UiO : Matematisk institutt

Det matematisk-naturvitenskapelige fakultet

STK-4051/9051 Computational Statistics Spring 2021
Variance reduction

Instructor: Odd Kolbjørnsen, oddkol@math.uio.no

Additional reference person

- liliana.vazquezfernandez@fhi.no
  (or shorter version: livf@fhi.no).

Today

- Variance reduction
- SMC- retake

Monte Carlo methods

- Aim (following notation from book):

$$\mu = E^{f(\mathbf{X})}[h(\mathbf{X})] = \begin{cases} \int_{\mathbf{x}} h(\mathbf{x})f(\mathbf{x}))d\mathbf{x} & \mathbf{x} \text{ continuous} \\ \sum_{\mathbf{x}} h(\mathbf{x})f(\mathbf{x}) & \mathbf{x} \text{ discrete} \end{cases}$$

- Monte Carlo:
  1. Simulate $\mathbf{X}_i \sim f(\mathbf{x}), i = 1, ..., n$
  2. Approximate μ by

$$\hat{\mu}_{MC} = \frac{1}{n}\sum_{i=1}^{n} h(\mathbf{x}_i)$$

- Properties:
  - Unbiased $E[\hat{\mu}_{MC}] = \mu$
  - If $X_1, ..., X_n$ are independent
    - Variance: $\text{var}[\hat{\mu}_{MC}] = \frac{1}{n}\text{var}[h(\mathbf{X})]$
    - Consistent: $\hat{\mu}_{MC} \rightarrow \mu$ as $n \rightarrow \infty$ if $\text{var}[h(\mathbf{X})] < \infty$
  - Estimate of variance:

$$\widehat{\text{var}}[\hat{\mu}_{MC}] = \frac{1}{n-1}\sum_{i=1}^{n}(h(\mathbf{x}_i) - \hat{\mu}_{MC})^2$$

- Can we do better than this?

Recap

- Exact methods
  - Inversion/transformation methods
  - Rejection sampling
- Approximate methods
  - Sampling importance resampling
  - Sequential Monte Carlo
  - Markov chain Monte Carlo (Chapter 7 and 8)
- Variance reduction methods
  - Importance sampling
  - Antithetic sampling
  - Control variates
  - Rao-blackwellization
  - Common random numbers

Variance Reduction

For importance sampling we have seen two options:

$$\mu = \int h(\mathbf{x})f(\mathbf{x})d\mathbf{x} = \int \frac{h(\mathbf{x})f(\mathbf{x})}{g(\mathbf{x})}g(\mathbf{x})d\mathbf{x} = \frac{\int \frac{h(\mathbf{x})f(\mathbf{x})}{g(\mathbf{x})}g(\mathbf{x})d\mathbf{x}}{\int \frac{f(\mathbf{x})}{g(\mathbf{x})}g(\mathbf{x})d\mathbf{x}}$$

- Assume $X_1, \dots, X_n$ iid from $g(\mathbf{x})$.
- Two alternative estimates

$$\hat{\mu}_{IS}^* = \frac{1}{n}\sum_{i=1}^n h(\mathbf{X}_i)w^*(\mathbf{X}_i), \quad w^*(X_i) = \frac{f(\mathbf{X}_i)}{g(\mathbf{X}_i)}$$

$$\hat{\mu}_{IS} = \sum_{i=1}^n h(\mathbf{X}_i)w(\mathbf{X}_i), \quad w(\mathbf{X}_i) = \frac{w^*(\mathbf{X}_i)}{\sum_{j=1}^n w^*(\mathbf{X}_i)}$$

- $w^*(\mathbf{X}_i)$ called importance weights
- $w(\mathbf{X}_i)$ called the normalized importance weights

We should use normalized importance weight if:

$$\mathrm{cov}[t(\mathbf{X}), w^*(\mathbf{X})] > \frac{\mu \mathrm{var}[w^*(\mathbf{X})]}{2}$$

$$\Updownarrow$$

$$\mathrm{cor}[t(\mathbf{X}), w^*(\mathbf{X})] > \frac{\sqrt{\mathrm{var}[w^*(\mathbf{X})]}}{2\sqrt{\mathrm{var}[t(\mathbf{X})]/\mu}} = \frac{\mathrm{cv}[w^*(\mathbf{X})]}{2\mathrm{cv}[t(\mathbf{X})]}$$

- Example: `imp_samp_beta.R`

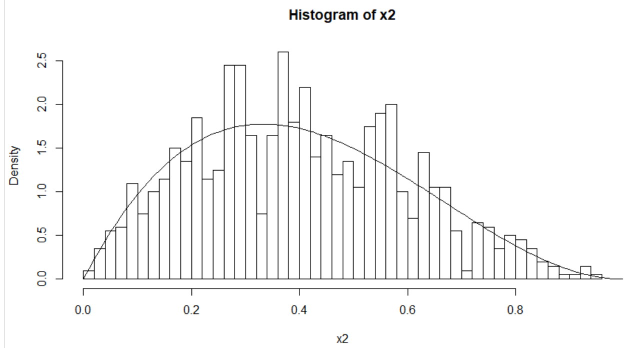
- Taylor approximation of $1/\bar{w}^*$ around 1:

$$\frac{1}{\bar{w}^*} \approx 1 - (\bar{w}^* - 1) + (\bar{w}^* - 1)^2$$

$$t(\mathbf{X}_i) = h(\mathbf{X}_i) w^*(\mathbf{X}_i)$$

Imp_sam_beta.R

- Sample: uniform(0,1)
- Target: beta(2,3)
- Estimate mean



$$\mathrm{cor}[t(\mathbf{X}), w^*(\mathbf{X})] > \frac{\sqrt{\mathrm{var}[w^*(\mathbf{X})]}}{2\sqrt{\mathrm{var}[t(\mathbf{X})]}/\mu} = \frac{\mathrm{cv}[w^*(\mathbf{X})]}{2\mathrm{cv}[t(\mathbf{X})]}$$

$$0.7120781 > \frac{0.6043563}{2*0.6547519} = 0.4615155$$

Bootstrapping to get uncertainty:

	Estimate	SD	RMSE
mu.hat.star	0.399958	0.008107	0.008103
mu.hat	0.399971	0.006126	0.006123

Control variates

- Of interest $\mu = E[h(\mathbf{X})]$
- $\theta = E[c(\mathbf{Y})]$ known
- $(\mathbf{X}, \mathbf{Y})$ correlated (often $\mathbf{Y} = \mathbf{X}$)
- $\{(\mathbf{X}_i, \mathbf{Y}_i), i = 1, ..., N\}$ iid pairs
- Possible alternative estimate

$$
\begin{aligned}
\hat{\mu}_{CV} &= \hat{\mu}_{MC} + \lambda(\hat{\theta}_{MC} - \theta) \\
\hat{\mu}_{MC} &= N^{-1}\sum_{i=1}^{N} h(\mathbf{X}_i) \\
\hat{\theta}_{MC} &= N^{-1}\sum_{i=1}^{N} c(\mathbf{Y}_i)
\end{aligned}
$$

- Properties:

$$
\begin{aligned}
E[\hat{\mu}_{CV}] &= \mu \\
\mathrm{var}[\hat{\mu}_{CV}] &= \mathrm{var}[\hat{\mu}_{MC}] + \lambda^2 \mathrm{var}[\hat{\theta}_{MC}] + 2\lambda \mathrm{cov}[\hat{\mu}_{MC}, \hat{\theta}_{MC}]
\end{aligned}
$$

- Can choose λ to minimize this:

$$
\lambda = -\frac{\mathrm{cov}[\hat{\mu}_{MC}, \hat{\theta}_{MC}]}{\mathrm{var}[\hat{\theta}_{MC}]}
$$

Importance sampling

- Importance sampling:
  - $\mu = E^f[h(\mathbf{X})] = E^g[w^*(\mathbf{X})h(\mathbf{X})]$ with $w^*(\mathbf{x}) = f(\mathbf{x})/g(\mathbf{x})$
  - We known $\theta = E[w^*(\mathbf{X})] = 1$
- Possible alternative estimate

$$\hat{\mu}_{ISCV} = \hat{\mu}^*_{IS} + \lambda(\bar{w}^* - 1) \qquad \lambda = -\frac{\text{cov}[\hat{\mu}_{MC}, \hat{\theta}_{MC}]}{\text{var}[\hat{\theta}_{MC}]}$$

Imp_sam_beta.R $\lambda = -0.3110773$

	Estimate	SD	RMSE
mu.hat.star	0.400116	0.008029	0.008026
mu.hat	0.400106	0.005725	0.005723
mu.hat.cv	0.400083	0.005466	0.005464

Improvement also for:

$\lambda = -0.25 \rightarrow 0.005587$
$\lambda = -0.40 \rightarrow 0.005719$

In example:
Robust towards
estimation of λ

Questions

Rao- Blackwellization

- Of interest $\mu = E[h(\mathbf{X})]$
- Assume $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$ and $E[h(\mathbf{X})|\mathbf{X}_2]$ is known.
- Alternative estimate

$$\hat{\mu}_{RB} = N^{-1}\sum_{i=1}^{N} E[h(\mathbf{X})|\mathbf{X}_2 = \mathbf{x}_{i2}]$$

- Properties:

$$E[\hat{\mu}_{RB}] = E[E[h(\mathbf{X})|\mathbf{X}_2]] = E[h(\mathbf{X})] = \mu$$

law of:
Total Expectation

- Note

$$\text{var}[h(\mathbf{X}_i)] = E[\text{var}[h(\mathbf{X}_i)|\mathbf{X}_2]] + \text{var}[E[h(\mathbf{X})|\mathbf{X}_2]] \geq \text{var}[E[h(\mathbf{X})|\mathbf{X}_2]$$

implying

$$\text{var}[\hat{\mu}_{MC}] \geq \text{var}[\hat{\mu}_{RB}]$$

law of:
Total Variance

Example 1

- Model

$$\begin{aligned} \sigma \sim& \mathrm{Unif}[0,2] \\ X|\sigma \sim& N(0,\sigma) \\ E[X] =& E[E[X|\sigma]] = E[0] = 0 \end{aligned}$$

- What is $\tau = \mathrm{Var}[X] = E[X^2]$?
- Direct MC:
  - Simulate $X_1, ..., X_N$ from model
  - $\hat{\tau}_{MC} = \frac{1}{N}\sum_{i=1}^N X_i^2$
- Rao-Blackwellization:
  - $E[X^2|\sigma^2] = \sigma^2$
  - Simulate $\sigma_1, ..., \sigma_N \sim \mathrm{Unif}[0,2]$
  - $\hat{\tau}_{RB} = \frac{1}{N}\sum_{i=1}^N \sigma_i^2$
- `Example_RB.R`
- Analytically: $\tau = \mathrm{Var}[X] = E[X^2] = 8/6 = 1.333$

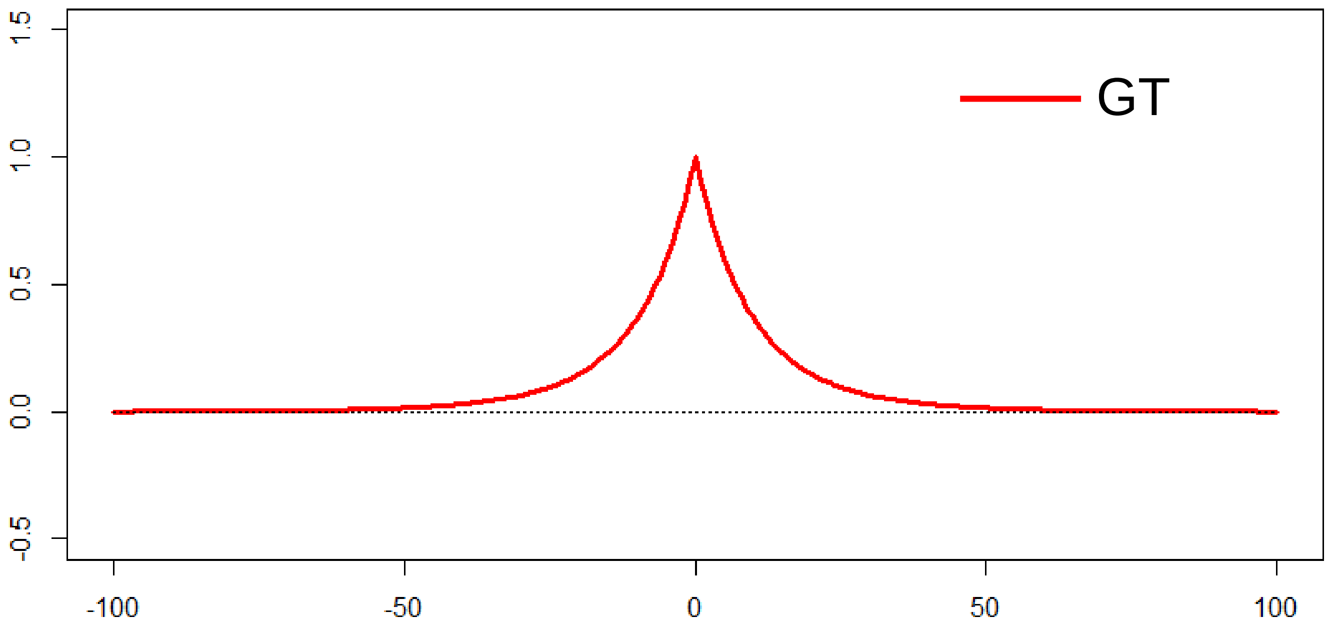
	Estimate	SD
Direct	1.32951	0.008805905
RB	1.33198	0.003777348

Example 2 (spatial statistics)

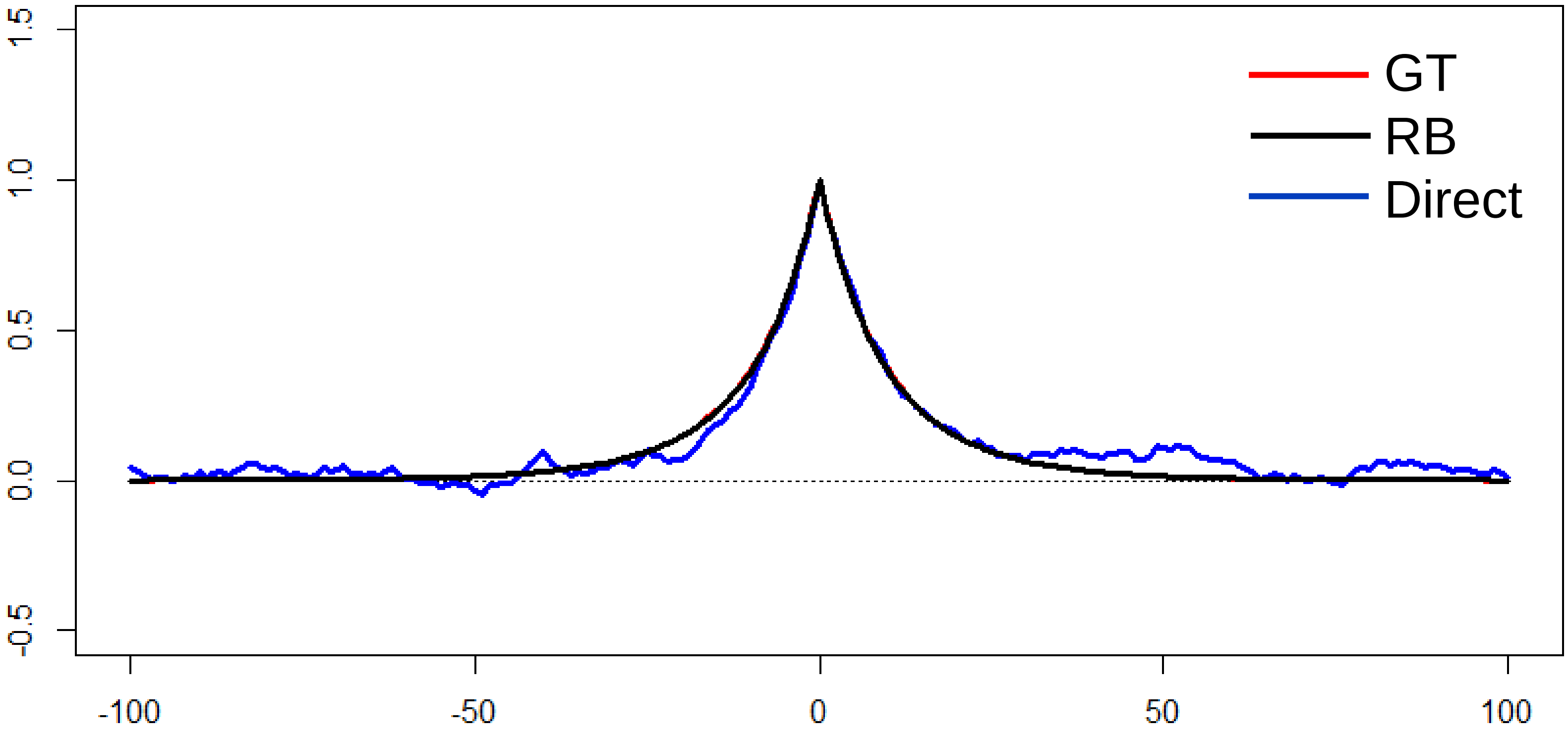
- Spatial dependency:

$$E\{z(x)\} = 0$$
$$\operatorname{cov}\big(z(x), z(x+h)\big) = \rho^h$$
$$\rho \sim \operatorname{unif}[0.85\,, 0.95]$$

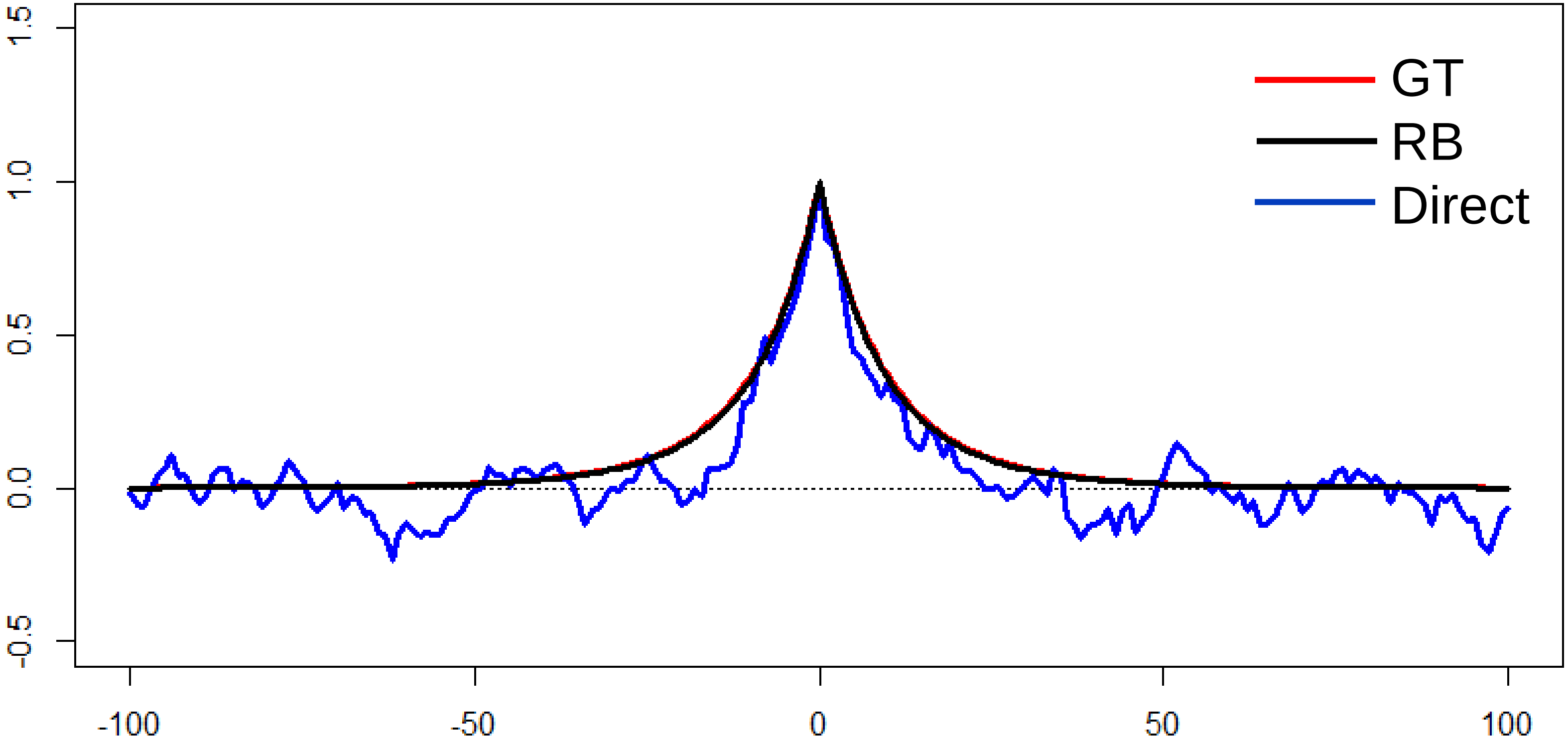
- What is: $\mathrm{E}\{z(x)|z(0) = 1\}$?



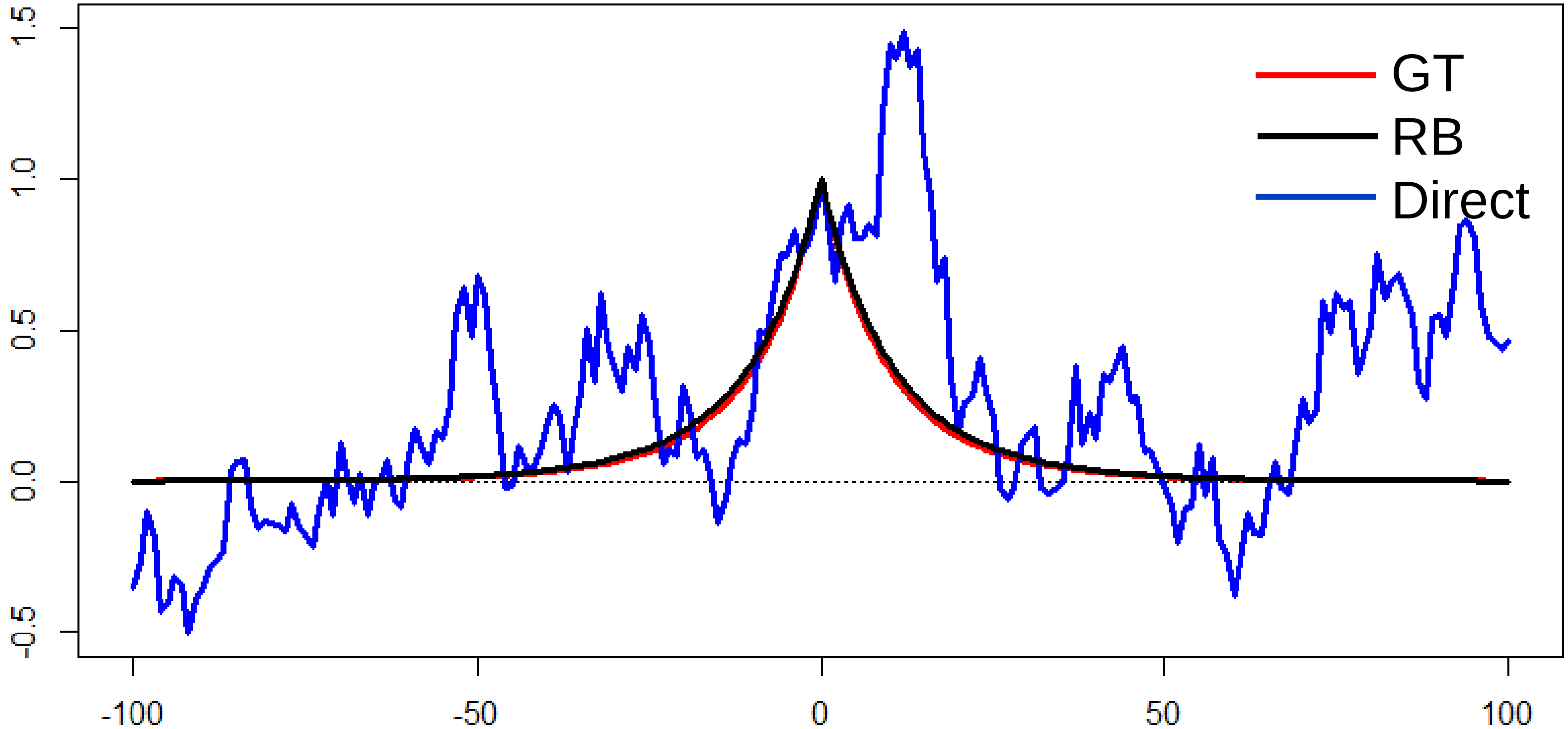
Direct vs Rao-Blackwellization N=1000



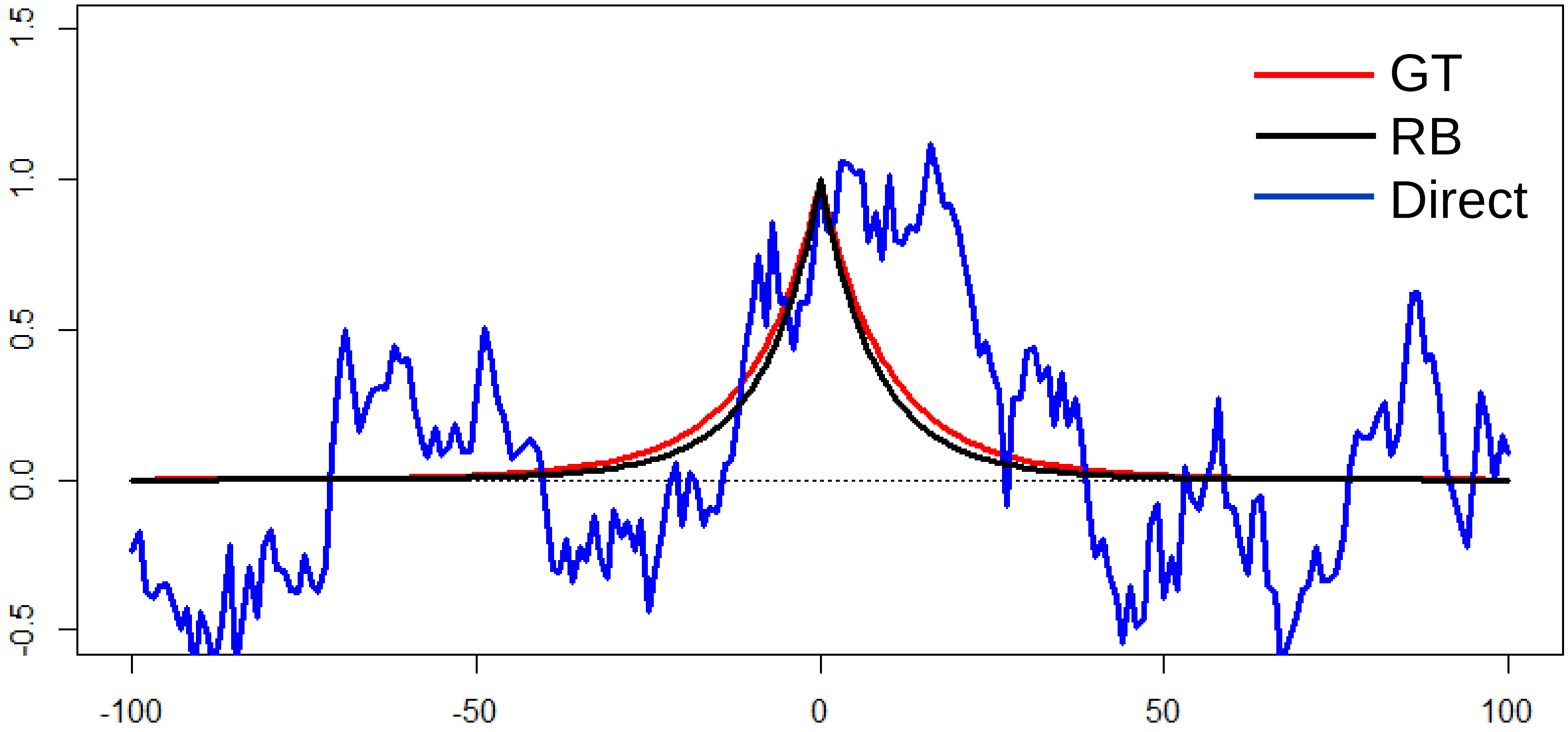
Direct vs Rao-Blackwellization N=100



Direct vs Rao-Blackwellization N=10



Direct vs Rao-Blackwellization N=10



Questions

Antithetic sampling

Things that are **antithetic** to one another contradict or oppose each other.

- Assume available $\hat{\mu}_1$ and $\hat{\mu}_2$, identically distributed with $\text{var}[\hat{\mu}_j] = \sigma^2/n$
- Assume $\text{cov}[\hat{\mu}_1, \hat{\mu}_2] < 0$.
- Define $\hat{\mu}_{AS} = \frac{1}{2}(\hat{\mu}_1 + \hat{\mu}_2)$

$$\begin{aligned}\text{var}[\hat{\mu}_{AS}] =& \tfrac{1}{4}(\text{var}[\hat{\mu}_1] + \text{var}[\hat{\mu}_2]) + \tfrac{1}{2}\text{cov}[\hat{\mu}_1, \hat{\mu}_2]\\ =& \tfrac{(1+\rho)\sigma^2}{2n}\end{aligned}$$

  where $\rho = \text{cor}[\hat{\mu}_1, \hat{\mu}_2]$.
- Gain by including $\hat{\mu}_2$ a factor of $\frac{1+\rho}{2}$!

Possible to construct such $\hat{\mu}_1, \hat{\mu}_2$?

Antithetic sampling

- Main idea: Most simulation procedures for generating $\mathbf{X} \sim f(\mathbf{x})$ is based on some transformation $X = h(\mathbf{U})$ where $\mathbf{U} = (U_1, ..., U_m)$ are iid uniform variables
- If U_j is uniform[0,1], then also $1 - U_j$ is uniform[0,1]
- $h(\mathbf{U})$ and $h(\mathbf{1} - \mathbf{U})$ will typically have negative correlation.
- Choose $\mathbf{X}_i = h(\mathbf{U}_i)$, $\mathbf{Y}_i = h(\mathbf{1} - \mathbf{U}_i)$

$$\hat{\mu}_1 = n^{-1} \sum_{i=1}^{n} h(\mathbf{U}_i)$$

$$\hat{\mu}_2 = n^{-1} \sum_{i=1}^{n} h(\mathbf{1} - \mathbf{U}_i)$$

- Can be generalized to other settings as well.
- The following slides:
  - Proof of $\mathrm{cor}[h(\mathbf{U}_i), h(\mathbf{1} - \mathbf{U}_i)] \leq 0$ for h monotone function in each U_j.

Antithetic sampling-theoretical derivations

- Assume $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$ iid sample
- Assume $\hat{\mu}_j = n^{-1} \sum_{i=1}^{n} h_j(\mathbf{x}_i)$ with $E[h_j(\mathbf{x}_i)] = \mu$.
- Assume $h_j(\mathbf{X}_i)$ is increasing in each argument
- Result: $\text{cor}[h_1(\mathbf{X}_i), h_2(\mathbf{X}_i)] \geq 0$.

If $h_1(\boldsymbol{x})$, $h_2(\boldsymbol{x})$ is non decreasing in each argument $\boldsymbol{x} = (x_1, \ldots, x_m)$
$h_j(\boldsymbol{x}) > h_j(\boldsymbol{x} - \boldsymbol{h})$, for all $\boldsymbol{h}$ such that $h_i > 0, i = 1, \ldots, m$
then $\text{cor}(h_1(\boldsymbol{X}_i), h_2(\boldsymbol{X}_i)) \geq 0$

Proof by induction on dimension:

1) Prove that it is true in dimension 1
2) Prove that if it is true for dimension $m-1$ then it is true for dimension m

Note slightly confusing the way we use the index on $\boldsymbol{X}_i$

Could have had $\mathrm{E}\left(h_j(\boldsymbol{X})\right) = \mu_j$ but this is not the case in question In antithetic sampling

Antithetic sampling-theoretical derivations

First: dimension 1

$$[h_1(X) - h_1(Y)][h_2(X) - h_2(Y)] \geq 0$$ Same sign

$$\Downarrow$$

$$E[[h_1(X) - h_1(Y)][h_2(X) - h_2(Y)]] \geq 0$$ For any X and Y

$$\Downarrow$$

$$E[h_1(X) - \mu - (h_1(Y) - \mu)][h_2(X) - \mu - (h_2(Y) - \mu)] \geq 0$$

$$\Downarrow$$ Assuming X, Y ind

$$\mathrm{cov}[h_1(X), h_2(X)] + \mathrm{cov}[h_1(Y), h_2(Y)] \geq 0$$

$$\Downarrow$$ Assuming X, Y iid

$$\mathrm{cov}[h_1(X), h_2(X)] \geq 0$$

Select joint distribution of X and Y to suit us

X and Y is selected to have the same distribution as X_i and X and Y are selected to be independent

Antithetic sampling-theoretical derivations

- Practical application in dimension 1

  - If h_1 increasing, h_2 decreasing:

$$\text{cor}[h_1(X), h_2(X)] = -\text{cor}[h_1(X), -h_2(X)] \leq 0$$

  - If X uniform: Then choose $h_1(X) = h(X)$, $h_2(X) = h(1-X)$

$$\begin{aligned}\text{Var}[\tfrac{1}{2}(h_1(X) + h_2(X))] =& \tfrac{1}{4}\text{Var}[h_1(X)] + \tfrac{1}{4}\text{Var}[h_2(X)] + \tfrac{1}{2}\text{Cov}[h_1(X), h_2(X)]\\ \leq& \tfrac{1}{2}\text{Var}[h_1(X)]\end{aligned}$$

  - If X Gaussian: Then choose $h_1(X) = h(X)$, $h_2(X) = h(-X)$

Works the same way, since: $\Phi(\cdot)$ is monotone and

$$\Phi(x) = u \Leftrightarrow \Phi(-x) = 1 - u$$

Warm up computations

$h_j(\mathbf{X}_i)$ is increasing in each argument

$E[h_j(\mathbf{X})|X_m] = \tilde{h}_j(X_m)$ is an increasing function in X_m.

For any $(x_1, x_2, \dots, x_{m-1})$, we have that

$$h_j(x_1, \dots, x_{m-1}, x_m) \geq h_j\,(x_1, \dots, x_{m-1}, x_m - h), \text{ for } h > 0$$

$$E\{h_j(X_1, \dots, X_{m-1}, x_m)\} \geq E\{h_j\,(X_1, \dots, X_{m-1}, x_m - h)\} \text{ for } h > 0$$

The relation is valid for any distribution,
we select: $f(\boldsymbol{x}|x_m)$ which gives the result

- $(E[\tilde{h}_j(X_m)] = E[E[h_j(\mathbf{X})|X_m]] = E[h_j(\mathbf{X})] = \mu)$

law of:
Total Expectation

- Assume $\mathrm{cor}[h_1(\mathbf{X}), h_2(\mathbf{X})] \geq 0$ for $\mathbf{X}$ $(m-1)$ dimensional. Then

$$\mathrm{cov}[h_1(\mathbf{X}), h_2(\mathbf{X})|X_m] \geq 0$$

Use result for $m-1$ for $f(\boldsymbol{x}|x_m)$

Taking expectations gives

$$\begin{aligned}0 \leq & E[\mathrm{cov}[h_1(\mathbf{X}), h_2(\mathbf{X})|X_m] \\ = & E[E[h_1(\mathbf{X})h_2(\mathbf{X})|X_m]] - \underbrace{E[E[h_1(\mathbf{X})|X_m] \cdot E[h_2(\mathbf{X})|X_m]]}\end{aligned}$$

$$\begin{aligned}& E[E[h_1(\mathbf{X})|X_m] \cdot E[h_2(\mathbf{X})|X_m]] \\ & = E[\tilde{h}_1(X_m)\tilde{h}_2(X_m)] \\ & = \mathrm{cov}[\tilde{h}_1(X_m)\tilde{h}_2(X_m)] + E[\tilde{h}_1(X_m)]E[\tilde{h}_2(X_m)] \\ & \geq E[\tilde{h}_1(X_m)]E[\tilde{h}_2(X_m)] = \mu^2\end{aligned}$$

Use result for 1 dimension

which gives

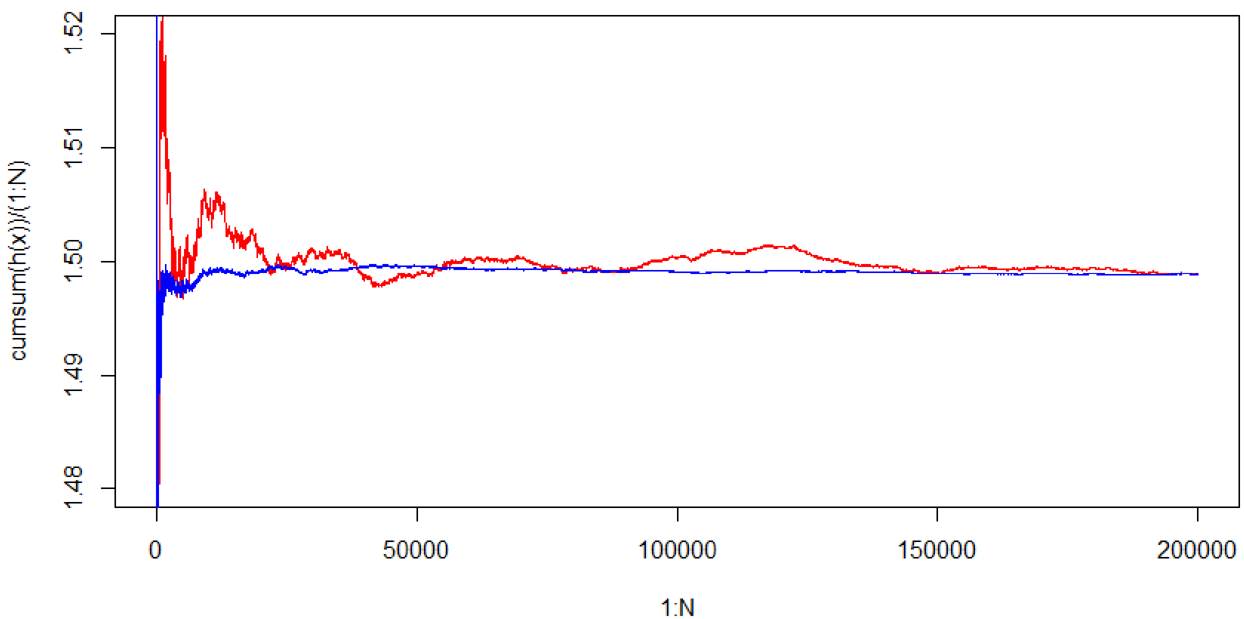
$$0 \leq E[h_1(\mathbf{X})h_2(\mathbf{X})] - \mu^2 = \mathrm{cov}[h_1(\mathbf{X}), h_2(\mathbf{X})]$$

Example

- $\mu = E[x/(2^x - 1)]$ for $x \sim N(0,1)$
- Note: $x \sim N(0,1)$ imply $-x \sim N(0,1)$
- `Example_6_10.R`

```
set.seed(231171)
N = 2e5
x = rnorm(N)
N2 = N/2
x1 = x[1:N2]
x12=matrix(t(cbind(x1,-x1)), 1,N)

plot(1:N, cumsum(h(x))/(1:N),type='l',col='red'
lines(1:N, cumsum(h(x12))/(1:N)  , col='blue')
```



Questions