Lecture 4. Variance Reduction

Tiejun Li[1,2]

[1]School of Mathematical Sciences (SMS),
&
[2]Center for Machine Learning Research (CMLR),
Peking University,
Beijing 100871,
P.R. China
tieli@pku.edu.cn

Office: No. 1 Science Building, Room 1376E

Table of Contents

Necessity of Variance Reduction

Importance Sampling

Control Variates

Rao-Blackwellization

Antithetic Variables

Stratified Sampling

Necessity

- The standard MC for computing $I(f) = \int_0^1 f(x)dx$ is

$$I_N(f) = \frac{1}{N}\sum_{i=1}^{N} f(X_i), \qquad X_i \sim i.i.d.\ \mathcal{U}[0,1].$$

Necessity

- The standard MC for computing $I(f) = \int_0^1 f(x)dx$ is

$$I_N(f) = \frac{1}{N}\sum_{i=1}^{N} f(X_i), \qquad X_i \sim i.i.d.\ \mathcal{U}[0,1].$$

- The mean square error

$$\mathbb{E}|e_N|^2 = \mathbb{E}(I_N(f) - I(f))^2 = \frac{1}{N}\mathrm{Var}\,(f),$$

where

$$\mathrm{Var}(f) = \int_0^1 (f(x) - I(f))^2 dx.$$

Necessity

- The standard MC for computing $I(f) = \int_0^1 f(x)dx$ is

$$I_N(f) = \frac{1}{N}\sum_{i=1}^{N} f(X_i), \qquad X_i \sim i.i.d.\ \mathcal{U}[0,1].$$

- The mean square error

$$\mathbb{E}|e_N|^2 = \mathbb{E}(I_N(f) - I(f))^2 = \frac{1}{N}\mathrm{Var}\,(f),$$

  where

$$\mathrm{Var}(f) = \int_0^1 (f(x) - I(f))^2 dx.$$

- If $\mathrm{Var}(f) \gg 1$, the accuracy will be very poor!

Necessity

- The standard MC for computing $I(f) = \int_0^1 f(x)dx$ is

$$I_N(f) = \frac{1}{N}\sum_{i=1}^{N} f(X_i), \qquad X_i \sim i.i.d.\ \mathcal{U}[0,1].$$

- The mean square error

$$\mathbb{E}|e_N|^2 = \mathbb{E}(I_N(f) - I(f))^2 = \frac{1}{N}\mathrm{Var}\,(f),$$

where

$$\mathrm{Var}(f) = \int_0^1 (f(x) - I(f))^2 dx.$$

- If $\mathrm{Var}(f) \gg 1$, the accuracy will be very poor!
- Consider the example $f(x) = 1/\varepsilon \mathbf{1}_{[0,\varepsilon]}(x)$ for $x \in [0,1]$. We have $\mathbb{E}f(X) = 1$, while $\mathrm{Var}(f) \sim 1/\varepsilon$, where $X \sim \mathcal{U}[0,1]$.

Variance reduction

- We see from $\frac{1}{N}\mathrm{Var}\left(f\right)$ that there are two factors that affect the error of Monte Carlo method: the sampling size N and the variance of f. N is clearly limited by the computational cost we are willing to afford. But the variance can be manipulated in order to reduce the size of the error.

Variance reduction

- We see from $\frac{1}{N}\mathrm{Var}\,(f)$ that there are two factors that affect the error of Monte Carlo method: the sampling size N and the variance of f. N is clearly limited by the computational cost we are willing to afford. But the variance can be manipulated in order to reduce the size of the error.
- The essence of variance reduction: to utilize some prior information about the integrand and try to extract the part which can be efficiently and accurately estimated through other ways.

Table of Contents

Necessity of Variance Reduction

Importance Sampling

Control Variates

Rao-Blackwellization

Antithetic Variables

Stratified Sampling

Importance Sampling

- Consider for example the numerical evaluation of $\int_{-10}^{10} e^{-\frac{1}{2}x^2} dx$. Straightforward application of Monte Carlo method would give

  $$\int_{-10}^{10} e^{-\frac{1}{2}x^2} dx \approx \frac{20}{N} \sum_{i=1}^{N} e^{-\frac{1}{2}X_i^2},$$

  where $\{X_i\}_{i=1}^{N}$ are i.i.d. random variables that are uniformly distributed on $[-10, 10]$.

Importance Sampling

- Consider for example the numerical evaluation of $\int_{-10}^{10} e^{-\frac{1}{2}x^2} dx$. Straightforward application of Monte Carlo method would give

$$\int_{-10}^{10} e^{-\frac{1}{2}x^2} dx \approx \frac{20}{N} \sum_{i=1}^{N} e^{-\frac{1}{2}X_i^2},$$

  where $\{X_i\}_{i=1}^N$ are i.i.d. random variables that are uniformly distributed on $[-10, 10]$.

- However notice that the integrand $e^{-\frac{1}{2}x^2}$ is an extremely non-uniform function, whose value is very small (and hence will have little contribution to the integral) everywhere except a small neighborhood of $x = 0$, most of the samples will be wasted in the region where the integrand is small.

Importance Sampling

In other words, the uniform distribution ignores the importance of the integrand and thus the numerical quadrature is inefficient. The *importance sampling* embodies this idea by utilizing special distributions, which is schematically shown in Figure.

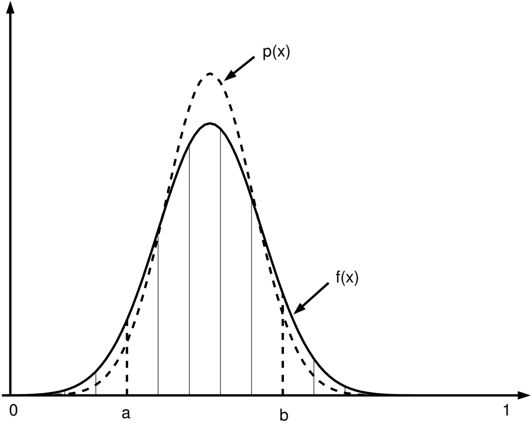

Figure: Schematics of importance sampling

Importance Sampling

Now if instead the $\{X_i\}$'s are distributed, differentially, say with density function $p(x)$, then we can use the fact that

$$\int f(x)dx = \int \frac{f(x)}{p(x)} p(x)dx = \mathbb{E}\Big(\frac{f}{p}(X)\Big)$$
$$= \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \frac{f(X_i)}{p(X_i)}$$

and approximate $\int f(x)dx$ by

$$I_N^p(f) = \frac{1}{N} \sum_{i=1}^{N} \frac{f(X_i)}{p(X_i)},$$

where $X \sim p(x)$. The error can be estimated in the same way as before, and we get

$$\mathbb{E}(I(f) - I_N^p(f))^2 = \frac{1}{N}\mathrm{Var}\Big(\frac{f}{p}\Big) = \frac{1}{N}\Big(\int \frac{f^2(x)}{p(x)}dx - I^2(f)\Big).$$

Importance Sampling

The Cauchy-Schwartz inequality shows

$$\Big(\int \frac{f(x)}{\sqrt{p(x)}}\sqrt{p(x)}dx\Big)^2 \leq \int \frac{f^2}{p}dx \int p(x)dx$$

and the equality holds iff $p(x) = cf(x)$. Now we get an ideal importance function

$$p(x) = Z^{-1}f(x)$$

if f is nonnegative, where Z is the normalization factor $Z = \int f(x)dx$. In this case

$$I(f) = I_N^p(f).$$

This is not a miracle since all the necessary work has gone into computing Z which was our original task.

Importance Sampling

- Though the perfect importance function is a mission impossible, it shows the direction toward which the sampled distribution should be constructed.

Importance Sampling

- Though the perfect importance function is a mission impossible, it shows the direction toward which the sampled distribution should be constructed.
- For the example discussed earlier, we can pick $p(x)$ that behaves as $e^{-\frac{1}{2}x^2}$ and at the same time can be sampled with a reasonable cost.

Importance Sampling

Now let us discuss a slight variant of above direct implementation of the importance sampling (Book: Monte Carlo Strategies in Scientific Computing). Suppose we are interested in evaluating

$$I = \int f(x)\pi(x)dx,$$

we can proceed as the following steps.

- Draw $X_1, \ldots, X_n$ i.i.d. from a distribution $g(x)$.

Importance Sampling

Now let us discuss a slight variant of above direct implementation of the importance sampling (Book: Monte Carlo Strategies in Scientific Computing). Suppose we are interested in evaluating

$$I = \int f(x)\pi(x)dx,$$

we can proceed as the following steps.

- Draw $X_1, \ldots, X_n$ i.i.d. from a distribution $g(x)$.
- Calculate the importance weight

$$w_j = \frac{\pi(X_j)}{g(X_j)}, \quad \text{for } j = 1, 2, \ldots, n.$$

Importance Sampling

Now let us discuss a slight variant of above direct implementation of the importance sampling (Book: Monte Carlo Strategies in Scientific Computing). Suppose we are interested in evaluating

$$I = \int f(x)\pi(x)dx,$$

we can proceed as the following steps.

- Draw $X_1, \ldots, X_n$ i.i.d. from a distribution $g(x)$.
- Calculate the importance weight

$$w_j = \frac{\pi(X_j)}{g(X_j)}, \quad \text{for } j = 1, 2, \ldots, n.$$

- Approximate the expectation by

$$\hat{I} = \frac{\sum_{i=1}^{n} w_i f(X_i)}{\sum_{i=1}^{n} w_i}.$$

Importance Sampling

Note that the expectation of $\hat{I}$ is not I, but we have by SLLN

$$\tilde{I} = \frac{1}{n}\sum_{i=1}^{n} w_i f(X_i) \to \mu \qquad \text{and} \qquad \frac{1}{n}\sum_{i=1}^{n} w_i \to 1$$

as $n \to \infty$. In this sense we call $\hat{I}$ a *biased estimator*. A major advantage of using $\hat{I}$ instead of $\tilde{I}$ is that in using the former, we *need only the ratio* $\pi(x)/g(x)$ *up to a multiplicative constant*, which is a usual case in statistics; whereas in the latter, the ratio needs to be known explicitly.

Importance Sampling

- Toy example for importance sampling. Suppose we want to compute

$$I = \int\int_{\mathcal{X}} f(x,y)dxdy,$$

where $\mathcal{X} = [-1,1] \times [-1,1]$ and

$$\begin{aligned} f(x,y) =& 0.5 \exp\Big(-90(x-0.5)^2 - 45(y+0.1)^4 \Big) \\ &+ \exp\Big(-45(x+0.4)^2 - 60(y-0.5)^2 \Big). \end{aligned}$$

Importance Sampling

- Toy example for importance sampling. Suppose we want to compute

$$I = \int\int_{\mathcal{X}} f(x,y)dxdy,$$

where $\mathcal{X} = [-1,1] \times [-1,1]$ and

$$\begin{aligned} f(x,y) =& 0.5\exp\Big(-90(x-0.5)^2 - 45(y+0.1)^4\Big) \\ &+ \exp\Big(-45(x+0.4)^2 - 60(y-0.5)^2\Big). \end{aligned}$$

- The integrand resembles some renormalized Gaussian mixture distribution except the power 4 appearing in the first part for y variable. So the first step is to choose a suitable "Gaussian" to approximate the first part suitably.

Importance Sampling

- Here we take the trial distribution

$$\begin{aligned} g(x,y) \propto & 0.5 \exp\Big(-90(x-0.5)^2 - 10(y+0.1)^2 \Big) \\ & + \exp\Big(-45(x+0.4)^2 - 60(y-0.5)^2 \Big). \end{aligned}$$

  The reason that we take the number 10 before the y variable is as follows.

Importance Sampling

- Here we take the trial distribution

$$\begin{aligned} g(x,y) \propto & 0.5 \exp\Big(-90(x-0.5)^2 - 10(y+0.1)^2 \Big) \\ & + \exp\Big(-45(x+0.4)^2 - 60(y-0.5)^2 \Big). \end{aligned}$$

  The reason that we take the number 10 before the y variable is as follows.

- Suppose we approximate $\exp(-10) \approx 0$, then from $45y^4 = 10$ we have the support radius for y is approximately $r = (10/45)^{1/4}$. With $kr^2 = 10$ we have $k = \sqrt{450} \gtrsim \mathcal{O}(10)$. A conservative choice may be $k = 10$.

Importance Sampling

- With the constraint $(x, y) \in \mathcal{X}$, it corresponds to a truncated mixture of Gaussian distribution

$$0.46N\left[\left(\begin{array}{c} 0.5 \\ -0.1 \end{array}\right), \left(\begin{array}{cc} \frac{1}{180} & 0 \\ 0 & \frac{1}{20} \end{array}\right)\right] + 0.54N\left[\left(\begin{array}{c} -0.4 \\ 0.5 \end{array}\right), \left(\begin{array}{cc} \frac{1}{90} & 0 \\ 0 & \frac{1}{120} \end{array}\right)\right].$$

Importance Sampling

- With the constraint $(x, y) \in \mathcal{X}$, it corresponds to a truncated mixture of Gaussian distribution

$$0.46N\left[\begin{pmatrix} 0.5 \\ -0.1 \end{pmatrix}, \begin{pmatrix} \frac{1}{180} & 0 \\ 0 & \frac{1}{20} \end{pmatrix}\right] + 0.54N\left[\begin{pmatrix} -0.4 \\ 0.5 \end{pmatrix}, \begin{pmatrix} \frac{1}{90} & 0 \\ 0 & \frac{1}{120} \end{pmatrix}\right].$$

- We can sample $\boldsymbol{X}_n$ from this Gaussian mixture and compute the importance weight as

$$w_i = \frac{f(\boldsymbol{X}_i)}{g(\boldsymbol{X}_i)} \cdot \mathbf{1}_{\mathcal{X}}(\boldsymbol{X}_i).$$

Importance Sampling

- One particular interesting specification of the importance sampling is the cross-entropy method. Suppose we want to compute
$$I(f) = \int f(x)\pi(x)dx$$
We assume $f \geq 0$. Then the perfect importance function will be $\mu(x) \propto f(x)\pi(x)$ but unable to sample in general.

Importance Sampling

- One particular interesting specification of the importance sampling is the cross-entropy method. Suppose we want to compute
$$I(f) = \int f(x)\pi(x)dx$$
We assume $f \geq 0$. Then the perfect importance function will be $\mu(x) \propto f(x)\pi(x)$ but unable to sample in general.
- To compute a good sample average, one can assume a parameterized pdf with the form $\mu_u(x) = \mu(x;u)$ where u are the prescribed parameters. We choose u to minimize the cross-entropy (or Kullback-Leibler "distance", or relative entropy)
$$\min_u D(\mu||\mu_u) = \int \mu(x) \ln \frac{\mu(x)}{\mu_u(x)} dx.$$
Note the order matters here and it is important for the following derivations.

Importance Sampling

- We have

$$\begin{aligned} D(\mu||\mu_u) &= \int \mu(x)\ln\mu(x)dx - \int \mu(x)\ln\mu_u(x)dx \\ &= \int \mu(x)\ln\mu(x)dx - \frac{1}{I(f)}\int f(x)\pi(x)\ln\mu_u(x)dx \end{aligned}$$

  So minimizing cross-entropy is equivalent to maximize $F(x) = \int f(x)\pi(x)\ln\mu_u(x)dx$. The extremal point satisfies

$$\nabla F(x) = \int \frac{f(x)\pi(x)}{\mu_u(x)}\nabla_u\mu(x;u)dx = 0.$$

  Solving this equation we obtain u^*, thus have a good candidate importance distribution μ_{u^*}.

Importance Sampling

The above argument is very useful for estimating the *rare events* such as the the small probability $p = \mathbb{P}(X \geq \gamma) = \mathbb{E}1_{\{X \geq \gamma\}}$. We have the relative error

$$\frac{\sqrt{\mathrm{Var}(1_{\{X \geq \gamma\}})}}{I} = \sqrt{\frac{1-p}{p}} \gg 1 \quad \text{when} \quad p \ll 1.$$

One should introduce a multileveled version of the cross-entropy method to relax this issue with a step-by-step version.

Table of Contents

Necessity of Variance Reduction

Importance Sampling

Control Variates

Rao-Blackwellization

Antithetic Variables

Stratified Sampling

Control Variates

Consider another form of $I(f)$

$$I(f) = \int f(\boldsymbol{x})d\boldsymbol{x} = \int (f(\boldsymbol{x}) - g(\boldsymbol{x}))d\boldsymbol{x} + \int g(\boldsymbol{x})d\boldsymbol{x}.$$

The idea of *control variates* is quite simple.

- If $g(\boldsymbol{x})$ is very similar to $f(\boldsymbol{x})$, and $I(g)$ is known or can be obtained in a highly accurate manner, then $\mathrm{Var}(f - g) < \mathrm{Var}(f)$, we will obtain a variance reduced estimator of $I(f)$.

Control Variates

Consider another form of $I(f)$

$$I(f) = \int f(\boldsymbol{x})d\boldsymbol{x} = \int (f(\boldsymbol{x}) - g(\boldsymbol{x}))d\boldsymbol{x} + \int g(\boldsymbol{x})d\boldsymbol{x}.$$

The idea of *control variates* is quite simple.

- If $g(\boldsymbol{x})$ is very similar to $f(\boldsymbol{x})$, and $I(g)$ is known or can be obtained in a highly accurate manner, then $\mathrm{Var}(f - g) < \mathrm{Var}(f)$, we will obtain a variance reduced estimator of $I(f)$.
- Similarly, an ideal control variates will be f itself, but we don't know $I(f)$! This is similar to the importance sampling. Though the perfect control variates is not practical, it tells us the direction toward which the approximate control variates should be constructed.

Control Variates

Another form of control variates is as follows.

- Suppose we have an unbiased estimator

$$U = \frac{1}{N} \sum_{i=1}^{N} f(X_i)$$

  for the integral $I(f)$, and we have another statistic V with known expectation $\mathbb{E}V = \mu$.

Control Variates

Another form of control variates is as follows.

- Suppose we have an unbiased estimator

$$U = \frac{1}{N}\sum_{i=1}^{N} f(X_i)$$

for the integral $I(f)$, and we have another statistic V with known expectation $\mathbb{E}V = \mu$.

- Define a new static

$$\tilde{U} = U + c(V - \mu)$$

where c is to be determined. It is obvious that $\tilde{U}$ is also an unbiased estimator of $I(f)$. We have

$$\mathrm{Var}(\tilde{U}) = \mathrm{Var}(U) + c^2\mathrm{Var}(V) + 2c\mathrm{Cov}(U, V).$$

Control Variates

- The optimal parameter for the minimization of the variance is

$$c^* = -\mathrm{Cov}(U, V)/\mathrm{Var}(V).$$

In this case

$$\mathrm{Var}(\tilde{U}^*) = (1 - \rho^2_{U,V})\mathrm{Var}(U)$$

where $\rho_{U,V}$ is the correlation coefficient between U and V. So the more the introduced estimator V correlates with U, the more accurate the result will be. The constant c^* is usually computed from simulations in practice, e.g.

$$C^*_N = -\frac{\sum_{i=1}^{N}(U_i - \bar{U})(V_i - \bar{V})}{\sum_{i=1}^{N}(V_i - \bar{V})^2}.$$

Control Variates

There are also nonlinear version of control variates like

$$\bar{X} \cdot \frac{\mathbb{E}Y}{\bar{Y}} \quad \text{or} \quad \bar{X} \exp(\bar{Y} - \mathbb{E}Y)$$

in estimating $\mathbb{E}X$ through $\bar{X}$.

Example (Toy example for control variates)

Consider the following integral

$$I(f) = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} (1+r)^{-1} e^{-\frac{x^2}{2}} dx,$$

where $r = e^{\sigma x}, \ \sigma \gg 1$.

Control Variates

Notice that

$$(1+r)^{-1} \approx h(x) = \begin{cases} 1, & x \leq 0, \\ 0, & x > 0, \end{cases}$$

we have

$$I(f) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \big((1+r)^{-1} - h(x)\big) e^{-\frac{x^2}{2}} dx + \frac{1}{2}.$$

Here $h(x)$ plays the role of control variates. Applying standard normal distribution can reduce the variance more.

Table of Contents

Necessity of Variance Reduction

Importance Sampling

Control Variates

Rao-Blackwellization

Antithetic Variables

Stratified Sampling

Rao-Blackwellization

- This method reflects a basic principle in Monte Carlo computation: *One should carry out analytical computation as much as possible.* Indeed this principle is also embodied in the idea of control variates.

Rao-Blackwellization

- This method reflects a basic principle in Monte Carlo computation: *One should carry out analytical computation as much as possible.* Indeed this principle is also embodied in the idea of control variates.
- Suppose we have n independent samples $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ drawn from pdf $\pi(\boldsymbol{x})$ and we are interested in evaluating $I = \int f(\boldsymbol{x})\pi(\boldsymbol{x})d\boldsymbol{x}$. A straightforward estimator is

$$\hat{I} = \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{X}_i).$$

Rao-Blackwellization

- This method reflects a basic principle in Monte Carlo computation: *One should carry out analytical computation as much as possible.* Indeed this principle is also embodied in the idea of control variates.
- Suppose we have n independent samples $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ drawn from pdf $\pi(\boldsymbol{x})$ and we are interested in evaluating $I = \int f(\boldsymbol{x})\pi(\boldsymbol{x})d\boldsymbol{x}$. A straightforward estimator is

$$\hat{I} = \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{X}_i).$$

- Suppose that $\boldsymbol{x}$ can be decomposed into two parts $(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)})$ and the conditional expectation $\mathbb{E}(f(\boldsymbol{X})|\boldsymbol{x}^{(2)})$ can be obtained analytically or in a highly accurate manner. We can define another unbiased estimator of I as

$$\tilde{I} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(f(\boldsymbol{X})|\boldsymbol{X}_i^{(2)}).$$

Rao-Blackwellization

If the computational effort for obtaining the two estimates are similar, then $\tilde{I}$ should be preferred because of the variance identity (Book:Probability: Theory and Examples)

$$\mathrm{Var}(f(\boldsymbol{X})) = \mathrm{Var}(\mathbb{E}(f(\boldsymbol{X})|\boldsymbol{X}^{(2)})) + \mathbb{E}(\mathrm{Var}(f(\boldsymbol{X})|\boldsymbol{X}^{(2)})),$$

which implies that

$$\mathrm{Var}(\hat{I}) = \frac{\mathrm{Var}(f(\boldsymbol{X}))}{m} \geq \frac{\mathrm{Var}(\mathbb{E}(f(\boldsymbol{X})|\boldsymbol{X}^{(2)}))}{m} = \mathrm{Var}(\tilde{I}).$$

The above procedure is called Rao-Blackwellization. The readers may be referred to (Book: Monte Carlo Strategies in Scientific Computing) for more details.

Table of Contents

Necessity of Variance Reduction

Importance Sampling

Control Variates

Rao-Blackwellization

Antithetic Variables

Stratified Sampling

Antithetic Variables

Proposition

Suppose $X \sim \mathcal{U}[0,1]$, and $f(x)$ is monotone, then

$$\mathrm{Cov}(f(X), f(1-X)) \le 0.$$

Define

$$I_N(f) = \frac{1}{2N}\sum_{i=1}^{N}(f(X_i) + f(1-X_i)), \qquad X_i \sim i.i.d.\, \mathcal{U}[0,1],$$

then $\mathbb{E}I_N = I(f)$,

$$\mathrm{Var}(I_N) = \frac{1}{2N}(\mathrm{Var}(f) + \mathrm{Cov}(f(X), f(1-X))) \le \frac{1}{2N}\mathrm{Var}(f).$$

The variance is reduced!

Table of Contents

Necessity of Variance Reduction

Importance Sampling

Control Variates

Rao-Blackwellization

Antithetic Variables

Stratified Sampling

Stratified Sampling

Consider two numerical integration strategies:

1. Monte Carlo: $I_N^{(1)}(f) = \frac{1}{N} \sum_{i=1}^{N} f(X_i), \qquad X_i \sim i.i.d.\, \mathcal{U}[0,1]$.

Stratified Sampling

Consider two numerical integration strategies:

1. Monte Carlo: $I_N^{(1)}(f) = \frac{1}{N}\sum_{i=1}^{N} f(X_i), \qquad X_i \sim i.i.d.\, \mathcal{U}[0,1].$
2. Midpoint rule: $I_N^{(2)}(f) = \frac{1}{N}\sum_{i=1}^{N} f(Y_i), \qquad Y_i = \frac{1}{2N} + \frac{i-1}{N}.$

Stratified Sampling

Consider two numerical integration strategies:

1. Monte Carlo: $I_N^{(1)}(f) = \frac{1}{N}\sum_{i=1}^N f(X_i), \qquad X_i \sim i.i.d.\, \mathcal{U}[0,1].$
2. Midpoint rule: $I_N^{(2)}(f) = \frac{1}{N}\sum_{i=1}^N f(Y_i), \qquad Y_i = \frac{1}{2N} + \frac{i-1}{N}.$

- The error estimate:

$$|e_N^{(1)}| \sim \mathcal{O}(\frac{1}{\sqrt{N}}), \quad |e_N^{(2)}| \sim \mathcal{O}(\frac{1}{N^2}).$$

Stratified Sampling

Consider two numerical integration strategies:

1. Monte Carlo: $I_N^{(1)}(f) = \frac{1}{N}\sum_{i=1}^N f(X_i), \qquad X_i \sim i.i.d.\, \mathcal{U}[0,1]$.
2. Midpoint rule: $I_N^{(2)}(f) = \frac{1}{N}\sum_{i=1}^N f(Y_i), \qquad Y_i = \frac{1}{2N} + \frac{i-1}{N}$.

- The error estimate:

$$|e_N^{(1)}| \sim \mathcal{O}(\frac{1}{\sqrt{N}}), \quad |e_N^{(2)}| \sim \mathcal{O}(\frac{1}{N^2}).$$

- The comparison of accuracy:

$$\text{Uniform} > \text{Quasi-random} > \text{Random}.$$

To improve the accuracy, one applies

$$\text{Uniform} + \text{Adaptive} \to \text{Moving Mesh}.$$

$$\text{Random} + \text{Adaptive} \to \text{Importance Sampling}.$$

Stratified Sampling

- Idea: If we combine the uniform and random sample points, we obtain the stratified sampling, and the accuracy will be improved.

Stratified Sampling

- Idea: If we combine the uniform and random sample points, we obtain the stratified sampling, and the accuracy will be improved.
- Strategy: Divide $\Omega = [0,1]$ into M equi-partitions

$$\Omega_k = \Big[\frac{k-1}{M}, \frac{k}{M}\Big], \qquad k = 1, 2, \ldots, M.$$

Sample $N_k = N/M$ points uniformly in Ω_k, denoted as $X_i^{(k)}, \ i = 1, \ldots, N_k$. Define

$$\bar{f}(x) = \bar{f}_k = |\Omega_k|^{-1} \int_{\Omega_k} f(x) dx = \mathbb{E} f(X^{(k)}), \quad x \in \Omega_k$$

and

$$I_N = \frac{1}{N} \sum_{k=1}^{M} \sum_{i=1}^{N_k} f(X_i^{(k)}).$$

Stratified Sampling

We have

$$\begin{aligned}
\mathbb{E}I_N &= \frac{1}{N}\sum_{k=1}^{M}(N_k\cdot \bar{f}_k) = I(f),\\
\mathrm{Var}\,(I_N) &= \frac{1}{N^2}\sum_{i,k}\sum_{j,l}\mathbb{E}\Big[\big(f(X_i^{(k)})-\bar{f}_k\big)\big(f(X_j^{(l)})-\bar{f}_l\big)\Big]\\
&= \frac{1}{N^2}\sum_{k=1}^{M}\left(N_k\cdot|\Omega_k|^{-1}\int_{\Omega_k}(f(x)-\bar{f}_k)^2dx\right)\\
&= \frac{1}{N}\int_{\Omega}(f(x)-\bar{f}(x))^2dx.
\end{aligned}$$

Stratified Sampling

Proposition

Define $\sigma_s = \left(\int_\Omega (f(x) - \bar{f}(x))^2 dx\right)^{\frac{1}{2}}$, *then*

$$\sigma_s \le \sigma = \left(\int_\Omega (f(x) - I(f))^2 dx\right)^{\frac{1}{2}}.$$

Stratified Sampling

Proposition

Define $\sigma_s = \left(\int_\Omega (f(x) - \bar{f}(x))^2 dx\right)^{\frac{1}{2}}$, *then*

$$\sigma_s \leq \sigma = \left(\int_\Omega (f(x) - I(f))^2 dx\right)^{\frac{1}{2}}.$$

Proof.

The quadratic function of c, $g(c) = \int_{\Omega_k} (f(x) - c)^2 dx$ takes minimum at $c = \bar{f}_k$, so we have

$$\sigma_s^2 = \sum_k \int_{\Omega_k} (f(x) - \bar{f}_k)^2 dx \leq \sum_k \int_{\Omega_k} (f(x) - I(f))^2 dx = \sigma^2.$$

The variance is reduced! □

Stratified Sampling

The stratified sampling can be combined with importance sampling.
Let

$$\Omega = \bigcup_{k=1}^{M} \Omega_k,$$

take N_k points $\{X_i^{(k)}\}_{i=1}^{N_k}$ in Ω_k, $\sum_{k=1}^{M} N_k = N$. Assume $\{X_i^{(k)}\}_{i=1}^{N_k} \sim i.i.d.\ p^{(k)}(x) = p(x)/\bar{p}_k$, $x \in \Omega_k$, and $\bar{p}_k = \int_{\Omega_k} p(x)dx$, then

$$I_N = \sum_{k=1}^{M} \frac{\bar{p}_k}{N_k} \sum_{i=1}^{N_k} f(X_i^{(k)}).$$

Stratified Sampling

Define

$$\bar{f}(x) = \bar{f}_k = \mathbb{E} f(X^{(k)}) = \bar{p}_k^{-1} \int_{\Omega_k} f(x) p(x) dx, \quad x \in \Omega_k,$$

we have

$$\mathbb{E} I_N = \sum_{k=1}^{M} \int_{\Omega_k} f(x) p(x) dx = I(f),$$

$$\mathrm{Var}\,(I_N) = \sum_{k=1}^{M} \frac{\bar{p}_k}{N_k} \int_{\Omega_k} (f(x) - \bar{f}_k)^2 p(x) dx = \sum_{k=1}^{M} \frac{\bar{p}_k}{N_k} \sigma_k^2,$$

where $\sigma_k^2 \triangleq \int_{\Omega_k} (f(x) - \bar{f}_k)^2 p(x) dx$.

Stratified Sampling

Proposition

If the balance condition $\bar{p}_k/N_k = \frac{1}{N}$ *is satisfied, the variance is reduced.*

Stratified Sampling

Proposition

If the balance condition $\bar{p}_k/N_k = \frac{1}{N}$ is satisfied, the variance is reduced.

- In a nutshell, the stratified sampling can be described as

$$\begin{aligned}\mathbb{E}Y &= \sum_{k=1}^{K} \mathbb{P}(Y \in A_k)\mathbb{E}(Y|Y \in A_k) = \sum_{k=1}^{K} p_k \mathbb{E}(Y|Y \in A_k) \\ &= \sum_{k=1}^{K} \frac{n_k}{n}\mathbb{E}(Y|Y \in A_k) \approx \frac{1}{n}\sum_{k=1}^{K}\sum_{j=1}^{n_k} Y_{kj}\end{aligned}$$

  where $Y_{kj} \sim Y|Y \in A_k$, and $n_k = np_k$ which is enforced in the partition.

Stratified Sampling

- When the stratified sampling is applied to the realistic high dimensional problems, rather than attempt to stratify all the dimensions, it is better to identify which variables (if any) carry most of the variation of the integrand and stratify these.

Stratified Sampling

- When the stratified sampling is applied to the realistic high dimensional problems, rather than attempt to stratify all the dimensions, it is better to identify which variables (if any) carry most of the variation of the integrand and stratify these.
- Significant reduction in the variance can sometimes be achieved by stratifying a single dimension in a many-dimensional integral.