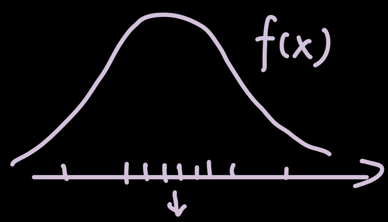
Monte Carlo Integration

Statistical paradigm

population ⟶ random sample

$N (\approx \infty)$ points $\quad X_1, \dots X_n \overset{iid}{\sim} f(x)$



population average $\xleftarrow{\text{estimation}}$ sample average

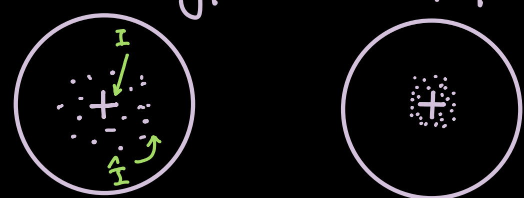
- population average $\quad$ approx. $\quad$ • sample average

$I = E(h(X)) = \int h(x) f(x)\, dx \qquad \hat{I} = \frac{1}{n} \sum_{i=1}^{n} h(X_i)$

or $\left[\frac{1}{N} \sum_{i=1}^{N} h(X_i) \right]$ for a very large N.

$\boxed{\text{fixed}} \qquad \boxed{\text{random}}$

under hypothetical repetitions



one method $\qquad$ another method (better)

$E(\hat{I}) = \frac{1}{n} \sum_{i=1}^{n} \underbrace{E(h(X_i))}_{\text{each one is}: I}$

$= \frac{1}{n} \cdot n I$

$= I$

$\text{var}(\hat{I}) = \frac{1}{n^2} \text{Var}\left(\sum_{i=1}^{n} h(X_i)\right)$

unbiased

$= \frac{1}{n^2} \sum_{i=1}^{n} \text{Var}(h(X_i))$ ← since all X_i are indep.

$= \frac{1}{n^2}\, n\, \text{var}(h(X))$

$= \frac{\text{Var}(h(X))}{n}$ → we can estimate population variance w/ sample variance

increasing sample size decreases variance

$\sigma^2 = \mathrm{var}(h(X))$

$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(h(X_i) - \hat{I})^2$ average over squared deviation from center

$\mathrm{var}(\hat{I}) \approx \frac{\hat{\sigma}^2}{n}$ based on hypothetical repeated sampling

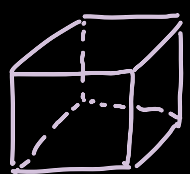
Buffon needle experiment

1901, Mario lazzarini performed the experiment $n = 3408$ times, and the Monte Carlo approximation to π : $\hat{\pi} = \frac{353}{113} = 3.14159\ldots$

too good to be true

$\vec{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} \sim \mathrm{Unif}[0,1]^d$ d-dimensional cube

density $f(\vec{x}) = 1$ for x in the unit cube



E.g. 3d

density $= 1$. difficult to calculate.

$I = E(h(\vec{x})) = \int\int\cdots\int h(\vec{x})\,d\vec{x}$

$X_1, X_2, \ldots, X_n \overset{iid}{\sim} \mathrm{Unif}[0,1]^d$

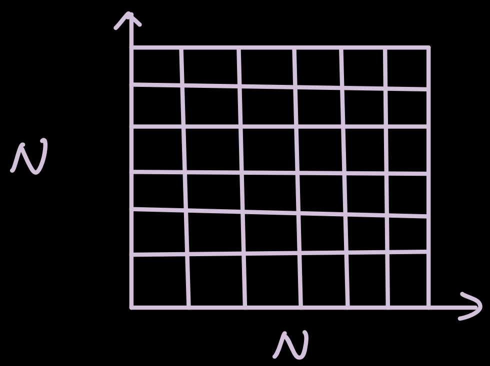
$\hat{I} = \frac{1}{n}\sum_{i=1}^{n} h(\vec{X}_i)$

has nothing to do with d

$E(\hat{I}) = I$, $\mathrm{var}(\hat{I}) = \frac{\mathrm{Var}(h(\vec{X}))}{n}$

avoids the curse of dimensionality since has nothing to do with d.

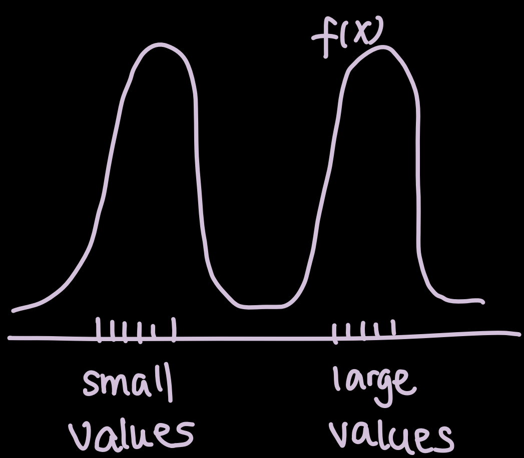
Deterministic methods are "cursed":



$d = 2$

N^d cells

Intuition about average



$X_1, X_2 \overset{iid}{\sim} f(x)$

$\bar{X} = \frac{X_1 + X_2}{2}$

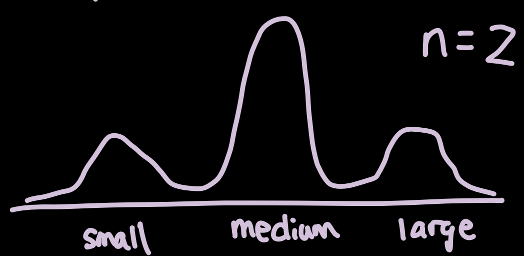
Distribution of $\bar{x}$:

$x_1 \backslash x_2$	small	large
small	small $p = \frac{1}{4}$	medium
large	medium	large $p = \frac{1}{4}$

$P(\text{med. avg}) = \frac{1}{2}$

Graph of $\bar{x}$:

$n = 2$

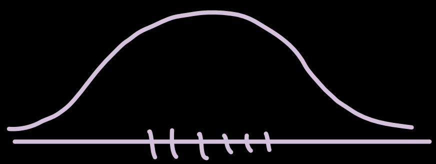


Effects of averaging

(1) variance becomes smaller

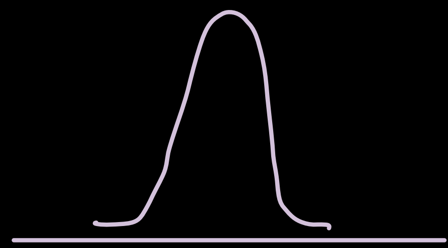
(2) Shape is smoother

$\bar{X}$, n=10



medium

$\bar{X}$, n=100



$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$

$\mu = E(X)$

$\sigma^2 = var(X)$

Central Limit Theorem

Law of large numbers.

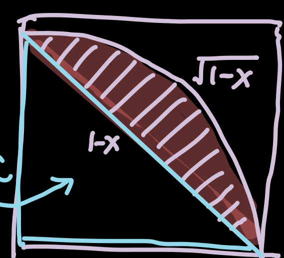
Recall,

$$Var(X+Y) = Var(X) + Var(Y) + 2cov(X, Y)$$

Control Variance method

$$\hat{I} = \frac{1}{n} \sum_{i=1}^{n} \left(\sqrt{1-X^2} - (1-X)\right) + \frac{1}{2}$$

deterministic



We do this, but we want:

$$Var\left(\sqrt{1-X_i^2} - (1-X_i)\right) < Var\left(\sqrt{1-X_i}\right)$$

If X and Y are highly correlated, and we subtract something highly correlated to the original, we can reduce the variance.

Compute shaded part using Monte Carlo.

Most General Formulation

$$I = \int a(X)\,d(X)$$

where · X can be multi-dimensional
· no density

$$= \int \boxed{\frac{a(x)}{f(x)}}^{h(x)} \underset{\sim}{f(x)}\,dx$$

bring in a density function (continuous) or probability mass function (discrete)

$$= E_f\left[\frac{a(x)}{f(x)}\right]$$

→ avg over n points

design a f(x) to do Monte-Carlo

$$X_1, X_2, \ldots, X_n \overset{iid}{\sim} f(x)$$

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

$$E(\hat{I}) = I$$

$$\mathrm{var}(\hat{I}) = \underline{\frac{1}{n}\mathrm{Var}_f\left(\frac{a(X)}{f(X)}\right)}$$

f(x) should be designed so that when a(X) is constant, variance is smallest

Importance Sampling

$$I = E_f[h(X)] = \int h(x)f(x)\,dx$$

$$\downarrow \qquad\qquad \downarrow$$

$$\frac{1}{N}\sum_{i=1}^{N} h(X_i) = \sum_{\text{bins }(x,\,x+\Delta x)} h(x)\underbrace{f(x)\Delta x}_{N(x)/N}$$

Analogy: population average — sum over large # of bins

Sometimes, $f(x)$ may be inefficient and we want to sample from $g(x)$ instead.

$$I = \int \underbrace{h(x) f(x)}_{\text{"}a(x)\text{"}} dx = \int h(x) \frac{f(x)}{g(x)} g(x)\, dx$$

$$\downarrow \qquad\qquad\qquad\qquad \downarrow$$

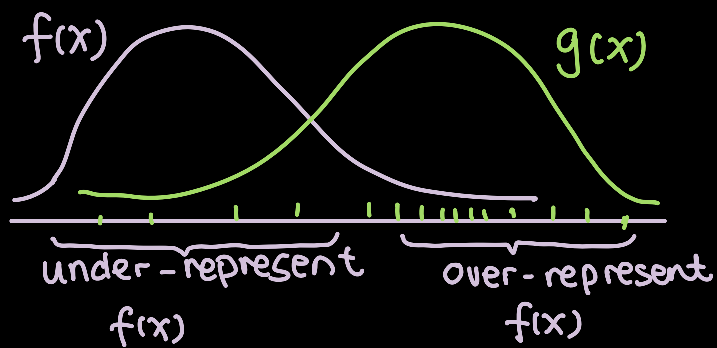
$$E_f[h(X)] = E_g\left[h(X) \boxed{\frac{f(X)}{g(X)}}\right]$$

$\uparrow$

$w(X)$, importance

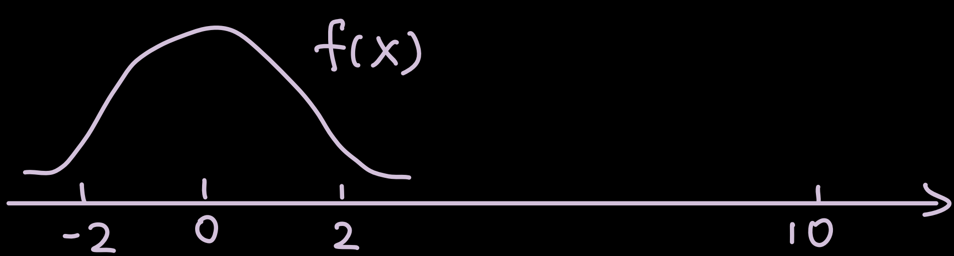
Why do we have importance ?

Suppose :



Example 1:

$X \sim N(0, 1)$



What is $P(X>10)$?

indicator function

$$P(X>10) = E(\underbrace{1(X>10)}) \qquad 1(x) = \begin{cases} 1 & \text{if } x > 10 \\ 0 & \text{if } x \leq 10 \end{cases}$$

$$= \int 1(x>10) f(x)\, dx$$

$$= \int_{10}^{\infty} f(x)\, d(x)$$

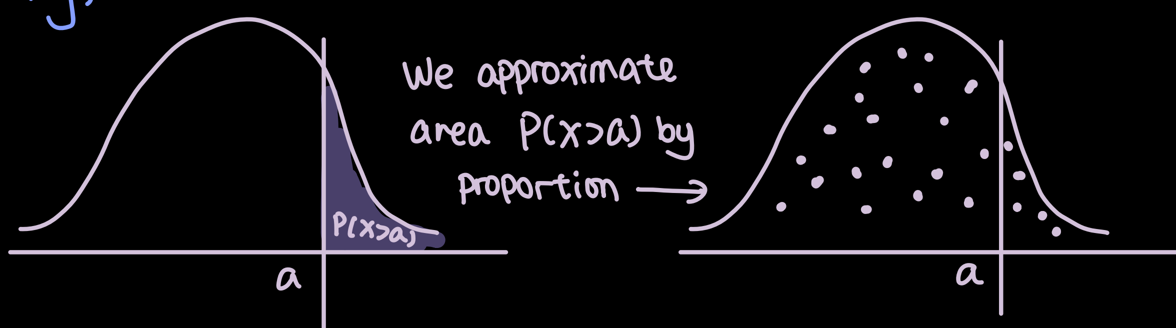
How do we calculate population average ?

$$X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(0,1)$$

$$\hat{I} = \frac{1}{n} \sum_{i=1}^{n} 1(X_i > 10)$$

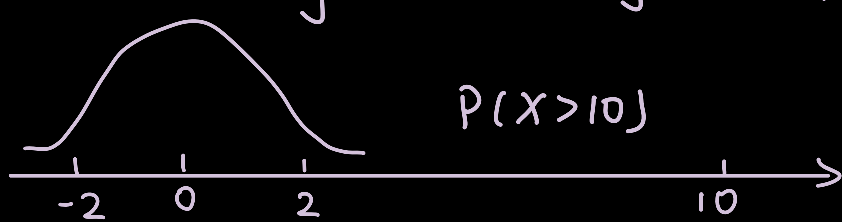
= count the proportion of the occurrence $X_i > 10$ over the total # of events

Normally,

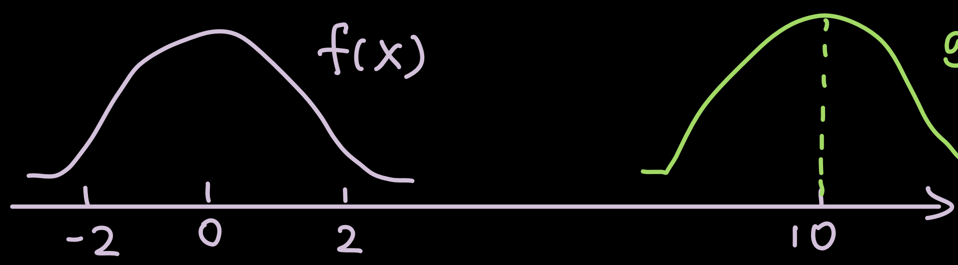


But for $x = 10$ ...

Like in the insurance company's scenario, Capturing that event may lead to very small probability.



What if we sampled from a different distribution, like centering at 10:



$X_1, X_2, \dots X_n \overset{iid}{\sim} N(\underset{\sim}{10}, 1)$ g(x)

$\hat{I} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(X_i > 10)$ would give a Probability of 50% ... incorrect.

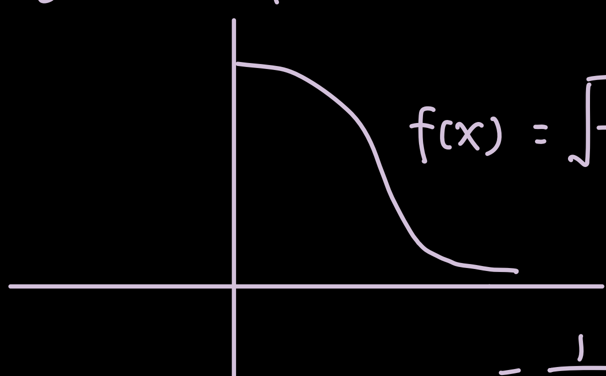
With importance sampling,

$\hat{I} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(X_i > 10)\, w_i$, where

$$w_i = \frac{f(X_i)}{g(X_i)} = \frac{\frac{1}{\sqrt{2\pi}} e^{-\frac{x_i^2}{2}}}{\frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i-10)^2}{2}}}$$

Example 2. Normalizing Constant Estimation

$X \sim N_+(0,1)$



$$f(x) = \sqrt{\frac{2}{\pi}}\, e^{-\frac{x^2}{2}}$$

$$= \frac{1}{Z} e^{-\frac{x^2}{2}}, \text{ where } \underset{\text{Normalizing constant}}{Z} = \int_0^\infty e^{-\frac{x^2}{2}}\, dx$$

$$Z = \int_0^{\infty} e^{-\frac{x^2}{2}} dx = \int_0^{\infty} \tilde{f}(x)\, dx$$

$$g(x) = e^{-x} \quad (x \geq 0)$$

$$X \sim g(x)$$

$X = -\log U$ generated x using inversion method

Using importance sampling :

$$Z = \int_0^{\infty} \tilde{f}(x)\, dx$$

$$= \int_0^{\infty} \tilde{f}(x)\, g(x)\, dx$$

$$= E_g\left[\frac{\tilde{f}(x)}{g(x)}\right]$$

$$X_1, X_2, \ldots, X_n \overset{iid}{\sim} g(x)$$

$$\hat{Z} = \frac{1}{n} \sum_{i=1}^{n} \frac{\tilde{f}(x_i)}{g(x_i)}$$

$$= \frac{1}{n} \sum_{i=1}^{n} e^{-\frac{x_i^2}{2} + x_i}$$

$$E_f[h(x)] = \int h(x)\, f(x)\, dx$$

$$= \int h(x)\, \frac{1}{Z}\, \tilde{f}(x)\, dx$$

for instance $\stackrel{e.g.}{=}$ $\int_0^\infty \sqrt{x}\,\frac{1}{z}\,e^{-\frac{x^2}{2}}\,dx$

If we only know $\hat{f}(x)$ but we don't know z, how can we calculate E_f?

$$= \frac{1}{z}\int_0^\infty \sqrt{x}\,e^{-\frac{x^2}{2}}\,dx$$

how to estimate the integral?

$$E_f[h(x)] = \frac{1}{z}\int h(x)\hat{f}(x)\,dx$$

$$\tilde{I} = \int h(x)\hat{f}(x)\,dx = \int h(x)\frac{\hat{f}(x)}{g(x)}g(x)\,dx$$

$$\hat{\tilde{I}} = \frac{1}{n}\sum_{i=1}^{n} h(X_i)\underbrace{\frac{\hat{f}(x_i)}{g(x_i)}}_{\tilde{w}_i}$$

$$I = E_f[h(x)] = \frac{1}{z}\int h(x)\hat{f}(x)\,dx = \frac{\tilde{I}}{z}$$

So,

$$\hat{I} = \frac{\frac{1}{n}\sum_{i=1}^{n} h(X_i)\tilde{w}_i}{\frac{1}{n}\sum_{i=1}^{n}\tilde{w}_i}$$

$$= \frac{\sum_{i=1}^{n} h(X_i)\tilde{w}_i}{\sum_{i=1}^{n}\tilde{w}_i}$$