratio policy gradient. In *Advances in Neural Information Processing Systems*, pp. 1000–1008.

Kloek, T. and H. K. Van Dijk (1978). Bayesian estimates of equation system parameters: an application of integration by monte carlo. *Econometrica: Journal of the Econometric Society*, 1–19.

Lavenberg, S. S. and P. D. Welch (1981). A perspective on the use of control variables to increase the efficiency of Monte Carlo simulations. *Management Sci. 27*(3), 322–335.

McCulloch, C. E. and S. R. Searle (2001). Generalized, linear, mixed models.

Bibliography III

McDiarmid, C. (1998). Concentration. In *Probabilistic methods for algorithmic discrete mathematics*, Volume 16 of *Algorithms Combin.*, pp. 195–248. Springer, Berlin.

Novak, E. (2016). Some results on the complexity of numerical integration. In *Monte Carlo and Quasi-Monte Carlo Methods*, pp. 161–183. Springer.

Oates, C. J., M. Girolami, and N. Chopin (2016). Control functionals for monte carlo integration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Oh, M.-S. and J. O. Berger (1992). Adaptive importance sampling in Monte Carlo integration. *J. Statist. Comput. Simulation 41*(3-4), 143–168.

Owen, A. B. (2013). Monte Carlo Theory, Methods and Examples. `http://statweb.stanford.edu/~owen/mc/`.

P., F. and J. Segers (2018). Monte carlo integration with a growing number of control variates. *arXiv preprint arXiv:1801.01797*.

Richard, J.-F. and W. Zhang (2007). Efficient high-dimensional importance sampling. *J. Econometrics 141*(2), 1385–1411.

Robert, C. P. (2010). The metropolis–hastings algorithm. *Wiley StatsRef: Statistics Reference Online*.

Robert, C. P. and G. Casella (2004). *Monte Carlo statistical methods* (Second ed.). Springer Texts in Statistics. Springer-Verlag, New York.

Talagrand, M. (1996). New concentration inequalities in product spaces. *Inventiones mathematicae 126*(3), 505–563.

Wang, C., X. Chen, A. J. Smola, and E. P. Xing (2013). Variance reduction for stochastic gradient optimization. In *Advances in Neural Information Processing Systems*, pp. 181–189.