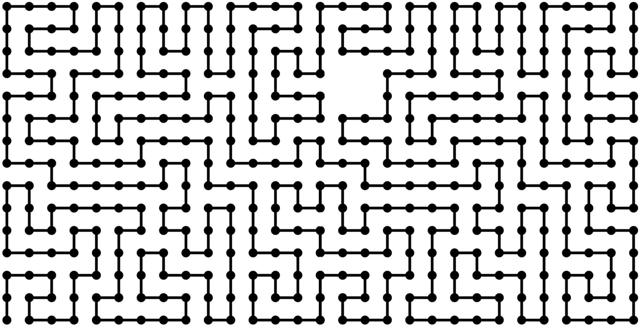
Winsorized Importance Sampling

Paulo Orenstein

February 8, 2019

Stanford University



Introduction

- Let $f(x)$ be an arbitrary function, $p(x)$, $q(x)$ probability densities. Suppose we are interested in

$$\theta = \mathbb{E}_p[f(X)] = \int_{\mathbb{R}} f(x)p(x)dx.$$

Introduction

- Let $f(x)$ be an arbitrary function, $p(x)$, $q(x)$ probability densities. Suppose we are interested in

$$\theta = \mathbb{E}_p[f(X)] = \int_{\mathbb{R}} f(x)p(x)dx.$$

- Assume we can only sample from q, which is called the *sampling distribution*; p is the *target distribution*.

Introduction

- Let $f(x)$ be an arbitrary function, $p(x)$, $q(x)$ probability densities. Suppose we are interested in

$$\theta = \mathbb{E}_p[f(X)] = \int_{\mathbb{R}} f(x)p(x)dx.$$

- Assume we can only sample from q, which is called the *sampling distribution*; p is the *target distribution*.

- The *importance sampling estimator* for θ is

$$\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n} f(X_i)\frac{p(X_i)}{q(X_i)}, \qquad X_i \sim q.$$

Introduction

- The importance sampling (IS) estimator is unbiased:

$$\hat{\theta}_n \xrightarrow{n\to\infty} \mathbb{E}\left[f(x)\frac{p(X)}{q(X)}\right] = \int f(x)\frac{p(x)}{q(x)}q(x)dx = \int f(x)p(x)dx = \theta,$$

as long as $q(x) > 0$ whenever $f(x)p(x) \neq 0$.

Introduction

- The importance sampling (IS) estimator is unbiased:

$$\hat{\theta}_n \xrightarrow{n\to\infty} \mathbb{E}\left[f(x)\frac{p(X)}{q(X)}\right] = \int f(x)\frac{p(x)}{q(x)}q(x)dx = \int f(x)p(x)dx = \theta,$$

as long as $q(x) > 0$ whenever $f(x)p(x) \neq 0$.

- But it can have huge or even infinite variance, leading to terrible estimates.

Introduction

- The importance sampling (IS) estimator is unbiased:

$$\hat{\theta}_n \xrightarrow{n\to\infty} \mathbb{E}\left[f(x)\frac{p(X)}{q(X)}\right] = \int f(x)\frac{p(x)}{q(x)}q(x)dx = \int f(x)p(x)dx = \theta,$$

  as long as $q(x) > 0$ whenever $f(x)p(x) \neq 0$.

- But it can have huge or even infinite variance, leading to terrible estimates.

- Can we control the variance of the terms

$$Y_i = f(X_i)\frac{p(X_i)}{q(X_i)}$$

  by sacrificing some small amount of bias?

Winsorizing

- Can we improve on the IS estimator by winsorizing, or capping, the weights?

Winsorizing

- Can we improve on the IS estimator by winsorizing, or capping, the weights?
- Denote the random variables winsorized at levels $-M$ and M by

$$Y_i^M = \max(-M, \min(Y_i, M)).$$

Winsorizing

- Can we improve on the IS estimator by winsorizing, or capping, the weights?
- Denote the random variables winsorized at levels $-M$ and M by

$$Y_i^M = \max(-M, \min(Y_i, M)).$$

- Define the *winsorized importance sampling estimator* at level M as

$$\hat{\theta}_n^M = \frac{1}{n}\sum_{i=1}^{n} Y_i^M.$$

Winsorizing

- Can we improve on the IS estimator by winsorizing, or capping, the weights?
- Denote the random variables winsorized at levels $-M$ and M by

$$Y_i^M = \max(-M, \min(Y_i, M)).$$

- Define the *winsorized importance sampling estimator* at level M as

$$\hat{\theta}_n^M = \frac{1}{n} \sum_{i=1}^{n} Y_i^M.$$

- Picking the right threshold level M is crucial.

Winsorizing

- Can we improve on the IS estimator by winsorizing, or capping, the weights?
- Denote the random variables winsorized at levels $-M$ and M by

$$Y_i^M = \max(-M, \min(Y_i, M)).$$

- Define the *winsorized importance sampling estimator* at level M as

$$\hat{\theta}_n^M = \frac{1}{n}\sum_{i=1}^{n} Y_i^M.$$

- Picking the right threshold level M is crucial.
- Bias-variance trade-off: smaller M implies less variance but more bias.

How to pick M?

- Let $\{Y_i\}_{i=1}^n$ be random variables distributed iid with mean θ.

How to pick M?

- Let $\{Y_i\}_{i=1}^n$ be random variables distributed iid with mean θ.
- Consider winsorizing Y_i at different threshold levels in a pre-chosen set $\Lambda = \{M_1, \ldots, M_k\}$ to obtain winsorized samples $\{Y_i^{M_j}\}_{i=1}^n$, $j = 1, \ldots, k$.

How to pick M?

- Let $\{Y_i\}_{i=1}^n$ be random variables distributed iid with mean θ.
- Consider winsorizing Y_i at different threshold levels in a pre-chosen set $\Lambda = \{M_1, \ldots, M_k\}$ to obtain winsorized samples $\{Y_i^{M_j}\}_{i=1}^n$, $j = 1, \ldots, k$.
- Pick the threshold level according to the rule

$$M_* = \mathbf{min}\left\{M \in \Lambda : \forall M', M'' \geq M, |\overline{Y^{M'}} - \overline{Y^{M''}}| \leq \alpha \cdot \left(\frac{\hat{\sigma}^{M'} + \hat{\sigma}^{M''}}{2}\right)\right\},$$

where:

How to pick M?

- Let $\{Y_i\}_{i=1}^n$ be random variables distributed iid with mean θ.
- Consider winsorizing Y_i at different threshold levels in a pre-chosen set $\Lambda = \{M_1, \ldots, M_k\}$ to obtain winsorized samples $\{Y_i^{M_j}\}_{i=1}^n$, $j = 1, \ldots, k$.
- Pick the threshold level according to the rule

$$M_* = \mathbf{min}\left\{M \in \Lambda : \forall M', M'' \geq M, |\overline{Y^{M'}} - \overline{Y^{M''}}| \leq \alpha \cdot \left(\frac{\hat{\sigma}^{M'} + \hat{\sigma}^{M''}}{2}\right)\right\},$$

where:

  - $\alpha = c \cdot \frac{t}{\sqrt{n-t}}$
  - c, t are chosen constants
  - $\overline{Y^M} = \frac{1}{n}\sum_{i=1}^n Y_i^M$
  - $\hat{\sigma}^M = \sqrt{\frac{1}{n}\sum_{i=1}^n (Y_i^M - \overline{Y^M})^2}$.

Why?

- Why is this rule sensible?

Why?

- Why is this rule sensible?

- Intuitively, if we have truncation levels $M' > M''$, we are willing to truncate further to M'' if the increase in bias $|\frac{1}{n}\sum_{i=1}^{n} Y_i^{M'} - \frac{1}{n}\sum_{i=1}^{n} Y_i^{M''}|$ is small relative to the standard deviation.

Why?

- Why is this rule sensible?

- Intuitively, if we have truncation levels $M' > M''$, we are willing to truncate further to M'' if the increase in bias $|\frac{1}{n}\sum_{i=1}^{n} Y_i^{M'} - \frac{1}{n}\sum_{i=1}^{n} Y_i^{M''}|$ is small relative to the standard deviation.

- The actual rule can be thought of as a concrete version of the Balancing Principle (or Lepski's Method), which is reminiscent of oracle inequalities.

Why?

- Why is this rule sensible?

- Intuitively, if we have truncation levels $M' > M''$, we are willing to truncate further to M'' if the increase in bias $|\frac{1}{n}\sum_{i=1}^{n} Y_i^{M'} - \frac{1}{n}\sum_{i=1}^{n} Y_i^{M''}|$ is small relative to the standard deviation.

- The actual rule can be thought of as a concrete version of the Balancing Principle (or Lepski's Method), which is reminiscent of oracle inequalities.

- With high probability, the mean-squared error using M_* is less than 5 times the error roughly incurred by choosing the best threshold level in the set.

Theorem

Let Y_i be iid with mean θ. Consider winsorizing Y_i at different levels in $\Lambda = \{M_1, \ldots, M_k\}$ to obtain samples $Y_i^{M_j}$. Pick the threshold level

$$M_* = \min\left\{ M \in \Lambda \ : \ \forall M', M'' \geq M, \quad |\overline{Y^{M'}} - \overline{Y^{M''}}| \leq \alpha \cdot \left(\frac{\hat{\sigma}^{M'} + \hat{\sigma}^{M''}}{2} \right) \right\},$$

where $\alpha = c \cdot \frac{t}{\sqrt{n} - t}$ with c, t chosen constants. Let $K > 0$ be such that $\mathbb{E}[|Y_i^{M_j} - \mathbb{E}[Y_i^{M_j}]|^3] \leq K(\mathbb{V}[Y_i^{M_j}])^{3/2}$ for all j. Then, with probability

$$1 - 2|\Lambda| \left(1 + \frac{50K}{\sqrt{n}} - \Phi\left(t \sqrt{\frac{n}{(\sqrt{n} - t)^2 + t^2}} \right) \right),$$

it holds

$$|\overline{Y^{M_*}} - \theta| \leq C \min_{M \in \Lambda} \left\{ |\mathbb{E}[Y_i^M] - \theta| + \frac{t\sqrt{n}}{\sqrt{n} - t} \frac{\hat{\sigma}^M}{\sqrt{n}} \right\},$$

where $C = C(c)$ can be made less than 4.25.

Theorem

Let Y_i be iid with mean θ. Consider winsorizing Y_i at different levels in $\Lambda = \{M_1, \ldots, M_k\}$ to obtain samples $Y_i^{M_j}$. Pick the threshold level

$$M_* = \min\left\{M \in \Lambda \;:\; \forall M', M'' \geq M, \quad |\overline{Y^{M'}} - \overline{Y^{M''}}| \leq \alpha \cdot \left(\frac{\hat{\sigma}^{M'} + \hat{\sigma}^{M''}}{2}\right)\right\},$$

where $\alpha = c \cdot \frac{t}{\sqrt{n}-t}$ with c, t chosen constants. Let $K > 0$ be such that $\mathbb{E}[|Y_i^{M_j} - \mathbb{E}[Y_i^{M_j}]|^3] \leq K(\mathbb{V}[Y_i^{M_j}])^{3/2}$ for all j. Then, with probability

$$1 - 2|\Lambda|\left(1 + \frac{50K}{\sqrt{n}} - \Phi\left(t\sqrt{\frac{n}{(\sqrt{n}-t)^2 + t^2}}\right)\right),$$

it holds

$$|\overline{Y^{M_*}} - \theta| \leq C \min_{M \in \Lambda}\left\{|\mathbb{E}[Y_i^M] - \theta| + \frac{t\sqrt{n}}{\sqrt{n}-t}\frac{\hat{\sigma}^M}{\sqrt{n}}\right\},$$

where $C = C(c)$ can be made less than 4.25.

Theorem

Let Y_i be iid with mean θ. Consider winsorizing Y_i at different levels in $\Lambda = \{M_1, \ldots, M_k\}$ to obtain samples $Y_i^{M_j}$. Pick the threshold level

$$M_* = \min \left\{ M \in \Lambda \ : \ \forall M', M'' \geq M, \quad |\overline{Y^{M'}} - \overline{Y^{M''}}| \leq \alpha \cdot \left(\frac{\hat{\sigma}^{M'} + \hat{\sigma}^{M''}}{2} \right) \right\},$$

where $\alpha = c \cdot \frac{t}{\sqrt{n} - t}$ with c, t chosen constants. Let $K > 0$ be such that $\mathbb{E}[|Y_i^{M_j} - \mathbb{E}[Y_i^{M_j}]|^3] \leq K(\mathbb{V}[Y_i^{M_j}])^{3/2}$ for all j. Then, with probability

$$1 - 2|\Lambda| \left(1 + \frac{50K}{\sqrt{n}} - \Phi \left(t \sqrt{\frac{n}{(\sqrt{n} - t)^2 + t^2}} \right) \right),$$

it holds

$$|\overline{Y^{M_*}} - \theta| \leq C \min_{M \in \Lambda} \left\{ |\mathbb{E}[Y_i^M] - \theta| + \frac{t\sqrt{n}}{\sqrt{n} - t} \frac{\hat{\sigma}^M}{\sqrt{n}} \right\},$$

where $C = C(c)$ can be made less than 4.25.

Theorem

Let Y_i be iid with mean θ. Consider winsorizing Y_i at different levels in $\Lambda = \{M_1, \ldots, M_k\}$ to obtain samples $Y_i^{M_j}$. Pick the threshold level

$$M_* = \min \left\{ M \in \Lambda \ : \ \forall M', M'' \geq M, \quad |\overline{Y^{M'}} - \overline{Y^{M''}}| \leq \alpha \cdot \left(\frac{\hat{\sigma}^{M'} + \hat{\sigma}^{M''}}{2} \right) \right\},$$

where $\alpha = c \cdot \frac{t}{\sqrt{n}-t}$ with c, t chosen constants. Let $K > 0$ be such that $\mathbb{E}[|Y_i^{M_j} - \mathbb{E}[Y_i^{M_j}]|^3] \leq K(\mathbb{V}[Y_i^{M_j}])^{3/2}$ for all j. Then, with probability

$$1 - 2|\Lambda| \left(1 + \frac{50K}{\sqrt{n}} - \Phi \left(t \sqrt{\frac{n}{(\sqrt{n}-t)^2 + t^2}} \right) \right),$$

it holds

$$|\overline{Y^{M_*}} - \theta| \leq C \min_{M \in \Lambda} \left\{ |\mathbb{E}[Y_i^M] - \theta| + \frac{t\sqrt{n}}{\sqrt{n}-t} \frac{\hat{\sigma}^M}{\sqrt{n}} \right\},$$

where $C = C(c)$ can be made less than 4.25.

Theorem

Let Y_i be iid with mean θ. Consider winsorizing Y_i at different levels in $\Lambda = \{M_1, \dots, M_k\}$ to obtain samples $Y_i^{M_j}$. Pick the threshold level

$$M_* = \min\left\{ M \in \Lambda \ : \ \forall M', M'' \geq M, \quad |\overline{Y^{M'}} - \overline{Y^{M''}}| \leq \alpha \cdot \left(\frac{\hat{\sigma}^{M'} + \hat{\sigma}^{M''}}{2} \right) \right\},$$

where $\alpha = c \cdot \frac{t}{\sqrt{n}-t}$ with c, t chosen constants. Let $K > 0$ be such that $\mathbb{E}[|Y_i^{M_j} - \mathbb{E}[Y_i^{M_j}]|^3] \leq K(\mathbb{V}[Y_i^{M_j}])^{3/2}$ for all j. Then, with probability

$$1 - 2|\Lambda| \left(1 + \frac{50K}{\sqrt{n}} - \Phi\left(t\sqrt{\frac{n}{(\sqrt{n}-t)^2 + t^2}} \right) \right),$$

it holds

$$|\overline{Y^{M_*}} - \theta| \leq C \min_{M \in \Lambda} \left\{ |\mathbb{E}[Y_i^M] - \theta| + \frac{t\sqrt{n}}{\sqrt{n}-t} \frac{\hat{\sigma}^M}{\sqrt{n}} \right\},$$

where $C = C(c)$ can be made less than 4.25.

Proof

Proof

- Apply the Balancing Theorem:

Balancing Theorem

Suppose $\theta \in \mathbb{R}$ is an unknown parameter, $\{\hat{E}^M\}_{M\in\Theta}$ is a sequence of estimators of θ indexed by $M \in \Theta \subset \mathbb{R}$, with Θ a finite set. Additionally, suppose that for each M we know $|\hat{E}^M - \theta| \leq \text{bias}(M) + \hat{s}(M)$, where we assume $\text{bias}(M)$ is unknown but non-increasing in M, and $\hat{s}(M) > 0$ is observed and non-decreasing in M. Fix $c > 2$, and take

$$M_* = \min\left\{M \in \Theta : \forall M', M'' \geq M,\ |\hat{E}^{M'} - \hat{E}^{M''}| \leq c\left(\frac{\hat{s}(M') + \hat{s}(M'')}{2}\right)\right\}.$$

Then we have that

$$|\hat{E}^{M_*} - \theta| \leq C \min_{M\in\Theta}\{\hat{s}(M) + \text{bias}(M)\},$$

where C is a constant depending on the chosen c, less than 4.25.

Proof

- Apply the Balancing Theorem:

Balancing Theorem

Suppose $\theta \in \mathbb{R}$ is an unknown parameter, $\{\hat{E}^M\}_{M \in \Theta}$ is a sequence of estimators of θ indexed by $M \in \Theta \subset \mathbb{R}$, with Θ a finite set. Additionally, suppose that for each M we know $|\hat{E}^M - \theta| \leq \text{bias}(M) + \hat{s}(M)$, where we assume bias$(M)$ is unknown but non-increasing in M, and $\hat{s}(M) > 0$ is observed and non-decreasing in M. Fix $c > 2$, and take

$$M_* = \min \left\{ M \in \Theta : \forall M', M'' \geq M,\ |\hat{E}^{M'} - \hat{E}^{M''}| \leq c \left(\frac{\hat{s}(M') + \hat{s}(M'')}{2} \right) \right\}.$$

Then we have that

$$|\hat{E}^{M_*} - \theta| \leq C \min_{M \in \Theta} \{ \hat{s}(M) + \text{bias}(M) \},$$

where C is a constant depending on the chosen c, less than 4.25.

- Then, use Berry-Esseen to get probabilistic bounds.

Proof (of Balancing Theorem)

- We must thus show that for all $M \in \Theta$, there exists $C \geq 0$ such that $|\hat{E}^{M_*} - \theta| \leq C(\hat{s}(M) + \textbf{bias}(M))$. For this we shall consider two cases.

Proof (of Balancing Theorem)

- We must thus show that for all $M \in \Theta$, there exists $C \geq 0$ such that $|\hat{E}^{M_*} - \theta| \leq C(\hat{s}(M) + \textbf{bias}(M))$. For this we shall consider two cases.

- (i) First, consider any fixed M such that $M > M_*$. Then, by our definition of M_*, and since $\hat{s}(M)$ is non-decreasing in M,

$$|\hat{E}^{M_*} - \hat{E}^{M}| \leq c \cdot \hat{s}(M).$$

Also, as $|\hat{E}^{M} - \theta| \leq \textbf{bias}(M) + \hat{s}(M)$, we get

$$\begin{aligned}|\hat{E}^{M_*} - \theta| \leq |\hat{E}^{M_*} - \hat{E}^{M}| + |\hat{E}^{M} - \theta| &\leq c\hat{s}(M) + \textbf{bias}(M) + \hat{s}(M) \\ &= \textbf{bias}(M) + (c+1)\hat{s}(M).\end{aligned}$$

This proves the case $M > M_*$.

Proof (of Balancing Theorem)

- We must thus show that for all $M \in \Theta$, there exists $C \geq 0$ such that $|\hat{E}^{M_*} - \theta| \leq C(\hat{s}(M) + \textbf{bias}(M))$. For this we shall consider two cases.

- (i) First, consider any fixed M such that $M > M_*$. Then, by our definition of M_*, and since $\hat{s}(M)$ is non-decreasing in M,

$$|\hat{E}^{M_*} - \hat{E}^{M}| \leq c \cdot \hat{s}(M).$$

  Also, as $|\hat{E}^{M} - \theta| \leq \textbf{bias}(M) + \hat{s}(M)$, we get

$$\begin{aligned} |\hat{E}^{M_*} - \theta| &\leq |\hat{E}^{M_*} - \hat{E}^{M}| + |\hat{E}^{M} - \theta| \leq c\hat{s}(M) + \textbf{bias}(M) + \hat{s}(M) \\ &= \textbf{bias}(M) + (c+1)\hat{s}(M). \end{aligned}$$

  This proves the case $M > M_*$.

- (ii) The other side is harder.

How well does this work in practice?

- We consider examples with real and synthetic data.

How well does this work in practice?

- We consider examples with real and synthetic data.
- Compare three estimators:
  - usual IS: no winsorization;
  - CV IS: winsorization with threshold chosen via CV;
  - Balanced IS: winsorization with threshold chosen via Balancing Theorem.

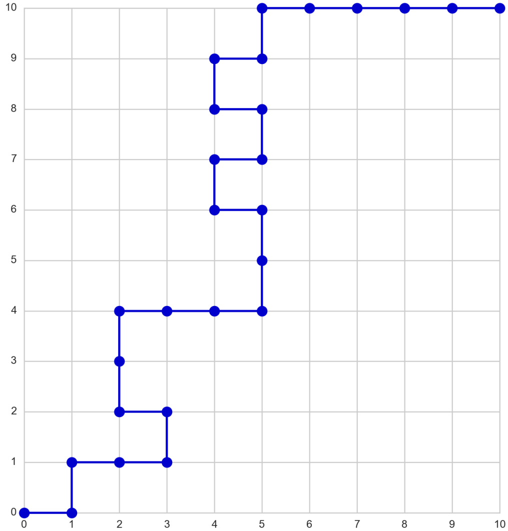
How well does this work in practice?

- We consider examples with real and synthetic data.
- Compare three estimators:
  - usual IS: no winsorization;
  - CV IS: winsorization with threshold chosen via CV;
  - Balanced IS: winsorization with threshold chosen via Balancing Theorem.
- CV IS takes 10-20$\times$ longer than Balanced IS and is usually worse.

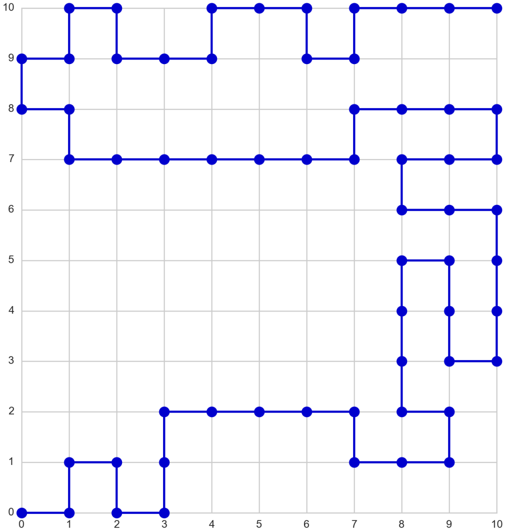
How well does this work in practice?

- We consider examples with real and synthetic data.
- Compare three estimators:
  - usual IS: no winsorization;
  - CV IS: winsorization with threshold chosen via CV;
  - Balanced IS: winsorization with threshold chosen via Balancing Theorem.
- CV IS takes 10-20$\times$ longer than Balanced IS and is usually worse.
- For small variances Balanced IS matches usual IS; as the proposal distribution gets worse, Balanced IS performs much better.

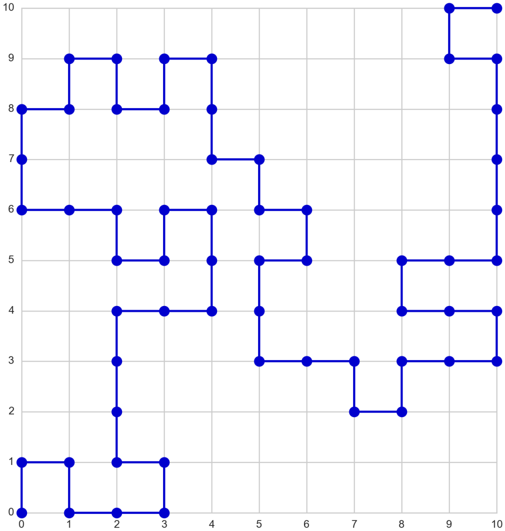
Example: self-avoiding walk [Knuth, 1976]



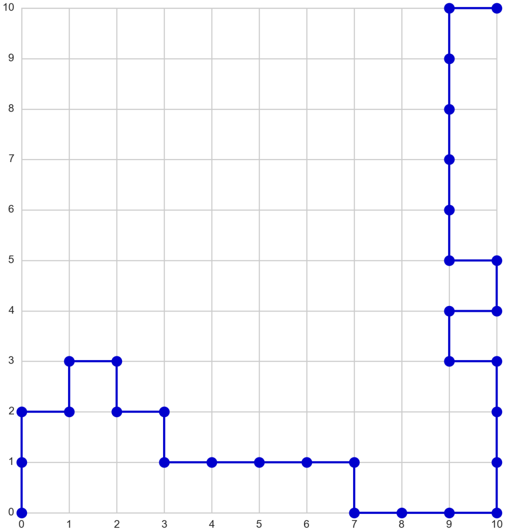
Example: self-avoiding walk [Knuth, 1976]



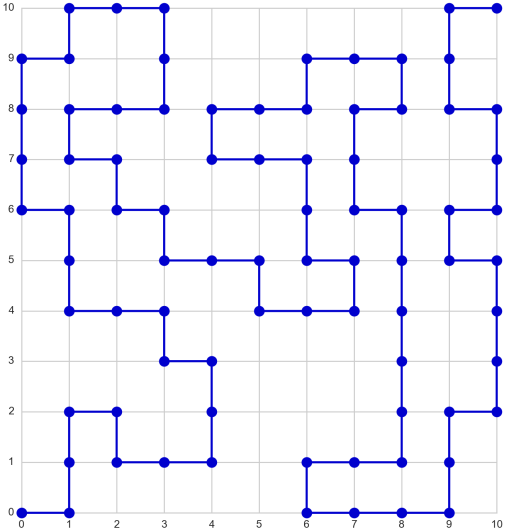
Example: self-avoiding walk [Knuth, 1976]



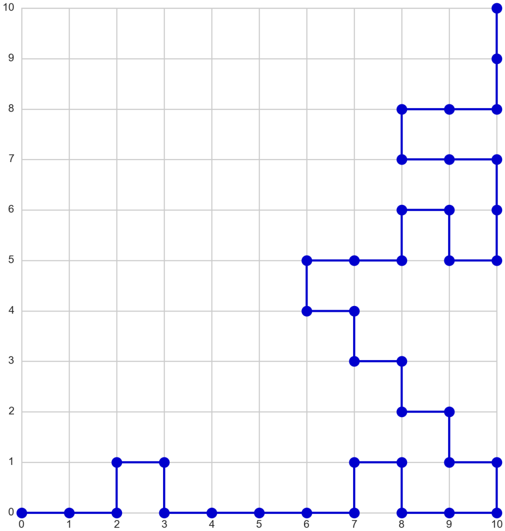
Example: self-avoiding walk [Knuth, 1976]



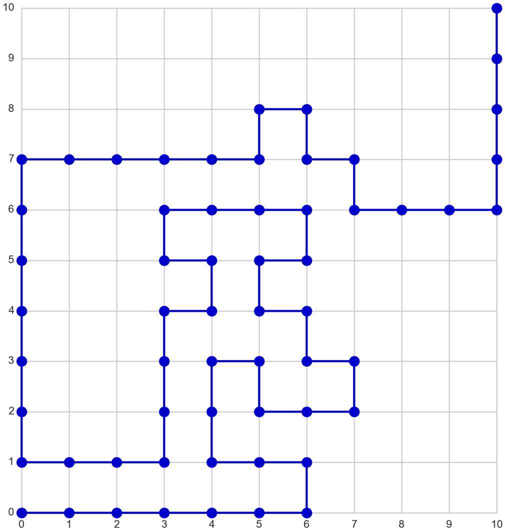
Example: self-avoiding walk [Knuth, 1976]



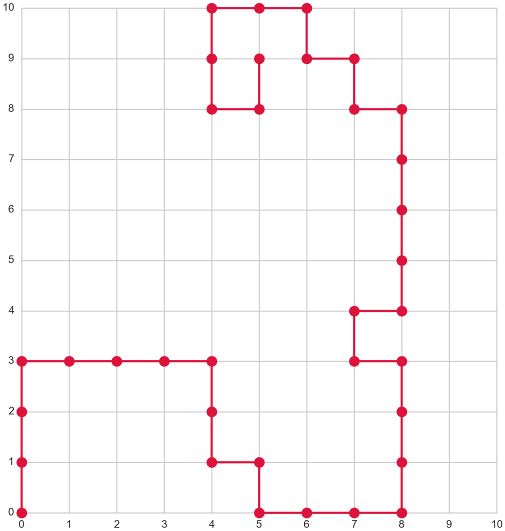
Example: self-avoiding walk [Knuth, 1976]



Example: self-avoiding walk [Knuth, 1976]



Example: self-avoiding walk [Knuth, 1976]



Example: self-avoiding walk [Knuth, 1976]

- Knuth suggested estimating the number of self-avoiding walks using importance sampling.

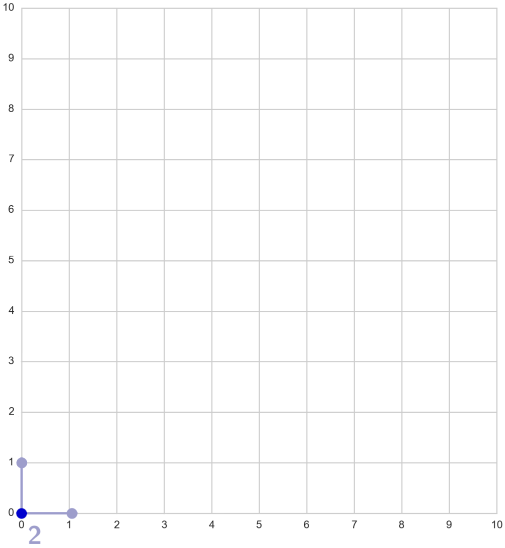
Example: self-avoiding walk [Knuth, 1976]

- Knuth suggested estimating the number of self-avoiding walks using importance sampling.
- For this, we need to choose a sampling distribution, $q(x)$, over the self-avoiding walks.

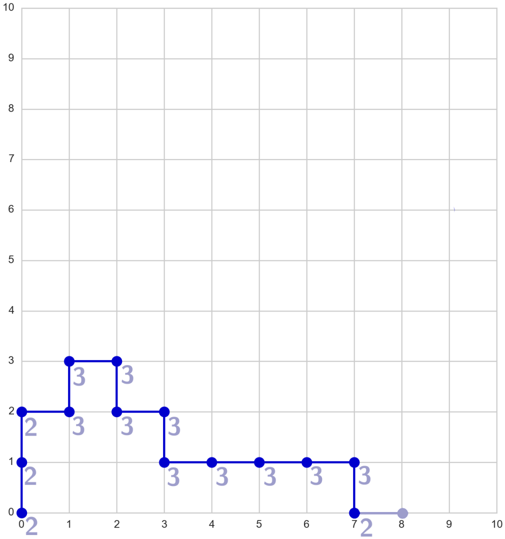
Example: self-avoiding walk [Knuth, 1976]

- Knuth suggested estimating the number of self-avoiding walks using importance sampling.
- For this, we need to choose a sampling distribution, $q(x)$, over the self-avoiding walks.
- Consider building one sequentially.

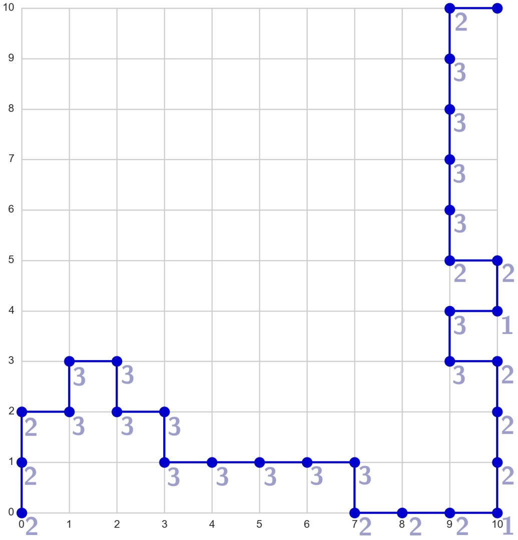
Example: self-avoiding walk [Knuth, 1976]



Example: self-avoiding walk [Knuth, 1976]

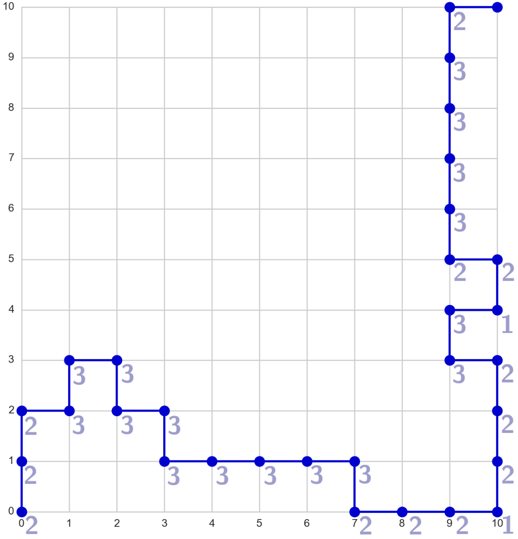


Example: self-avoiding walk [Knuth, 1976]



Example: self-avoiding walk [Knuth, 1976]

$$q(x) = 1^{-2} \times 2^{-12} \times 3^{-16}$$



Example: self-avoiding walk [Knuth, 1976]

- Define:
  - $p(x) = \frac{1}{Z_n}\mathbb{I}_{[SAW]}(x)$; note Z_n is the number of self-avoiding random walks;
  - $q(x) = \frac{1}{d_1 \cdot d_2 \cdots d_{m_x}}$; d_i is the number of available neighbors to i (could be 0);
  - $f(x) = Z_n$.

Example: self-avoiding walk [Knuth, 1976]

- Define:
  - $p(x) = \frac{1}{Z_n}\mathbb{I}_{[SAW]}(x)$; note Z_n is the number of self-avoiding random walks;
  - $q(x) = \frac{1}{d_1 \cdot d_2 \cdots d_{m_X}}$; d_i is the number of available neighbors to i (could be 0);
  - $f(x) = Z_n$.
- We would like to estimate

$$Z_n = \mathbb{E}_p[Z_n] = \mathbb{E}_p[f(X)] = \mathbb{E}_q\left[\frac{f(X)p(X)}{q(X)}\right] = \mathbb{E}_q\left[\frac{\mathbb{I}_{[SAW]}(X)}{q(X)}\right]$$

$$\approx \frac{1}{n}\sum_{i=1}^{n} d_1(X_i)d_2(X_i)\cdots d_{m_{X_i}}(X_i)\cdot \mathbb{I}_{[SAW]}(X).$$

Example: self-avoiding walk [Knuth, 1976]

- How does winsorization perform?

Example: self-avoiding walk [Knuth, 1976]

- How does winsorization perform?
- 1000 simulations of 1000 SAWs.
- $\theta = 1.56 \cdot 10^{24}$; $c = 1 + \sqrt{3}$, $t = 2$.
- $M \in \{10^{21}, 5 \cdot 10^{23}, 10^{25}, 5 \cdot 10^{26}, 10^{28}\}$.

Example: self-avoiding walk [Knuth, 1976]

- How does winsorization perform?
- 1000 simulations of 1000 SAWs.
- $\theta = 1.56 \cdot 10^{24}$; $c = 1 + \sqrt{3}$, $t = 2$.
- $M \in \{10^{21}, 5 \cdot 10^{23}, 10^{25}, 5 \cdot 10^{26}, 10^{28}\}$.

	IS	CV IS	Balanced IS
MSE	$2.075 \cdot 10^{49}$	$2.457 \cdot 10^{48}$	$2.437 \cdot 10^{48}$
MAD	$1.817 \cdot 10^{24}$	$1.567 \cdot 10^{24}$	$1.561 \cdot 10^{24}$

Procedure

- Procedure is run as follows:
  - Let $M_1 = 10^{28}$;
    - set $M_* = M_1$

Procedure

- Procedure is run as follows:
  - Let $M_1 = 10^{28}$;
    - set $M_* = M_1$
  - Let $M_2 = 5 \cdot 10^{26}$;

Procedure

- Procedure is run as follows:
  - Let $M_1 = 10^{28}$;
    - set $M_* = M_1$
  - Let $M_2 = 5 \cdot 10^{26}$;
    - if $|\overline{Y}^{M_1} - \overline{Y}^{M_2}| \leq \alpha \left(\frac{\hat{\sigma}^{M_1} + \hat{\sigma}^{M_2}}{2} \right)$, set $M_* = M_2$, and consider further truncation;

Procedure

- Procedure is run as follows:
  - Let $M_1 = 10^{28}$;
    - set $M_* = M_1$
  - Let $M_2 = 5 \cdot 10^{26}$;
    - if $|\overline{Y}^{M_1} - \overline{Y}^{M_2}| \leq \alpha \left(\frac{\hat{\sigma}^{M_1} + \hat{\sigma}^{M_2}}{2} \right)$, set $M_* = M_2$, and consider further truncation;
    - else, stop

Procedure

- Procedure is run as follows:
  - Let $M_1 = 10^{28}$;
    - set $M_* = M_1$
  - Let $M_2 = 5 \cdot 10^{26}$;
    - if $|\overline{Y}^{M_1} - \overline{Y}^{M_2}| \leq \alpha \left(\frac{\hat{\sigma}^{M_1} + \hat{\sigma}^{M_2}}{2} \right)$, set $M_* = M_2$, and consider further truncation;
    - else, stop
  - Let $M_3 = 10^{25}$

Procedure

- Procedure is run as follows:
  - Let $M_1 = 10^{28}$;
    - set $M_* = M_1$
  - Let $M_2 = 5 \cdot 10^{26}$;
    - if $|\overline{Y}^{M_1} - \overline{Y}^{M_2}| \leq \alpha \left(\frac{\hat{\sigma}^{M_1} + \hat{\sigma}^{M_2}}{2} \right)$, set $M_* = M_2$, and consider further truncation;
    - else, stop
  - Let $M_3 = 10^{25}$
    - if $|\overline{Y}^{M_1} - \overline{Y}^{M_3}| \leq \alpha \left(\frac{\hat{\sigma}^{M_1} + \hat{\sigma}^{M_3}}{2} \right)$ and $|\overline{Y}^{M_2} - \overline{Y}^{M_3}| \leq \alpha \left(\frac{\hat{\sigma}^{M_2} + \hat{\sigma}^{M_3}}{2} \right)$, set $M_* = M_3$, and consider further truncation;

Procedure

- Procedure is run as follows:
  - Let $M_1 = 10^{28}$;
    - set $M_* = M_1$
  - Let $M_2 = 5 \cdot 10^{26}$;
    - if $|\overline{Y}^{M_1} - \overline{Y}^{M_2}| \leq \alpha \left(\frac{\hat{\sigma}^{M_1} + \hat{\sigma}^{M_2}}{2} \right)$, set $M_* = M_2$, and consider further truncation;
    - else, stop
  - Let $M_3 = 10^{25}$
    - if $|\overline{Y}^{M_1} - \overline{Y}^{M_3}| \leq \alpha \left(\frac{\hat{\sigma}^{M_1} + \hat{\sigma}^{M_3}}{2} \right)$ and $|\overline{Y}^{M_2} - \overline{Y}^{M_3}| \leq \alpha \left(\frac{\hat{\sigma}^{M_2} + \hat{\sigma}^{M_3}}{2} \right)$, set $M_* = M_3$, and consider further truncation;
    - else, stop

Procedure

- Procedure is run as follows:
  - Let $M_1 = 10^{28}$;
    - set $M_* = M_1$
  - Let $M_2 = 5 \cdot 10^{26}$;
    - if $|\overline{Y}^{M_1} - \overline{Y}^{M_2}| \leq \alpha \left(\frac{\hat{\sigma}^{M_1} + \hat{\sigma}^{M_2}}{2} \right)$, set $M_* = M_2$, and consider further truncation;
    - else, stop
  - Let $M_3 = 10^{25}$
    - if $|\overline{Y}^{M_1} - \overline{Y}^{M_3}| \leq \alpha \left(\frac{\hat{\sigma}^{M_1} + \hat{\sigma}^{M_3}}{2} \right)$ and $|\overline{Y}^{M_2} - \overline{Y}^{M_3}| \leq \alpha \left(\frac{\hat{\sigma}^{M_2} + \hat{\sigma}^{M_3}}{2} \right)$, set $M_* = M_3$, and consider further truncation;
    - else, stop
  - ...

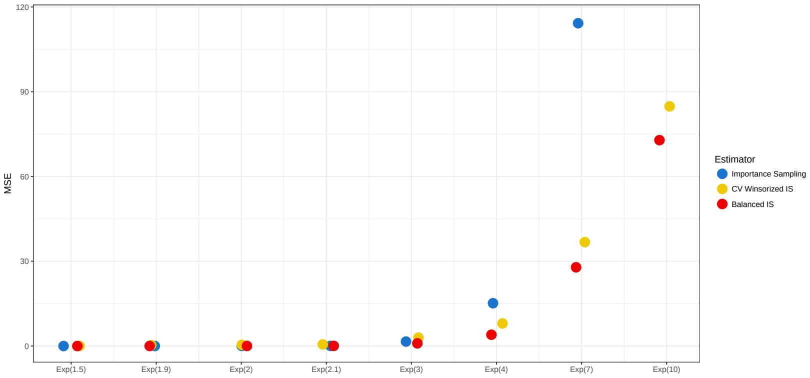
Procedure

- Procedure is run as follows:
  - Let $M_1 = 10^{28}$;
    - set $M_* = M_1$
  - Let $M_2 = 5 \cdot 10^{26}$;
    - if $|\overline{Y}^{M_1} - \overline{Y}^{M_2}| \leq \alpha \left(\frac{\hat{\sigma}^{M_1} + \hat{\sigma}^{M_2}}{2} \right)$, set $M_* = M_2$, and consider further truncation;
    - else, stop
  - Let $M_3 = 10^{25}$
    - if $|\overline{Y}^{M_1} - \overline{Y}^{M_3}| \leq \alpha \left(\frac{\hat{\sigma}^{M_1} + \hat{\sigma}^{M_3}}{2} \right)$ and $|\overline{Y}^{M_2} - \overline{Y}^{M_3}| \leq \alpha \left(\frac{\hat{\sigma}^{M_2} + \hat{\sigma}^{M_3}}{2} \right)$, set $M_* = M_3$, and consider further truncation;
    - else, stop
  - ...
- Computational complexity: $O(|\Lambda| \cdot (|\Lambda| + n))$

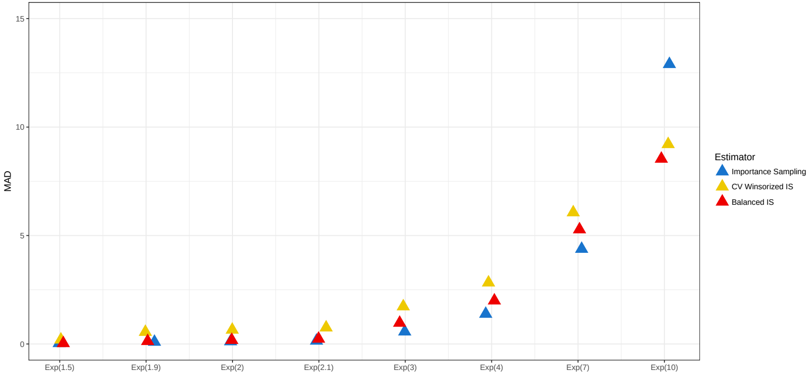
Simulation 1: Exponential

- $p = \frac{1}{\theta}\text{Expo}$,
- $q = \text{Expo}$,
- $f(x) = x$,
- $\theta \in \{1.3, 1.5, 1.9, 2, 2.1, 3, 4, 10\}$
- $M \in \{550, 500, 400, 200, 100, 10\}$

Simulation 1: Exponential



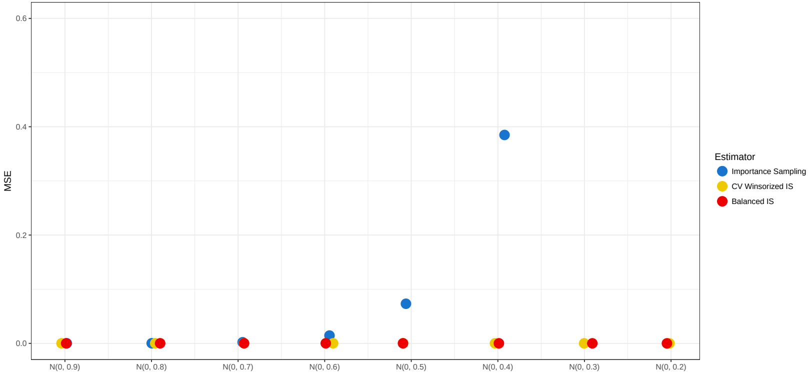
Simulation 1: Exponential



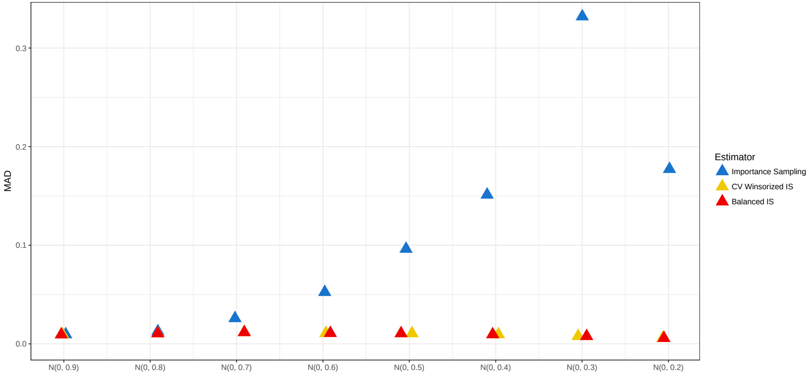
Simulation 2: Normal

- $p = N(0, 1)$,
- $q = N(0, \theta)$,
- $f(x) = x$,
- $\theta = \{0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.9\}$
- $M \in \{550, 500, 400, 200, 100, 10\}$

Simulation 2: Normal



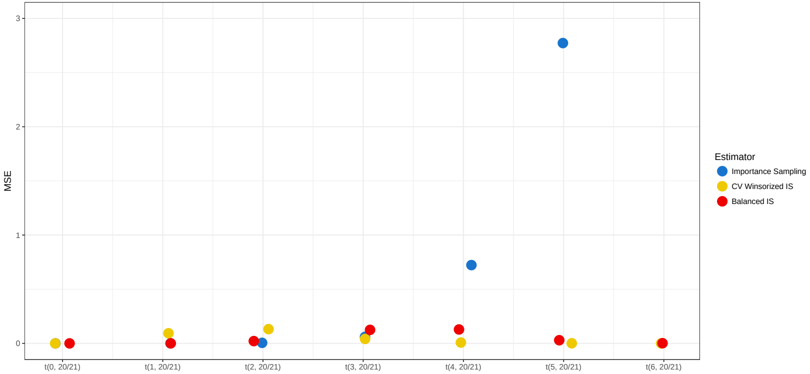
Simulation 2: Normal



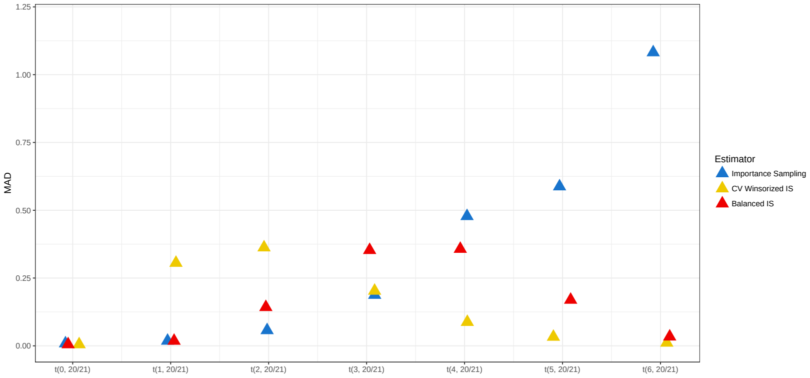
Simulation 3: t

- $p = t_{21}(0, 1)$,
- $q = t_{21}(\theta, 1 - 1/21)$,
- $f(x) = x$,
- $\theta = \{0, 0.5, 1, 1.5, 2, 2.5, 3\}$
- $M \in \{550, 500, 400, 200, 100, 50, 5, 1\}$

Simulation 3: t



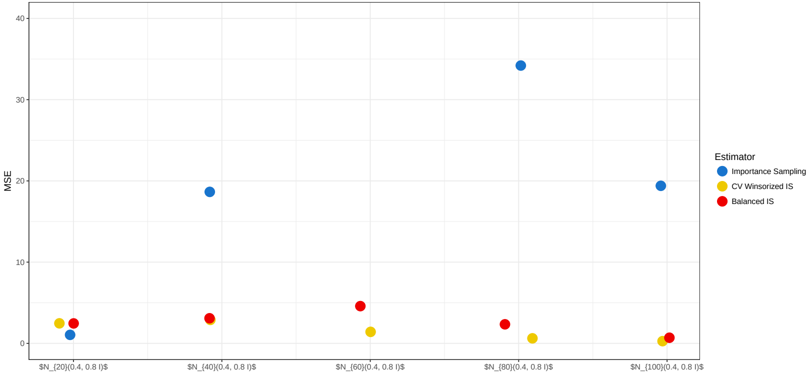
Simulation 3: t



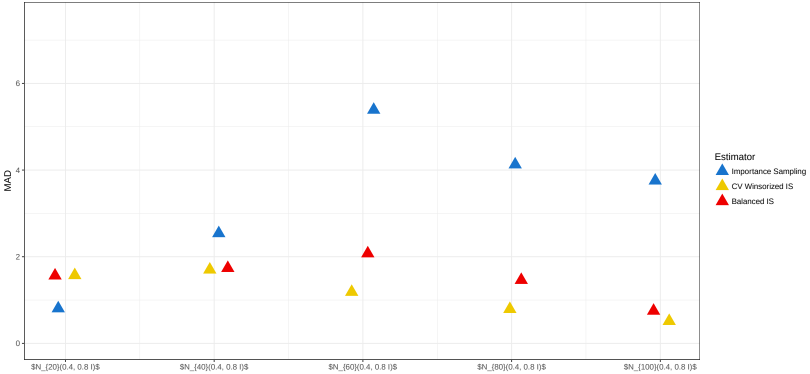
Simulation 4: Multivariate Normal

- $p = N_\theta(0, 1)$,
- $q = t_{21,\theta}(0.4 \cdot \mathbb{1}, 0.8 \cdot I)$,
- $f(\mathbf{x}) = \sum_{i=1}^{\theta} x_i$,
- $\theta = \{20, 40, 60, 80, 100\}$
- $M \in \{550, 500, 400, 200, 100, 50, 10\}$

Simulation 4: Multivariate Normal



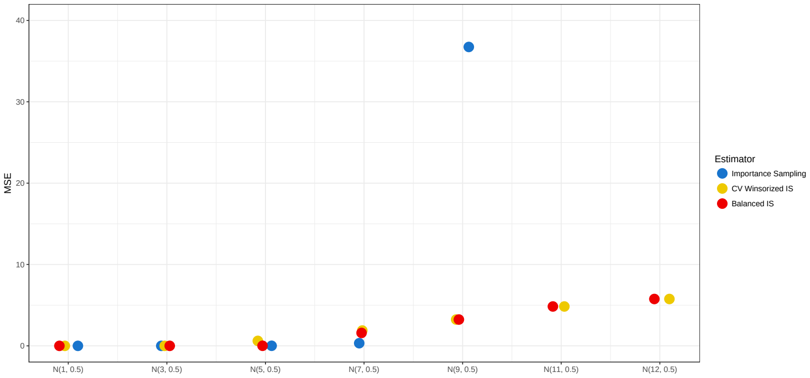
Simulation 4: Multivariate Normal



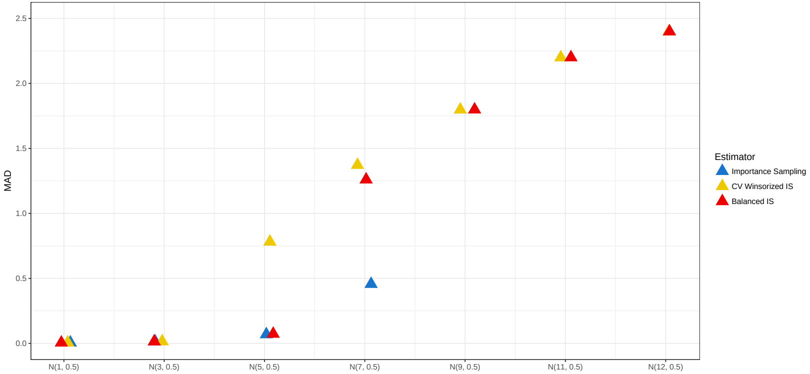
Simulation 5: Normal Mixture

- $p = 0.8 \cdot N(0, 0.5) + 0.2 \cdot N(\theta, 0.5)$,
- $q = N(0, 4)$,
- $f(x) = x$,
- $\theta = \{1, 3, 5, 7, 9, 11, 12\}$
- $M \in \{550, 500, 400, 200, 100, 10\}$

Simulation 5: Normal Mixture



Simulation 5: Normal Mixture



Is it worth winsorizing?

- Negative aspects:
  - theory requires high n, at least 10^8 (but can be improved);
  - must be provided truncation values;
  - why winsorize symmetrically around 0?

Is it worth winsorizing?

- Negative aspects:
  - theory requires high n, at least 10^8 (but can be improved);
  - must be provided truncation values;
  - why winsorize symmetrically around 0?
- Positive aspects:
  - works well in practice;
  - adaptive to the sample;
  - comes with finite-sample optimality properties.

Conclusion

- Importance sampling should not rely only on sample mean.

Conclusion

- Importance sampling should not rely only on sample mean.
- We need robust, adaptive alternatives.

Conclusion

- Importance sampling should not rely only on sample mean.
- We need robust, adaptive alternatives.
- Balanced IS has theoretical guarantees and performs well in practice:
  - in high-variance settings, it outperforms usual IS
  - in low-variance settings, it matches it.

Conclusion

- Importance sampling should not rely only on sample mean.
- We need robust, adaptive alternatives.
- Balanced IS has theoretical guarantees and performs well in practice:
  - in high-variance settings, it outperforms usual IS
  - in low-variance settings, it matches it.
- Many future extensions.

References

- Ionides, E. L. (2008). Truncated importance sampling. *Journal of Computational and Graphical Statistics*, 17(2).
- Mathé, P. (2006). The Lepskii principle revisited. *Inverse problems*, 22(3).
- Orenstein, P. (2018). Finite-sample Guarantees for Winsorized Importance Sampling. *arXiv preprint arXiv:1810.11130*.
- Shao, Q.-M. (2005). An explicit berry–esseen bound for student's t-statistic via Stein's Method. *Stein's Method and Applications*, 5:143.
- Vehtari, A., Gelman, A., and Gabry, J. (2015). Pareto smoothed importance sampling. *arXiv preprint arXiv:1507.02646*