Scenario Generation and Sampling Methods

Tito Homem-de-Mello
Güzin Bayraksan

SVAN 2016 – IMPA
9–13 May 2106

Let's Start with a Recap

Sample Average Approximation

We want to solve the "true" problem

$$\min_{x \in X} \{g(x) := \mathbb{E}[G(x,\xi)]\}, \tag{SP}$$

but it is typically very difficult or impossible to solve.

Sample Average Approximation

We want to solve the "true" problem

$$\min_{x \in X} \{ g(x) := \mathbb{E}[G(x,\xi)] \}, \qquad \text{(SP)}$$

but it is typically very difficult or impossible to solve.

Generate a sample $\{\xi^1, \xi^2, \ldots \xi^N\}$ and solve the Sample Average Approximation (SAA):

$$\min_{x \in X} \left\{ \hat{g}_N(x) \;:=\; \frac{1}{N} \sum_{j=1}^{N} G(x, \xi^j) \right\}. \qquad (\mathrm{SP}_N)$$

Sample Average Approximation

Recall

$$\begin{aligned} x^* &:= \text{an optimal solution of (SP)} \\ S^* &:= \text{the set of optimal solutions of (SP)} \\ \nu^* &:= \text{the optimal value of (SP)} \end{aligned}$$

and

$$\begin{aligned} \hat{x}_N &:= \text{an optimal solution of } (\mathrm{SP}_N) \\ S_N &:= \text{the set of optimal solutions of } (\mathrm{SP}_N) \\ \nu_N &:= \text{the optimal value of } (\mathrm{SP}_N) \end{aligned}$$

View (x^*, S^*, ν^*) as "statistical estimators" of $(\hat{x}_N, S_N, \nu_N)$

Sample Average Approximation

So far... **Standard Monte Carlo**: $\{\xi^1, \xi^2, \ldots \xi^N\}$ generated independent and identically distributed (iid) as ξ

And... Looked at the properties of statistical estimators

- Negative Bias: $\mathbb{E}[\nu_N] \leq \nu^*$
- Strong Consistency: e.g., $\nu_N \to \nu^*$, w.p.1.
- Rates of Convergence: e.g., $\sqrt{N}(\nu_N - \nu^*) \ \xrightarrow{d}\ \text{Normal}(0, \sigma(x^*))$
- Large Deviations; Exponential Rates of Convergence; ...

Why Variance Reduction?

Why Variance Reduction?

- Consider $g(x) = \mathbb{E}[G(x, \xi)]$ for a fixed $x \in X$.
- Estimate $g(x)$ by $\hat{g}_N(x) = \frac{1}{N} \sum_{j=1}^{N} G(x, \xi^j)$

Why Variance Reduction?

- Consider $g(x) = \mathbb{E}[G(x,\xi)]$ for a fixed $x \in X$.
- Estimate $g(x)$ by $\hat{g}_N(x) = \frac{1}{N}\sum_{j=1}^{N} G(x,\xi^j)$
- For illustration purposes, suppose we provide a Confidence Interval (CI) on the value of $g(x)$ as

$$\left[\hat{g}_N(x) - 2\sqrt{\mathrm{Var}\left[\hat{g}_N(x)\right]},\ \hat{g}_N(x) + 2\sqrt{\mathrm{Var}\left[\hat{g}_N(x)\right]}\right]$$

Why Variance Reduction?

- Consider $g(x) = \mathbb{E}[G(x,\xi)]$ for a fixed $x \in X$.
- Estimate $g(x)$ by $\hat{g}_N(x) = \frac{1}{N}\sum_{j=1}^{N} G(x,\xi^j)$
- For illustration purposes, suppose we provide a Confidence Interval (CI) on the value of $g(x)$ as

$$\left[\hat{g}_N(x) - 2\sqrt{\text{Var}\left[\hat{g}_N(x)\right]},\ \hat{g}_N(x) + 2\sqrt{\text{Var}\left[\hat{g}_N(x)\right]}\right]$$

Suppose we have 2 estimators:

Estimator	$\hat{\mathbf{g}}_{\mathbf{N}}(\mathbf{x})$	Var$[\hat{\mathbf{g}}_{\mathbf{N}}(\mathbf{x})]$	**CI**
1	10	25	[0, 20]
2	10	0.25	[9, 11]

Which one is preferable? Clearly. . .

Why Variance Reduction?

Suppose we use standard Monte Carlo with iid $\{\xi^1, \xi^2, \ldots \xi^N\}$. In this case:

$$\mathrm{Var}[\hat{g}_N(x)] \; = \mathrm{Var}\left[\frac{1}{N}\sum_{j=1}^{N} G(x, \xi^j)\right] = \; \frac{\mathrm{Var}[G(x,\xi)]}{N}.$$

So. . . to decrease variance we can increase the sample size N

Why Variance Reduction?

Suppose we use standard Monte Carlo with iid $\{\xi^1, \xi^2, \ldots \xi^N\}$. In this case:

$$\mathrm{Var}[\hat{g}_N(x)] \ = \mathrm{Var}\left[\frac{1}{N}\sum_{j=1}^{N} G(x,\xi^j)\right] = \ \frac{\mathrm{Var}[G(x,\xi)]}{N}.$$

So. . . to decrease variance we can increase the sample size N

NOTE: Although $\mathrm{Var}[G(x,\xi)]$ is typically unknown, it can be estimated by a sample variance as follows:

$$S_N^2(x) \ := \ \frac{\sum_{j=1}^{N}[G(x,\xi^j) - \hat{g}_N(x)]^2}{N-1}.$$

The above estimator is unbiased, i.e., $\mathbb{E}[S_N^2(x)] = \mathrm{Var}[G(x,\xi)]$.

Why Variance Reduction?

However, increasing the sample size is not desirable:

1. Estimation of $G(x, \xi^j)$ for a fixed $x \in X$ could be very expensive

2. When we also optimize ($\min_{x \in X} \hat{g}_N(x)$), computational burden of optimization can significantly increase with N

Why Variance Reduction?

However, increasing the sample size is not desirable:

1. Estimation of $G(x, \xi^j)$ for a fixed $x \in X$ could be very expensive
2. When we also optimize $(\min_{x \in X} \hat{g}_N(x))$, computational burden of optimization can significantly increase with N

WANT: decrease variance without increasing sample size

- A well studied topic in statistics and simulation
- When we optimize $\min_{x \in X} \hat{g}_N(x)$, variance reduction can be more important
- Poor estimates of objective and/or gradients can slow down the convergence of Monte Carlo simulation-based methods

Variance Reduction Techniques

Variance Reduction Techniques

We will discuss:

- Antithetic Variates
- Latin Hypercube Sampling
- Quasi-Monte Carlo (QMC) and Randomized QMC
- Importance Sampling
- Control Variates

Variance Reduction Techniques

We will:

- Introduce these techniques in the context of estimating $G(x, \xi)$ for a fixed $x \in X$
- Point to literature that uses them for stochastic optimization
- Discuss the properties of resulting statistical esimators when used in the context of stochastic optimization

Variance Reduction Techniques: Common Themes

Some common themes to reduce variance:

1. Exploitation of Correlations
2. Sampling more "uniformly" than random sampling (or "filling in the space better")
3. Concentrating the sampling to important regions

Antithetic Variates (AV)

Antithetic Variates

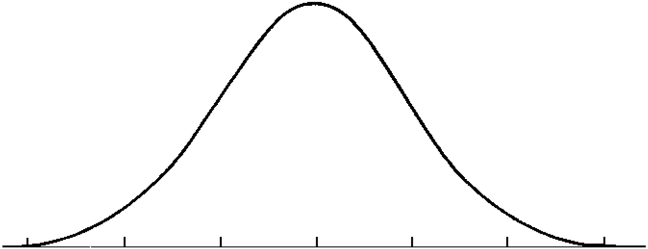
Idea: use pairs of negatively correlated random variables to reduce variance

Suppose N is even and components of ξ are independent:

1. Sample observations $\{U^1, \ldots, U^{\frac{N}{2}}\}$ from a $U(0,1)^{d_\xi}$ distribution.
2. Calculate the antithetic pairs $\{U^{1'}, \ldots, U^{\frac{N}{2}'}\} = \{1 - U^1, \ldots, 1 - U^{\frac{N}{2}}\}$.
3. Apply the inverse cumulative distribution function to obtain N observations $\{\xi^1, \xi^{1'},\ \ \xi^2, \xi^{2'} \ldots, \xi^{\frac{N}{2}}, \xi^{\frac{N}{2}'}\}$.

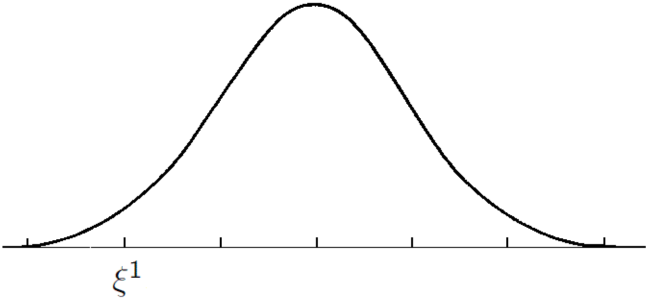
Antithetic Variates

Idea: use pairs of negatively correlated random variables to reduce variance



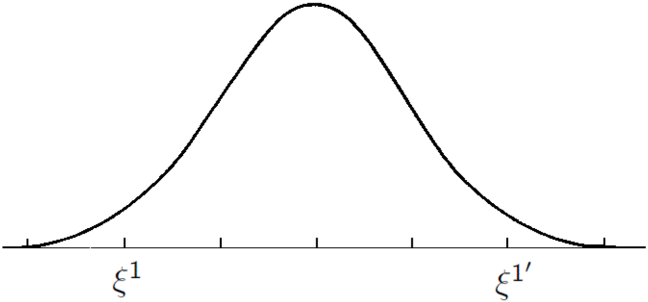
Antithetic Variates

Idea: use pairs of negatively correlated random variables to reduce variance



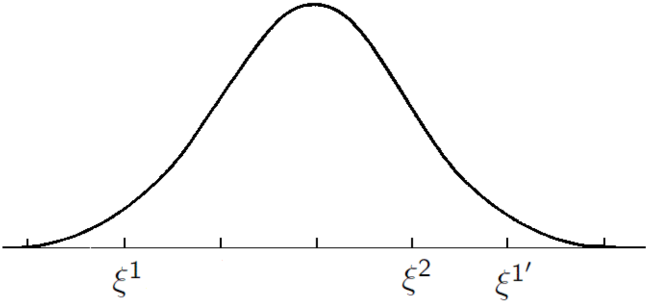
Antithetic Variates

Idea: use pairs of negatively correlated random variables to reduce variance



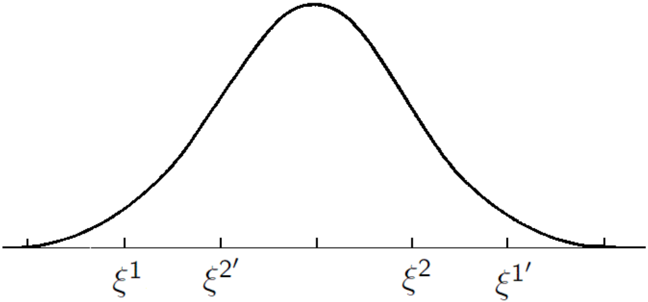
Antithetic Variates

Idea: use pairs of negatively correlated random variables to reduce variance



Antithetic Variates

Idea: use pairs of negatively correlated random variables to reduce variance



Antithetic Variates

AV estimator of $\mathbb{E}[G(x, \xi)]$ is:

$$\hat{g}_{N,\mathrm{AV}}(x) = \frac{1}{N/2} \sum_{j=1}^{N/2} \frac{G(x, \xi^j) + G(x, \xi^{j'})}{2}$$

Antithetic Variates

AV estimator of $\mathbb{E}[G(x,\xi)]$ is:

$$\hat{g}_{N,\mathrm{AV}}(x) = \frac{1}{N/2} \sum_{j=1}^{N/2} \frac{G(x,\xi^j) + G(x,\xi^{j'})}{2}$$

- Unbiased: $\mathbb{E}\left[\hat{g}_{N,\mathrm{AV}}(x)\right] = \mathbb{E}\left[G(x,\xi)\right]$
- Has variance:

$$\mathrm{Var}\left[\hat{g}_{N,\mathrm{AV}}(x)\right] = \underbrace{\frac{\mathrm{Var}\left[G(x,\xi)\right]}{N}}_{\text{Var of Standard MC!}} + \underbrace{\frac{1}{N}\mathrm{Cov}\left[G(x,\xi^j), G(x,\xi^{j'})\right]}_{\text{If Cov} < 0\text{, AV reduces var!}}$$

AV: Amount of Variance Reduction

The amount of variance reduction depends on how much negative correlation between U and $U^{'}$ is preserved when:

(i) Transforming to ξ and $\xi^{'}$

(ii) Applying $G(x, \cdot)$

It is well known that $G(x, \cdot)$ preserves negative correlation when:

(i) $G(x, \cdot)$ is bounded and monotone in each component of ξ

(ii) $G(x, \cdot)$ is not constant in the interior of Ξ

AV: Amount of Variance Reduction

The amount of variance reduction depends on how much negative correlation between U and $U^{'}$ is preserved when:

(i) Transforming to ξ and $\xi^{'}$

(ii) Applying $G(x, \cdot)$

It is well known that $G(x, \cdot)$ preserves negative correlation when:

(i) $G(x, \cdot)$ is bounded and monotone in each component of ξ

(ii) $G(x, \cdot)$ is not constant in the interior of Ξ

AV: Amount of Variance Reduction

When do we have monotononicity?

AV: Amount of Variance Reduction

When do we have monotononicity?

Example: Two-stage stochastic linear programs with recourse

$$\begin{aligned} \min_x \quad & \mathbb{E}\left[G(x,\xi)\right] = \mathbb{E}\left[cx + h(x,\xi)\right] \\ \text{s.t.} \quad & Ax = b, \quad x \geq 0, \end{aligned}$$

where $h(x,\xi)$ is the optimal value of the linear program

$$\begin{aligned} h(x,\xi) = \min_y \quad & \tilde{q}y \\ \text{s.t.} \quad & Wy \geq \tilde{r} - Tx, \quad y \geq 0. \end{aligned}$$

Here, ξ is a random vector that is comprised of random elements of $\tilde{q}$, $\tilde{r}$.

AV: Amount of Variance Reduction

When do we have monotononicity?

Example: Two-stage stochastic linear programs with recourse

$$\begin{aligned} \min_x \quad & \mathbb{E}\left[G(x,\xi)\right] = \mathbb{E}\left[cx + h(x,\xi)\right] \\ \text{s.t.} \quad & Ax = b, \quad x \geq 0, \end{aligned}$$

where $h(x,\xi)$ is the optimal value of the linear program

$$\begin{aligned} h(x,\xi) = \min_y \quad & \tilde{q}y \\ \text{s.t.} \quad & Wy \geq \tilde{r} - Tx, \quad y \geq 0. \end{aligned}$$

Here, ξ is a random vector that is comprised of random elements of $\tilde{q}$, $\tilde{r}$.

AV: Amount of Variance Reduction

When do we have monotononicity?

Example: Two-stage stochastic linear programs with recourse

$$\begin{aligned} \min_{x} \quad & \mathbb{E}\left[G(x,\xi)\right] = \mathbb{E}\left[cx + h(x,\xi)\right] \\ \text{s.t.} \quad & Ax = b, \quad x \geq 0, \end{aligned}$$

where $h(x,\xi)$ is the optimal value of the linear program

$$\begin{aligned} h(x,\xi) = \min_{y} \quad & \tilde{q}y \\ \text{s.t.} \quad & Wy \geq \tilde{r} - Tx, \quad y \geq 0. \end{aligned}$$

Here, ξ is a random vector that is comprised of random elements of $\tilde{q}$, $\tilde{r}$.

AV: Amount of Variance Reduction

Example Cont'd: Two-Stage Stochastic Linear Programs are Monotone when:

- The recourse matrix W is fixed
- The recourse decision vector y is non-negative
- Constraints are defined using inequalities
- The components of $\tilde{\xi}$ are independent

AV: Amount of Variance Reduction

Example Cont'd: Two-Stage Stochastic Linear Programs are Monotone when:

- The recourse matrix W is fixed
- The recourse decision vector y is non-negative
- Constraints are defined using inequalities
- The components of $\tilde{\xi}$ are independent

Be careful! AV can backfire

shows when monotonicity is lost, AV can increase variance

AV: In the Context of Optimization

When used for optimization:

- Analytically show the extent of variance reduction using AV on a newsvendor problem
  - Decreases bias of ν_N $\left(\mathbb{E}\left[\nu_N\right] \leq \nu^*\right)$
  - Can increase or decrease variance depending on parameters

AV: In the Context of Optimization

When used for optimization:

- **?** analytically show the extent of variance reduction using AV on a newsvendor problem
  - Decreases bias of ν_N $(\mathbb{E}[\nu_N] \leq \nu^*)$ –A nice side effect!
  - Can increase or decrease variance depending on parameters

AV: In the Context of Optimization

When used for optimization:

- **?** analytically show the extent of variance reduction using AV on a newsvendor problem
  - Decreases bias of ν_N $(\mathbb{E}[\nu_N] \le \nu^*)$ –A nice side effect!
  - Can increase or decrease variance depending on parameters
- When AV is effective, **?** uses it with other variance reduction techniques (QMC)
- Several studies show modest variance and bias reduction for most two-stage stochastic linear problems (**??**)

AV: In the Context of Optimization

When used for optimization:

- **?** analytically show the extent of variance reduction using AV on a newsvendor problem
  - Decreases bias of ν_N $\quad(\mathbb{E}\left[\nu_N\right] \leq \nu^*)$ –A nice side effect!
  - Can increase or decrease variance depending on parameters
- When AV is effective, uses it with other variance reduction techniques (QMC)
- Several studies show modest variance and bias reduction for most two-stage stochastic linear problems ()

Latin Hypercube Sampling (LHS)

Latin Hypercube Sampling

Idea: use statified sampling but in a less computationally intensive way

Suppose components of ξ are independent:

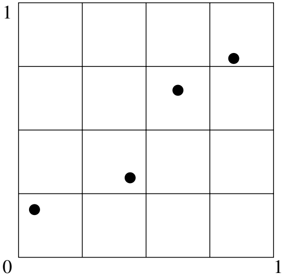
1. For each component of $\tilde{\xi}$:
   1.1 Sample observations $U^i \sim U(\frac{i-1}{N}, \frac{i}{N})$ for $i = 1, \ldots, N$.

   1.2 Randomly permute these N observations.
2. Apply the inverse cumulative distribution function to obtain $\{\xi^1, \ldots, \xi^N\}$.

Latin Hypercube Sampling

Idea: use statified sampling but in a less computationally intensive way

Example: Dimension $d_\xi = 2$ and sample size $N = 4$



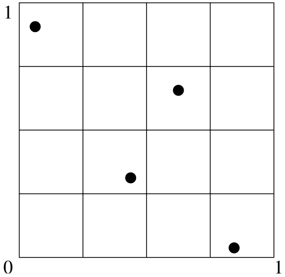
(a) Stratified

(Images:

Latin Hypercube Sampling

Idea: use statified sampling but in a less computationally intensive way

Example: Dimension $d_\xi = 2$ and sample size $N = 4$



(b) LHS

(Images:

Properties of LHS

- Each $G(x, \xi_{\mathcal{L}}^{j})$ is unbiased: $\mathbb{E}\left[G(x, \xi_{\mathcal{L}}^{j})\right] = \mathbb{E}\left[G(x, \xi)\right]$

Properties of LHS

- Each $G(x, \xi_{\mathcal{L}}^{j})$ is unbiased: $\mathbb{E}\left[G(x, \xi_{\mathcal{L}}^{j})\right] = \mathbb{E}\left[G(x, \xi)\right]$
- However, $G(x, \xi_{\mathcal{L}}^{j})$, $j = 1, 2 \ldots N$ are not independent!

Properties of LHS

- Each $G(x, \xi_{\mathcal{L}}^{j})$ is unbiased: $\mathbb{E}\left[G(x, \xi_{\mathcal{L}}^{j})\right] = \mathbb{E}\left[G(x, \xi)\right]$
- However, $G(x, \xi_{\mathcal{L}}^{j})$, $j = 1, 2 \ldots N$ are not independent!
  In fact, this correlation reduces variance!

Properties of LHS

- Each $G(x, \xi_{\mathcal{L}}^{j})$ is unbiased: $\mathbb{E}\left[G(x, \xi_{\mathcal{L}}^{j})\right] = \mathbb{E}\left[G(x, \xi)\right]$
- However, $G(x, \xi_{\mathcal{L}}^{j})$, $j = 1, 2 \ldots N$ are not independent!
  In fact, this correlation reduces variance!
- LHS estimator

$$\hat{g}_{N,\mathrm{LHS}}(x) = \frac{1}{N} \sum_{j=1}^{N} G(x, \xi_{\mathcal{L}}^{j})$$

is unbiased: $\mathbb{E}\left[\hat{g}_{N,\mathrm{LHS}}(x)\right] = \mathbb{E}\left[G(x, \xi)\right]$

Properties of LHS

- Each $G(x, \xi^j_{\mathcal{L}})$ is unbiased: $\mathbb{E}\left[G(x, \xi^j_{\mathcal{L}})\right] = \mathbb{E}\left[G(x, \xi)\right]$
- However, $G(x, \xi^j_{\mathcal{L}})$, $j = 1, 2 \dots N$ are not independent!
  In fact, this correlation reduces variance!
- LHS estimator

$$\hat{g}_{N,\mathrm{LHS}}(x) = \frac{1}{N} \sum_{j=1}^{N} G(x, \xi^j_{\mathcal{L}})$$

  is unbiased: $\mathbb{E}\left[\hat{g}_{N,\mathrm{LHS}}(x)\right] = \mathbb{E}\left[G(x, \xi)\right]$

- If $G(x, \cdot)$ monotone in each component of ξ, then
  $\mathrm{Var}\left[\hat{g}_{N,\mathrm{LHS}}(x)\right] \leq \mathrm{Var}\left[\hat{g}_N(x)\right]$

Properties of LHS

- Each $G(x, \xi_{\mathcal{L}}^{j})$ is unbiased: $\mathbb{E}\left[G(x, \xi_{\mathcal{L}}^{j})\right] = \mathbb{E}\left[G(x, \xi)\right]$
- However, $G(x, \xi_{\mathcal{L}}^{j})$, $j = 1, 2 \ldots N$ are not independent! In fact, this correlation reduces variance!
- LHS estimator

$$\hat{g}_{N,\mathrm{LHS}}(x) = \frac{1}{N} \sum_{j=1}^{N} G(x, \xi_{\mathcal{L}}^{j})$$

  is unbiased: $\mathbb{E}\left[\hat{g}_{N,\mathrm{LHS}}(x)\right] = \mathbb{E}\left[G(x, \xi)\right]$

- If $G(x, \cdot)$ monotone in each component of ξ, then $\mathrm{Var}\left[\hat{g}_{N,\mathrm{LHS}}(x)\right] \leq \mathrm{Var}\left[\hat{g}_{N}(x)\right]$
- More generally

$$\mathrm{Var}\left[\hat{g}_{N,\mathrm{LHS}}(x)\right] \leq \frac{N}{N-1} \mathrm{Var}\left[\hat{g}_{N}(x)\right]$$

LHS: In the Context of Optimization

When used for optimization:

- Analytically show the extent of variance reduction using LHS on a newsvendor problem
  - Completely removes bias of ν_N $\left(\mathbb{E}\left[\nu_N\right] = \nu^*\right)$
  - Decreases variance considerably

LHS: In the Context of Optimization

When used for optimization:

- **?** analytically show the extent of variance reduction using LHS on a newsvendor problem
    - Completely removes bias of ν_N $\left(\mathbb{E}\left[\nu_N\right] = \nu^*\right)$ –A really nice side effect!
    - Decreases variance considerably

LHS: In the Context of Optimization

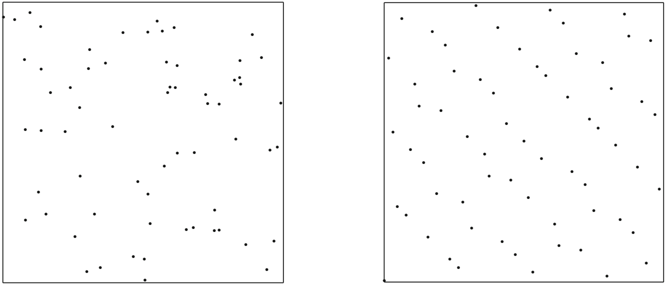
When used for optimization:

- **?** analytically show the extent of variance reduction using LHS on a newsvendor problem
  - Completely removes bias of ν_N $\left(\mathbb{E}\left[\nu_N\right] = \nu^*\right)$ –A really nice side effect!
  - Decreases variance considerably
- Many studies show LHS effectively reduces variance and bias for several classes of stochastic programs; e.g., (**????**)

Quasi-Monte Carlo (QMC) and Randomized Quasi-Monte Carlo (RQMC)

Quasi-Monte Carlo

Idea: use a low-discrepancy sequence to sample more uniformly. The sequence doesn't have to be independent or random.



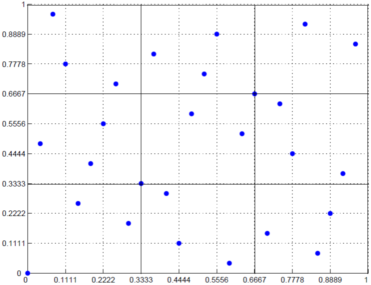
(a) Random Sequence

(b) Sobol' Sequence

(Images: Lemieux (2009))

Quasi-Monte Carlo

Idea: use a low-discrepancy sequence to sample more uniformly. The sequence doesn't have to be independent or random.



(c) (0,3,2)-net in base b=3

(Image: Homem-de-Mello and Bayraksan (2014))

Quasi-Monte Carlo

Many different ways to generate these sequences:

- Sobol'
- Niederreiter
- Lattice rules
- Digital nets and sequences
- . . .

For simplicity, we will assume ξ is a random vector with independent components, each with uniform distribution over $[0, 1]$

Quasi-Monte Carlo

How to measure the quality (or “uniformity”) of these sequences?

Quasi-Monte Carlo

How to measure the quality (or “uniformity”) of these sequences?

Star Discrepancy

- Consider a d_ξ dimensional unit cube $[0,1)^{d_\xi}$
- A point set $P = \{\xi^j\}_{j=1}^{\infty}$ inside $[0,1)^{d_\xi}$
- P_N denotes the first N points from the point set P
- Consider hyper-rectangles with a corner at the origin

$$B(\mathbf{v}) = \Pi_{i=1}^{d_\xi}[0, v_i]$$

where $\mathbf{v}$ is a vector $\left(v_1 \ v_2 \ \dots v_{d_\xi}\right)$

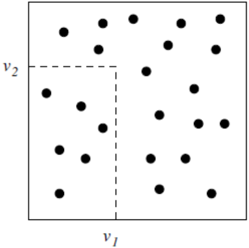
Star Discrepancy

Star Discrepancy is the Kolmogorov-Smirnov distance between the point set and the uniform distribution over the unit cube:

$$D^*(P_N) := \sup_{\mathbf{v}\in[0,1)^{d_\xi}} \left| \frac{\text{Number of } \xi^j \in B(\mathbf{v})}{N} - \Pi_{i=1}^{d_\xi} v_i \right|$$

Discrepancy

Example (Lemieux, 2009):



$v_1 = 0.4$, $v_2 = 0.7$ and $6/23$ points inside the box, gives a discrepancy of $|6/23 - 0.4 \times 0.7| = 0.019$

What About Star Discrepancy?

- Conjecture that for deterministic sequences, the best that can be achieved is $D^*(P_N) \in O\left(\frac{(\log N)^{d_\xi}}{N}\right)$

What About Star Discrepancy?

- Conjecture that for deterministic sequences, the best that can be achieved is $D^*(P_N) \in O\left(\frac{(\log N)^{d_\xi}}{N}\right)$
- Sequences that achieve this bound are known and called **low-discrepancy sequences**

What About Star Discrepancy?

- Conjecture that for deterministic sequences, the best that can be achieved is $D^*(P_N) \in O\left(\frac{(\log N)^{d_\xi}}{N}\right)$

- Sequences that achieve this bound are known and called **low-discrepancy sequences**

- *Koksma-Hlawka inequality* gives a bound on the error

$$|\hat{g}_{N,\mathrm{QMC}}(x) - g(x)| \leq D^*(P_N) V(G(x,\cdot))$$

  where $V(G(x,\cdot))$ is the total variation in the sense of Hardy and Krause

What About Star Discrepancy?

- Conjecture that for deterministic sequences, the best that can be achieved is $D^*(P_N) \in O\left(\frac{(\log N)^{d_\xi}}{N}\right)$
- Sequences that achieve this bound are known and called **low-discrepancy sequences**
- *Koksma-Hlawka inequality* gives a bound on the error

$$|\hat{g}_{N,\mathrm{QMC}}(x) - g(x)| \leq D^*(P_N)V(G(x,\cdot))$$

where $V(G(x,\cdot))$ is the total variation in the sense of Hardy and Krause

What About Star Discrepancy?

- Conjecture that for deterministic sequences, the best that can be achieved is $D^*(P_N) \in O\left(\frac{(\log N)^{d_\xi}}{N}\right)$
- Sequences that achieve this bound are known and called **low-discrepancy sequences**
- *Koksma-Hlawka inequality* gives a bound on the error

$$|\hat{g}_{N,\mathrm{QMC}}(x) - g(x)| \leq D^*(P_N) V(G(x, \cdot))$$

where $V(G(x, \cdot))$ is the total variation in the sense of Hardy and Krause

What About Star Discrepancy?

- Conjecture that for deterministic sequences, the best that can be achieved is $D^*(P_N) \in O\left(\frac{(\log N)^{d_\xi}}{N}\right)$
- Sequences that achieve this bound are known and called **low-discrepancy sequences**
- *Koksma-Hlawka inequality* gives a bound on the error

$$|\hat{g}_{N,\mathrm{QMC}}(x) - g(x)| \leq D^*(P_N) V(G(x,\cdot))$$

where $V(G(x,\cdot))$ is the total variation in the sense of Hardy and Krause

Low-Discrepancy Sequences

So if $V(G(x,\cdot)) < \infty$ and we use a low-discrepancy sequence

$$|\hat{g}_{N,\mathrm{QMC}}(x) - g(x)| \in O\left(\frac{(\log N)^{d_\xi}}{N}\right)$$

Low-Discrepancy Sequences

So if $V(G(x,\cdot)) < \infty$ and we use a low-discrepancy sequence

$$|\hat{g}_{N,\mathrm{QMC}}(x) - g(x)| \in O\left(\frac{(\log N)^{d_\xi}}{N}\right)$$

Low-Discrepancy Sequences

So if $V(G(x,\cdot)) < \infty$ and we use a low-discrepancy sequence

$$|\hat{g}_{N,\mathrm{QMC}}(x) - g(x)| \in O\left(\frac{(\log N)^{d_\xi}}{N}\right)$$

Compare with Standard Monte Carlo error bound: $O\left(\frac{1}{\sqrt{N}}\right)$

Low-Discrepancy Sequences

So if $V(G(x,\cdot)) < \infty$ and we use a low-discrepancy sequence

$$|\hat{g}_{N,\mathrm{QMC}}(x) - g(x)| \in O\left(\frac{(\log N)^{d_\xi}}{N}\right)$$

Compare with Standard Monte Carlo error bound: $O\left(\frac{1}{\sqrt{N}}\right)$

Observations:

Low-Discrepancy Sequences

So if $V(G(x,\cdot)) < \infty$ and we use a low-discrepancy sequence

$$|\hat{g}_{N,\mathrm{QMC}}(x) - g(x)| \in O\left(\frac{(\log N)^{d_\xi}}{N}\right)$$

Compare with Standard Monte Carlo error bound: $O\left(\frac{1}{\sqrt{N}}\right)$

Observations:

Low-Discrepancy Sequences

So if $V(G(x,\cdot)) < \infty$ and we use a low-discrepancy sequence

$$|\hat{g}_{N,\mathrm{QMC}}(x) - g(x)| \in O\left(\frac{(\log N)^{d_\xi}}{N}\right)$$

Compare with Standard Monte Carlo error bound: $O\left(\frac{1}{\sqrt{N}}\right)$

Observations:

- MC error bound is free of dimension

Low-Discrepancy Sequences

So if $V(G(x,\cdot)) < \infty$ and we use a low-discrepancy sequence

$$|\hat{g}_{N,\mathrm{QMC}}(x) - g(x)| \in O\left(\frac{(\log N)^{d_\xi}}{N}\right)$$

Compare with Standard Monte Carlo error bound: $O\left(\frac{1}{\sqrt{N}}\right)$

Observations:

- MC error bound is free of dimension
- If N is large and d_ξ is small, then, QMC is expected to give a better approximation

Low-Discrepancy Sequences

So if $V(G(x,\cdot)) < \infty$ and we use a low-discrepancy sequence

$$|\hat{g}_{N,\mathrm{QMC}}(x) - g(x)| \in O\left(\frac{(\log N)^{d_\xi}}{N}\right)$$

Compare with Standard Monte Carlo error bound: $O\left(\frac{1}{\sqrt{N}}\right)$

Observations:

- MC error bound is free of dimension
- If N is large and d_ξ is small, then, QMC is expected to give a better approximation
- For large d_ξ, QMC can backfire! (esp. with low N)

Low-Discrepancy Sequences

So if $V(G(x,\cdot)) < \infty$ and we use a low-discrepancy sequence

$$|\hat{g}_{N,\mathrm{QMC}}(x) - g(x)| \in O\left(\frac{(\log N)^{d_\xi}}{N}\right)$$

Compare with Standard Monte Carlo error bound: $O\left(\frac{1}{\sqrt{N}}\right)$

Observations:

- MC error bound is free of dimension
- If N is large and d_ξ is small, then, QMC is expected to give a better approximation
- For large d_ξ, QMC can backfire! (esp. with low N)
- One way to deal with this is to find the "effective" dimension of the problem and use LHS or Standard MC on the other dimensions; e.g., (Drew and Homem-de-Mello, 2006; Drew, 2007)

Randomized Quasi-Monte Carlo

Koksma-Hlawka inequality is difficult to compute, is only a bound; so, how to know the errors?

Randomized Quasi-Monte Carlo

Koksma-Hlawka inequality is difficult to compute, is only a bound; so, how to know the errors?

→ Include randomness to estimate errors

Randomized Quasi-Monte Carlo

Koksma-Hlawka inequality is difficult to compute, is only a bound; so, how to know the errors?

→ Include randomness to estimate errors

We want the randomization to be performed so that

(i) Each point follows Uniform distribution over the unit cube

(ii) Low discrepancy is preserved

Randomized Quasi-Monte Carlo

Koksma-Hlawka inequality is difficult to compute, is only a bound; so, how to know the errors?

→ Include randomness to estimate errors

We want the randomization to be performed so that

(i) Each point follows Uniform distribution over the unit cube

→So that the resulting estimator is unbiased!

(ii) Low discrepancy is preserved

Randomized Quasi-Monte Carlo

Koksma-Hlawka inequality is difficult to compute, is only a bound; so, how to know the errors?

→ Include randomness to estimate errors

We want the randomization to be performed so that

(i) Each point follows Uniform distribution over the unit cube

→So that the resulting estimator is unbiased!

(ii) Low discrepancy is preserved

→So that we do better than standard MC!

Randomized Quasi-Monte Carlo

Koksma-Hlawka inequality is difficult to compute, is only a bound; so, how to know the errors?

→ Include randomness to estimate errors

We want the randomization to be performed so that

(i) Each point follows Uniform distribution over the unit cube

→So that the resulting estimator is unbiased!

(ii) Low discrepancy is preserved

→So that we do better than standard MC!

Many different ways to do randomization

- Random shift
- Digital shift (for digital nets)
- Scrambling and Permutations, . . .

RQMC: Cranley Patterson Random Shift

Generate a random vector $\mathbf{u} \sim U([0.1)^{d_\xi})$

Let

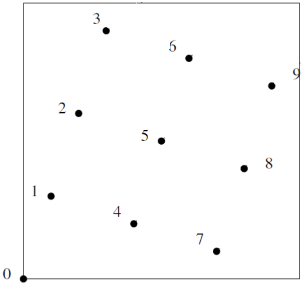
$$\tilde{\xi}^j = (\xi^j + \mathbf{u}) \mod 1$$

RQMC: Cranley Patterson Random Shift

Generate a random vector $\mathbf{u} \sim U([0.1)^{d_\xi})$

Let

$$\tilde{\xi}^j = (\xi^j + \mathbf{u}) \mod 1$$

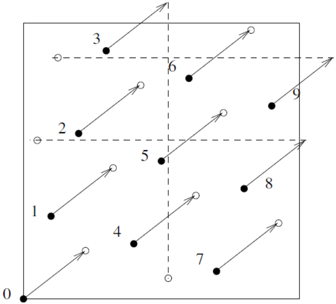


RQMC: Cranley Patterson Random Shift

Generate a random vector $\mathbf{u} \sim U([0.1)^{d_\xi})$

Let

$$\tilde{\xi}^j = (\xi^j + \mathbf{u}) \mod 1$$

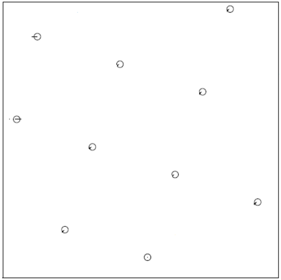


RQMC: Cranley Patterson Random Shift

Generate a random vector $\mathbf{u} \sim U([0.1)^{d_\xi})$

Let

$$\tilde{\xi}^j = (\xi^j + \mathbf{u}) \mod 1$$



RQMC: Estimation of Error

Generate m iid RQMC sequences $\tilde{P}_N$

$$\hat{g}^{l}_{N,\mathrm{RQMC}}(x) = \frac{1}{N}\sum_{j=1}^{N} G(x, \tilde{\xi}^{j}), \quad l = 1, 2, \ldots m$$

and use the estimator

$$\hat{g}_{N,\mathrm{RQMC}}(x) = \frac{1}{m}\sum_{l=1}^{m} \hat{g}^{l}_{N,\mathrm{RQMC}}(x)$$

RQMC: Estimation of Error

Generate m iid RQMC sequences $\tilde{P}_N$

$$\hat{g}^l_{N,\mathrm{RQMC}}(x) = \frac{1}{N}\sum_{j=1}^{N} G(x,\tilde{\xi}^j), \quad l = 1, 2, \dots m$$

and use the estimator

$$\hat{g}_{N,\mathrm{RQMC}}(x) = \frac{1}{m}\sum_{l=1}^{m} \hat{g}^l_{N,\mathrm{RQMC}}(x)$$

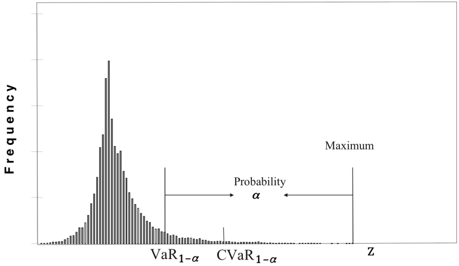
- This estimator is unbiased: $\mathbb{E}\left[\hat{g}_{N,\mathrm{RQMC}}(x)\right] = g(x)$
- And the error can be estimated through the variance

$$\frac{1}{m-1}\sum_{l=1}^{m}\left(\hat{g}^l_{N,\mathrm{RQMC}}(x) - \hat{g}_{N,\mathrm{RQMC}}(x)\right)^2$$

Importance Sampling (IS)

Importance Sampling: Motivation

Idea: Concentrate samples to most important areas. Especially can be useful for "rare" events.



$$\mathrm{CVaR}_{1-\alpha}[G(x,\xi)] = \min_{\eta}\left\{\eta + \frac{1}{\alpha}\mathbb{E}\left[(G(x,\xi)-\eta)_+\right]\right\}$$

If $\alpha = 0.05$, 95% of the standard Monte Carlo samples do not contribute to positive values in this expectation!

IS: How does it work?

For simplicity, suppose ξ has density f. Then,

$$g(x) = \mathbb{E}[G(x,\xi)] = \int_{\Xi} G(x,\xi) f(\xi) d\xi$$

IS: How does it work?

For simplicity, suppose ξ has density f. Then,

$$g(x) = \mathbb{E}[G(x,\xi)] = \int_{\Xi} G(x,\xi) f(\xi) d\xi$$

Now consider another density q over Ξ such that $q(E) = 0$ for every set E for which $f(E) = 0$ and rewrite

$$\mathbb{E}[G(x,\xi)] = \int_{\Xi} G(x,\xi) \mathcal{L}(\xi) q(\xi) d\xi$$

IS: How does it work?

For simplicity, suppose ξ has density f. Then,

$$g(x) = \mathbb{E}[G(x,\xi)] = \int_{\Xi} G(x,\xi) f(\xi) d\xi$$

Now consider another density q over Ξ such that $q(E) = 0$ for every set E for which $f(E) = 0$ and rewrite

$$\mathbb{E}[G(x,\xi)] = \int_{\Xi} G(x,\xi) \mathcal{L}(\xi) q(\xi) d\xi$$

IS: How does it work?

For simplicity, suppose ξ has density f. Then,

$$g(x) = \mathbb{E}[G(x,\xi)] = \int_{\Xi} G(x,\xi) f(\xi) d\xi$$

Now consider another density q over Ξ such that $q(E) = 0$ for every set E for which $f(E) = 0$ and rewrite

$$\mathbb{E}[G(x,\xi)] = \int_{\Xi} G(x,\xi) \mathcal{L}(\xi) q(\xi) d\xi$$

Here, $\mathcal{L}(\xi) = \frac{f(\xi)}{q(\xi)}$ is the likelihood ratio, which we assume is well defined
(for this, we may set $\mathcal{L}$ to zero whenever both f and q are zero).

IS: How does it work?

Standard MC: iid sample $\{\xi^1, \xi^2, \ldots, \xi^N\}$ from density f

Importance Sampling: iid sample $\{\tilde{\xi}_q^1, \tilde{\xi}_q^2, \ldots, \tilde{\xi}_q^N\}$ from density q

IS: How does it work?

Standard MC: iid sample $\{\xi^1, \xi^2, \ldots, \xi^N\}$ from density f

Importance Sampling: iid sample $\{\tilde{\xi}_q^1, \tilde{\xi}_q^2, \ldots, \tilde{\xi}_q^N\}$ from density q

IS: How does it work?

Standard MC: iid sample $\{\xi^1, \xi^2, \dots, \xi^N\}$ from density f

Importance Sampling: iid sample $\{\tilde{\xi}_q^1, \tilde{\xi}_q^2, \dots, \tilde{\xi}_q^N\}$ from density q

The importance sampling estimator:

$$\hat{g}_{N,\mathrm{IS}}(x) = \frac{1}{N} \sum_{j=1}^{N} G(x, \tilde{\xi}_q^j) \mathcal{L}(\tilde{\xi}_q^j)$$

- Unbiased: $\mathbb{E}\left[\hat{g}_{N,\mathrm{IS}}(x)\right] = g(x)$

Variance of the IS Estimator

How to obtain q?

Variance of the IS Estimator

How to obtain q?

$\rightarrow$ To reduce variance! To understand this better, let's take a look at the variance of the IS estimator.

Variance of the IS Estimator

How to obtain q?

→ To reduce variance! To understand this better, let's take a look at the variance of the IS estimator.

Consider the following facts:

- $\mathbb{E}[G(x,\tilde{\xi}_q)\mathcal{L}(\tilde{\xi}_q)] = \mathbb{E}[G(x,\xi)]$
- $\mathbb{E}[G^2(x,\tilde{\xi}_q)\mathcal{L}^2(\tilde{\xi}_q)] = \mathbb{E}[G^2(x,\xi)\mathcal{L}(\xi)]$
- Therefore, the variance of the IS estimator is

$$\mathrm{Var}\left[\hat{g}_{N,\mathrm{IS}}(x)\right] = \frac{1}{N}\left[\mathbb{E}[G^2(x,\xi)\mathcal{L}(\xi)] - (\mathbb{E}[G(x,\xi)])^2\right],$$

Variance of the IS Estimator

How to obtain q?

→ To reduce variance! To understand this better, let's take a look at the variance of the IS estimator.

Consider the following facts:

- $\mathbb{E}[G(x,\tilde{\xi}_q)\mathcal{L}(\tilde{\xi}_q)] = \mathbb{E}[G(x,\xi)]$
- $\mathbb{E}[G^2(x,\tilde{\xi}_q)\mathcal{L}^2(\tilde{\xi}_q)] = \mathbb{E}[G^2(x,\xi)\mathcal{L}(\xi)]$
- Therefore, the variance of the IS estimator is

$$\mathrm{Var}\left[\hat{g}_{N,\mathrm{IS}}(x)\right] = \frac{1}{N}\left[\mathbb{E}[G^2(x,\xi)\mathcal{L}(\xi)] - (\mathbb{E}[G(x,\xi)])^2\right],$$

→ Reduce Variance when $\mathbb{E}[G^2(x,\xi)\mathcal{L}(\xi)] < \mathbb{E}[G^2(x,\xi)]$

How to obtain IS distribution?

When $G(x,\cdot) \geq 0$, setting $\mathrm{Var}\left[\hat{g}_{N,\mathrm{IS}}(x)\right] = 0$ results in the **optimal zero-variance density**

$$\mathbf{q}^*(\xi) = \frac{f(\xi)G(x,\xi)}{\mathbb{E}[G(x,\xi)]}$$

How to obtain IS distribution?

When $G(x, \cdot) \geq 0$, setting $\mathrm{Var}\left[\hat{g}_{N,\mathrm{IS}}(x)\right] = 0$ results in the **optimal zero-variance density**

$$\mathbf{q}^*(\xi) = \frac{f(\xi)G(x,\xi)}{\mathbb{E}[G(x,\xi)]}$$

Too good to be true?

How to obtain IS distribution?

When $G(x,\cdot) \geq 0$, setting $\mathrm{Var}\left[\hat{g}_{N,\mathrm{IS}}(x)\right] = 0$ results in the **optimal zero-variance density**

$$\mathbf{q}^*(\xi) = \frac{f(\xi)G(x,\xi)}{\mathbb{E}[G(x,\xi)]}$$

Too good to be true? YES! Requires knowledge of unknown quantity!

Nevertheless, select q

$$q(\xi) \propto f(\xi)G(x,\xi)$$

to achieve variance reduction even though the proportionality constant may not be known

How to obtain IS distribution?

BE Careful! If q is not chosen properly, IS can backfire! Can actually increase the variance!

How to obtain IS distribution?

BE Careful! If q is not chosen properly, IS can backfire! Can actually increase the variance!

Also q should be easy to sample from

How to obtain IS distribution?

BE Careful! If q is not chosen properly, IS can backfire! Can actually increase the variance!

Also q should be easy to sample from

Many different ways to obtain the IS distribution

- Exponential Tilting
- Using Large Deviations
- Nonparametric methods, . . .
- Exploit the problem structure
- Still an open area of research

IS: Examples of Use in Optimization

Kozmík and Morton (2015) apply IS to solve multistage stochastic programs with Mean-CVaR objectives. Here is a 'gist':

IS: Examples of Use in Optimization

Kozmík and Morton (2015) apply IS to solve multistage stochastic programs with Mean-CVaR objectives. Here is a 'gist':

- Suppose ξ has a finite support, taking $|\Xi|$ realizations
- The 'nominal' distribution puts equal mass on each point. The nominal probability mass function (pmf) is: $f(\xi) = \frac{1}{|\Xi|}$

IS: Examples of Use in Optimization

Kozmík and Morton (2015) apply IS to solve multistage stochastic programs with Mean-CVaR objectives. Here is a 'gist':

- Suppose ξ has a finite support, taking $|\Xi|$ realizations
- The 'nominal' distribution puts equal mass on each point. The nominal probability mass function (pmf) is: $f(\xi) = \frac{1}{|\Xi|}$
- Instead of the expensive evaluations $G(x, \xi)$, suppose there is a good *approximation function* $H(x, \xi)$ that:
  - estimates the value of $G(x, \xi)$ cheaply and
  - orders the values $G(x, \xi)$ in the same way

IS: Examples of Use in Optimization

- Suppose at a given x, the Value at Risk at level $1 - \alpha$ of the approximation function is obtained $\rightarrow$ Let's denote it V_H

IS: Examples of Use in Optimization

- Suppose at a given x, the Value at Risk at level $1 - \alpha$ of the approximation function is obtained $\rightarrow$ Let's denote it V_H

IS: Examples of Use in Optimization

- Suppose at a given x, the Value at Risk at level $1 - \alpha$ of the approximation function is obtained $\rightarrow$ Let's denote it V_H
- The 'IS' pmf is

$$g(\xi) = \begin{cases} \frac{1}{2}\frac{1}{\lfloor \alpha |\Xi| \rfloor}, & \xi : H(x,\xi) \geq V_H \\ \frac{1}{2}\frac{1}{|\Xi| - \lfloor \alpha |\Xi| \rfloor}, & \xi : H(x,\xi) < V_H \end{cases}$$

NOTE: The IS distribution depends on x

IS: Examples of Use in Optimization

- Suppose at a given x, the Value at Risk at level $1-\alpha$ of the approximation function is obtained $\quad\rightarrow$ Let's denote it V_H
- The 'IS' pmf is

$$g(\xi) = \begin{cases} \frac{1}{2}\frac{1}{\lfloor \alpha|\Xi| \rfloor}, & \xi : H(x,\xi) \geq V_H \\ \frac{1}{2}\frac{1}{|\Xi| - \lfloor \alpha|\Xi| \rfloor}, & \xi : H(x,\xi) < V_H \end{cases}$$

NOTE: The IS distribution depends on x

IS: Examples of Use in Optimization

- Suppose at a given x, the Value at Risk at level $1-\alpha$ of the approximation function is obtained $\quad\rightarrow$ Let's denote it V_H
- The 'IS' pmf is

$$g(\xi) = \begin{cases} \frac{1}{2}\frac{1}{\lfloor \alpha|\Xi| \rfloor}, & \xi : H(x,\xi) \geq V_H \\ \frac{1}{2}\frac{1}{|\Xi| - \lfloor \alpha|\Xi| \rfloor}, & \xi : H(x,\xi) < V_H \end{cases}$$

NOTE: The IS distribution depends on x

IS: Examples of Use in Optimization

- Barrera et al. (2016) apply IS to a chance constrained problem with rare probabilities.
- They show consistency of the SAA problem formed using IS (optimal values converge, etc.)
- They make two important improvements:
  - The IS distribution typically depends on ξ. They enhance it to depend on x as well.
  - Find an IS distribution that works for a set of x
- Adapted IS yields significantly better results, but the resulting problem can get difficult to solve

IS: Examples of Use in Optimization

- Barrera et al. (2016) apply IS to a chance constrained problem with rare probabilities.
- They show consistency of the SAA problem formed using IS (optimal values converge, etc.)
- They make two important improvements:
  - The IS distribution typically depends on ξ. They enhance it to depend on x as well.

    Think of $\mathcal{L}_x(\xi)$ instead of $\mathcal{L}(\xi)$
  - Find an IS distribution that works for a set of x
- Adapted IS yields significantly better results, but the resulting problem can get difficult to solve

Control Variates (CV)

Control Variates

Idea: reduce variance by inducing correlations

Control Variates

Idea: reduce variance by inducing correlations

Let C be a *control variable* with

- $\mathbb{E}[C] = 0$
- C is correlated with $G(x, \xi)$ —can be positively or negatively correlated

Control Variates

Idea: reduce variance by inducing correlations

Let C be a *control variable* with

- $\mathbb{E}[C] = 0$
- C is correlated with $G(x, \xi)$ —can be positively or negatively correlated

Control Variates

Idea: reduce variance by inducing correlations

Let C be a *control variable* with

- $\mathbb{E}[C] = 0$
- C is correlated with $G(x, \xi)$ —can be positively or negatively correlated

Control Variates

Idea: reduce variance by inducing correlations

Let C be a *control variable* with

- $\mathbb{E}[C] = 0$
- C is correlated with $G(x, \xi)$ —can be positively or negatively correlated

With λ a scalar, the control variate estimator of $\mathbb{E}[G(x, \xi)]$ is given by

$$\hat{g}_{N,\mathrm{CV}}(x) = \frac{1}{N} \sum_{j=1}^{N} \left(G(x, \xi^j) + \lambda C^j \right)$$

Control Variates

The CV estimator (for a given λ):

- Unbiased: $\mathbb{E}\left[\hat{g}_{N,\mathrm{CV}}(x)\right] = g(x)$

Control Variates

The CV estimator (for a given λ):

- Unbiased: $\mathbb{E}\left[\hat{g}_{N,\mathrm{CV}}(x)\right] = g(x)$
- Has variance:

$$\mathsf{Var}\left[\hat{g}_{N,\mathrm{CV}}(x)\right] = \frac{1}{N}\left(\sigma^2(x) + \lambda^2 \mathsf{Var}[C] + 2\lambda \mathrm{Cov}[G(x,\xi), C]\right)$$

where $\sigma^2(x) = \mathsf{Var}\left[G(x,\xi)\right]$

Control Variates

The CV estimator (for a given λ):

- Unbiased: $\mathbb{E}\left[\hat{g}_{N,\mathrm{CV}}(x)\right] = g(x)$
- Has variance:

  $$\mathrm{Var}\left[\hat{g}_{N,\mathrm{CV}}(x)\right] = \frac{1}{N}\left(\sigma^2(x) + \lambda^2 \mathrm{Var}[C] + 2\lambda \mathrm{Cov}[G(x,\xi), C]\right)$$

  where $\sigma^2(x) = \mathrm{Var}\left[G(x,\xi)\right]$
- We can minimize this variance by setting λ to

  $$\lambda^* = \frac{-\mathrm{Cov}[G(x,\xi), C]}{\mathrm{Var}[C]}.$$

CV: Role of λ

Plug λ^* back in:

$$\mathrm{Var}[\hat{g}^*_{N,\mathrm{CV}}(x)] = \frac{1}{N}\left(\sigma^2(x) - \frac{\mathrm{Cov}^2[G(x,\xi), C]}{\mathrm{Var}[C]}\right)$$

If C and $G(x,\xi)$ are correlated, the variance of the CV estimator is less than the variance of the standard MC estimator

CV: Role of λ

Plug λ^* back in:

$$\mathrm{Var}[\hat{g}^*_{N,\mathrm{CV}}(x)] = \frac{1}{N}\left(\sigma^2(x) - \frac{\mathrm{Cov}^2[G(x,\xi), C]}{\mathrm{Var}[C]}\right)$$

If C and $G(x,\xi)$ are correlated, the variance of the CV estimator is less than the variance of the standard MC estimator

CV: Role of λ

A caveat: Even though $\text{Var}[C]$ may be known, $\mathrm{Cov}[G(x,\xi),C]$ is unknown

CV: Role of λ

A caveat: Even though $\mathsf{Var}[C]$ may be known, $\mathrm{Cov}[G(x,\xi), C]$ is unknown

Can be estimated, but when an estimator of λ^* is used:

- $\hat{g}_{N,\mathrm{CV}}(x)$ No longer unbiased
- Can still yield significant variance reduction
- Obeys a Central Limit Theorem (CLT) of the form (Nelson, 1990)

$$\sqrt{N}\left(\hat{g}_{N,\mathrm{CV}}(x) - \mathbb{E}[G(x,\xi)]\right) \xrightarrow{d} \mathrm{Normal}(0, \mathsf{Var}[\hat{g}^*_{N,\mathrm{CV}}(x)])$$

CV: Role of λ

A caveat: Even though Var$[C]$ may be known, $\mathrm{Cov}[G(x,\xi), C]$ is unknown

Can be estimated, but when an estimator of λ^* is used:

- $\hat{g}_{N,\mathrm{CV}}(x)$ No longer unbiased
- Can still yield significant variance reduction
- Obeys a Central Limit Theorem (CLT) of the form (Nelson, 1990)

$$\sqrt{N}\,(\hat{g}_{N,\mathrm{CV}}(x) - \mathbb{E}[G(x,\xi)]) \xrightarrow{d} \mathrm{Normal}(0, \mathrm{Var}[\hat{g}^*_{N,\mathrm{CV}}(x)])$$

CV: Role of λ

A caveat: Even though Var$[C]$ may be known, $\mathrm{Cov}[G(x,\xi), C]$ is unknown

Can be estimated, but when an estimator of λ^* is used:

- $\hat{g}_{N,\mathrm{CV}}(x)$ No longer unbiased
- Can still yield significant variance reduction
- Obeys a Central Limit Theorem (CLT) of the form (Nelson, 1990)

$$\sqrt{N}\left(\hat{g}_{N,\mathrm{CV}}(x) - \mathbb{E}[G(x,\xi)]\right) \xrightarrow{d} \mathrm{Normal}(0, \mathrm{Var}[\hat{g}^*_{N,\mathrm{CV}}(x)])$$

CV: Role of λ

A caveat: Even though Var$[C]$ may be known, $\mathrm{Cov}[G(x,\xi), C]$ is unknown

Can be estimated, but when an estimator of λ^* is used:

- $\hat{g}_{N,\mathrm{CV}}(x)$ No longer unbiased
- Can still yield significant variance reduction
- Obeys a Central Limit Theorem (CLT) of the form (Nelson, 1990)

$$\sqrt{N}\left(\hat{g}_{N,\mathrm{CV}}(x) - \mathbb{E}[G(x,\xi)]\right) \xrightarrow{d} \mathrm{Normal}(0, \mathrm{Var}[\hat{g}^*_{N,\mathrm{CV}}(x)])$$

CV: Use in Optimization

Use in Optimization: Only for a **fixed** $x \in X$ to estimate $\mathbb{E}\left[G(x,\xi)\right]$ or its subgradients, etc.

CV: Use in Optimization

Use in Optimization: Only for a **fixed** $x \in X$ to estimate $\mathbb{E}\left[G(x,\xi)\right]$ or its subgradients, etc.

Example (Pierre-Louis et al., 2011): Consider two-stage stochastic programs of the form:

$$\min_{x \in X} \; \{g(x) := c(x) + \mathbb{E}[Q(x,\xi)]\},$$

where

$$\begin{aligned} Q(x,\xi) = \quad & \min_{y \geq 0} \quad q(y) \\ & \text{s.t.} \quad\; g(y) \leq h(\xi) - T(x,\xi). \end{aligned}$$

CV: Example Use in Optimization

Assume:

(A1) $Q(x, \cdot)$ is convex on $\text{co}(\Xi)$ for all $x \in X$;

(A2) ξ has independent components, and $h(\cdot)$ and $T(x, \cdot)$ are affine on $\mathbb{R}^{d_\xi}$ for all $x \in X$.

CV: Example Use in Optimization

Assume:

(A1) $Q(x, \cdot)$ is convex on $\text{co}(\Xi)$ for all $x \in X$;

(A2) ξ has independent components, and $h(\cdot)$ and $T(x, \cdot)$ are affine on $\mathbb{R}^{d_\xi}$ for all $x \in X$.

This class of problems could be:

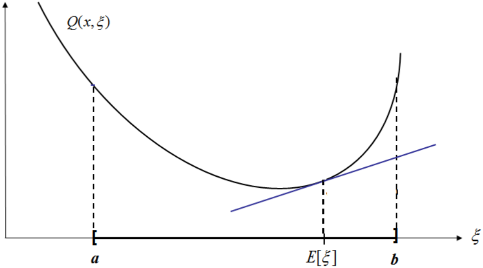
- Two-Stage Stochastic Linear Program
- Two-Stage Stochastic Convex Program
- X can have integrality restrictions, leading to a Stochastic Integer Program

CV: Example Use in Optimization

Use the first-order Taylor approximation of $Q(x,\cdot)$ as the control random variable

$$C(x,\xi) = Q(x,\bar{\xi}) + \nabla_{\xi} Q(x,\bar{\xi})(\xi - \bar{\xi})$$

where $\bar{\xi} = \mathbb{E}[\xi]$



To Sum Up. . .

Recap of Variance Reduction Techniques We Discussed

Some change the way we Sample:

- Antithetic Variates
- Latin Hypercube Sampling
- Quasi-Monte Carlo (deterministic!)
- Randomized Quasi-Monte Carlo
- Importance Sampling

Some applied typically for a fixed x and exploits problem structure:

- Importance Sampling
- Control Variates

Recap of Variance Reduction Techniques We Discussed

Some can backfire if not used properly:

- Antithetic Variates (if cannot induce negative correlation)
- Randomized Quasi-Monte Carlo (for high dimensions)
- Importance Sampling (if IS distribution not selected properly)

Can significantly improve the performance of SAA if used well (and sometimes with minimal effort)

How about properties of SAA with variance reduction?

Bias of ν_N

Bias of ν_N: Let $\xi^1, \xi^2, \ldots, \xi^N$ satisfy

$$\mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N} G(x,\xi^j)\right] = \mathbb{E}\left[G(x,\xi)\right], \ \forall x \in X$$

Then, $\mathbb{E}\left[\nu_N\right] \leq z^*$.

Unbiasedness condition above is satisfied by many variance reduction techniques we discussed: AV (with adjustments to estimator), LHS, RQMC, IS, CV (under certain conditions)

For AV, LHS, and RQMC Bias reduction have been observed.

Bias of ν_N

Bias of ν_N: Let $\xi^1, \xi^2, \ldots, \xi^N$ satisfy

$$\mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N} G(x,\xi^j)\right] = \mathbb{E}\left[G(x,\xi)\right], \ \forall x \in X$$

Then, $\mathbb{E}\left[\nu_N\right] \leq z^*$.

Unbiasedness condition above is satisfied by many variance reduction techniques we discussed: AV (with adjustments to estimator), LHS, RQMC, IS, CV (under certain conditions)

For AV, LHS, and RQMC Bias reduction have been observed.

Bias of ν_N

Bias of ν_N: Let $\xi^1, \xi^2, \ldots, \xi^N$ satisfy

$$\mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N} G(x,\xi^j)\right] = \mathbb{E}\left[G(x,\xi)\right], \ \forall x \in X$$

Then, $\mathbb{E}\left[\nu_N\right] \leq z^*$.

Unbiasedness condition above is satisfied by many variance reduction techniques we discussed: AV (with adjustments to estimator), LHS, RQMC, IS, CV (under certain conditions)

For AV, LHS, and RQMC Bias reduction have been observed.

Strong Consistency of ν_N and $\hat{x}_N$

Strong Consistency of Resulting Estimators: For many of these Variance Reduction Techniques (VRT), under appropriate conditions:

$$\nu_N \to \nu^*, \text{ wp1}$$

$$\text{dist}(S_N, S^*) \to 0, \text{ wp1}$$

Strong Consistency of ν_N and $\hat{x}_N$

Strong Consistency of Resulting Estimators: For many of these Variance Reduction Techniques (VRT), under appropriate conditions:

$\nu_N \to \nu^*$, wp1

$\text{dist}(S_N, S^*) \to 0$, wp1

- **AV:** Because AV pairs $(\xi^j, \xi^{j'})$ are iid, results typically follow from iid case with modifications to notation

Strong Consistency of ν_N and $\hat{x}_N$

Strong Consistency of Resulting Estimators: For many of these Variance Reduction Techniques (VRT), under appropriate conditions:

$\nu_N \to \nu^*$, wp1

$\mathrm{dist}(S_N, S^*) \to 0$, wp1

- **AV:** Because AV pairs $(\xi^j, \xi^{j'})$ are iid, results typically follow from iid case with modifications to notation
- **LHS:** Might require additional conditions. For instance, pointwise Strong Law of Large Numbers (SLLN) requires:

$$\mathbb{E}\left[(G(x,\xi))^2\right] < \infty$$

Conditions under which consistency results hold are discussed in, e.g., (Drew, 2007; Stockbridge, 2013).

Strong Consistency of ν_N and $\hat{x}_N$

- **RQMC:** One condition that is needed is that the star discrepancy of QMC sequence shrinks to zero

$$D^*(P_N) \searrow 0$$

  as $N \nearrow \infty$. Then, through *epi-convergence* and under additional assumptions, consistency results are shown (Koivu, 2005)

- **QMC:** Similar results through epi-convergence have been shown for QMC discretization, e.g, (Pennanen and Koivu, 2005)

Rates of Convergence

Rates of Convergence of Optimal Values: can be obtained for a class of problems that satisfy (Homem-de-Mello, 2008):

Assumption PLF

Suppose either

(i) X is convex and compact polyhedron

(ii) $G(\cdot, \xi)$ is convex and piecewise linear

(iii) ξ has finite support

or X is finite.

Rates of Convergence

Suppose assumption PLF holds and the “true” problem has a unique optimal solution x^*.

Rates of Convergence

Suppose assumption PLF holds and the "true" problem has a unique optimal solution x^*.

Suppose CLT holds pointwise for estimators $\hat{g}_N(x)$, esp. at x^*:

$$\frac{\hat{g}_N(x^*) - g(x^*)}{\sigma_N(x^*)} \xrightarrow{d} \text{Normal}(0, 1)$$

where $\sigma_N(x^*) = \text{Var}\left[\hat{g}_N(x^*)\right]$.

Rates of Convergence

Suppose assumption PLF holds and the “true” problem has a unique optimal solution x^*.

Suppose CLT holds pointwise for estimators $\hat{g}_N(x)$, esp. at x^*:

$$\frac{\hat{g}_N(x^*) - g(x^*)}{\sigma_N(x^*)} \xrightarrow{d} \text{Normal}(0, 1)$$

where $\sigma_N(x^*) = \text{Var}\left[\hat{g}_N(x^*)\right]$.

Then, optimal values also obey a CLT:

$$\frac{\nu_N - \nu^*}{\sigma_N(x^*)} \xrightarrow{d} \text{Normal}(0, 1)$$

Rates of Convergence

- **Standard MC:** We have seen earlier that the rate of convergence is $O(N^{-1/2})$

Rates of Convergence

- **Standard MC:** We have seen earlier that the rate of convergence is $O(N^{-1/2})$
- **AV:** Also by iid case above, $O(N^{-1/2})$

Rates of Convergence

- **Standard MC:** We have seen earlier that the rate of convergence is $O(N^{-1/2})$
- **AV:** Also by iid case above, $O(N^{-1/2})$
- **LHS:** Pointwise CLT for LHS holds when G is bounded

  $$\sup_{x \in X, \xi \in \Xi} |G(x, \xi)| < M \quad \text{for some } 0 < M < \infty$$

  and $G(x^*, \cdot)$ is not additive.

  Then, rate of convergence is same as standard MC: $O(N^{-1/2})$

Rates of Convergence

- **RQMC:** For a RQMC where pointwise CLT holds, rate of convergence is

$$O\left(\left[\frac{(\log N)^{d_\xi - 1}}{N^3}\right]^{\frac{1}{2}}\right)$$

  This rate is asymptotically better than standard MC.

Rates of Convergence

Rates of Convergence of Optimal Solutions: have similar properties (Homem-de-Mello, 2008).

For simplicity, let's assume assumption PLF holds.

Rates of Convergence

Rates of Convergence of Optimal Solutions: have similar properties (Homem-de-Mello, 2008).

For simplicity, let's assume assumption PLF holds.

If exponential rate of convergence holds pointwise

$$P(|\hat{g}_N(x) - g(x)| \geq \delta) \leq C_x e^{-N\gamma_x(\delta)}, \quad \forall x \in X$$

for all $N \geq 1$ and $\delta > 0$ with some constant $C_x > 0$ and function $\gamma_x(\cdot)$ such that $\gamma_x(0) = 0$ and $\gamma_x(z) > 0$ if $z > 0$

Rates of Convergence

Rates of Convergence of Optimal Solutions: have similar properties (Homem-de-Mello, 2008).

For simplicity, let's assume assumption PLF holds.

If exponential rate of convergence holds pointwise

$$P(|\hat{g}_N(x) - g(x)| \geq \delta) \leq C_x e^{-N\gamma_x(\delta)}, \quad \forall x \in X$$

for all $N \geq 1$ and $\delta > 0$ with some constant $C_x > 0$ and function $\gamma_x(\cdot)$ such that $\gamma_x(0) = 0$ and $\gamma_x(z) > 0$ if $z > 0$

Rates of Convergence

Then,

$$P(\hat{x}_N \notin S^*) \leq Ke^{-\alpha N} \quad \text{for all } N \geq 1$$

for some constants $K > 0$ and $\alpha > 0$

Rates of Convergence

Then,

$$P(\hat{x}_N \notin S^*) \leq Ke^{-\alpha N} \quad \text{for all } N \geq 1$$

for some constants $K > 0$ and $\alpha > 0$

LHS: Pointwise large deviations results (i.e., exponential rates of convergence) holds, for instance, when $G(x, \cdot)$ is monotone in each component of ξ

Some Final Remarks

Final Remarks

- Variance Reduction can be very important for optimization because it can significantly improve the statistical estimators
- Many asymptotic (and other) properties of SAA can be recovered when variance reduction techniques are used (sometimes, though, under more stringent conditions)
- If not used properly, some techniques may backfire
- Still more to do with respect to algorithmic (optimization wise) and application-based advances

Thank you

(bayraksan.1@osu.edu)

References I

Barrera, J., T. Homem-de Mello, E. Moreno, B. K. Pagnoncelli, and G. Canessa (2016). Chance-constrained problems and rare events: an importance sampling approach. *Mathematical Programming 157*(1), 153–189.

Drew, S. S. (2007). *Quasi-Monte Carlo Methods for Stochastic Programming*. Ph. D. thesis, Northwestern University.

Drew, S. S. and T. Homem-de-Mello (2006). Quasi-Monte Carlo strategies for stochastic optimization. In L. F. Perrone, F. P. Wieland, J. Liu, B. G. Lawson, D. M. Nicol, and R. M. Fujimoto (Eds.), *Proceedings of the 2006 Winter Simulation Conference*, pp. 774–782. IEEE Press.

Homem-de-Mello, T. (2008). On rates of convergence for stochastic optimization problems under non-i.i.d. sampling. *SIAM Journal on Optimization 19*(2), 524–551.

Homem-de-Mello, T. and G. Bayraksan (2014). Monte Carlo sampling-based methods for stochastic optimization. *Surveys in Operations Research and Management Science 19*(1), 56–85.

Koivu, M. (2005). Variance reduction in sample approximations of stochastic programs. *Mathematical Programming 103*(3), 463–485.

Kozmík, V. and D. P. Morton (2015). Evaluating policies in risk-averse multi-stage stochastic programming. *Math. Program. 152*(1-2), 275–300.

References II

Lemieux, C. (2009). *Monte Carlo and Quasi-Monte Carlo Sampling*. Springer Series in Statistics. New York: Springer.

Nelson, B. L. (1990). Control variate remedies. *Operations Research 38*, 359–375.

Pennanen, T. and M. Koivu (2005). Epi-convergent discretizations of stochastic programs via integration quadratures. *Numerische Mathematik 100*, 141–163.

Pierre-Louis, P., D. Morton, and G. Bayraksan (2011). A combined deterministic and sampling-based sequential bounding method for stochastic programming. In *Proceedings of the 2011 Winter Simulation Conference*, Piscataway, New Jersey, pp. 4172–4183. Institute of Electrical and Electronics Engineers, Inc.

Stockbridge, R. (2013). *Bias and Variance Reduction in Assessing Solution Quality for Stochastic Programs*. Ph. D. thesis, The University of Arizona.

Additional References

Freimer, M. B., J. T. Linderoth, and D. J. Thomas (2012). The impact of sampling methods on bias and variance in stochastic linear programs. *Computational Optimization and Applications 51*(1), 51–75.

Homem-de-Mello, T., V. L. de Matos, and E. C. Finardi (2011). Sampling strategies and stopping criteria for stochastic dual dynamic programming: a case study in long-term hydrothermal scheduling. *Energy Systems 2*, 1–31.

Koivu, M. (2005). Variance reduction in sample approximations of stochastic programs. *Mathematical Programming 103*(3), 463–485.

Lemieux, C. (2009). *Monte Carlo and Quasi-Monte Carlo Sampling*. Springer Series in Statistics. New York: Springer.

Linderoth, J. T., A. Shapiro, and S. J. Wright (2006). The empirical behavior of sampling methods for stochastic programming. *Annals of Operations Research 142*(1), 215–241.

Stockbridge, R. and G. Bayraksan (2016). Variance reduction in Monte Carlo sampling-based optimality gap estimators for two-stage stochastic linear programming. *Computational Optimization and Applications 64*(2), 407–431.