* Monte Carlo Methods:

Expected value of r.vs:

$$\hat{E}_{p_\theta}[h(X)] = \sum_{i=1}^{\infty} h(x_i)\, p_\theta(x_i) \qquad \text{(Discrete)}$$

$\hookrightarrow$ Distribution from which our r.v belongs to

$$E_{p_\theta}[h(X)] = \int_{\mathbb{R}} h(x)\, p_\theta(x)\, dx$$

Main challenge when computing E is that it's an integral from $-\infty$ to $+\infty$ & for discrete it's a sum from 1 to ∞

Solution: Law of Large Numbers: Average of the results obtained from a large number of trials should be close to the expected value & will tend to become closer to the expected value as more trials are performed.

$$\frac{1}{N} \sum_{i=1}^{N} h(x_i) \approx E_{p_\theta}[h(X)]$$

Let's explain LLN w/ an example:

I have a fair dice & I want its expected value

$$\Rightarrow E_{p_\theta}[h(X)] = \sum_{i=1}^{6} h(x_i)\, p_\theta(x_i) \quad \text{w/ } h(x_i) = x_i$$

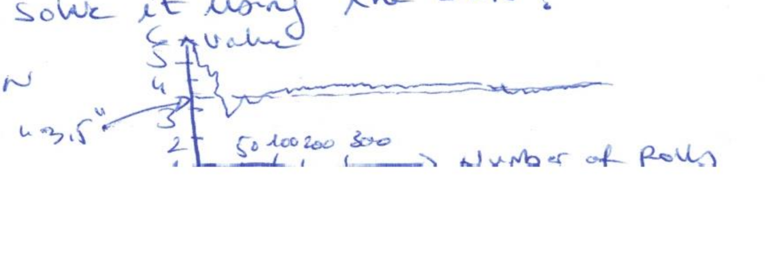
$$\text{w/ } p(x=1) = p(x=2) = p(x=3) = \dots = p(x=6) = \frac{1}{6}.$$

$$= 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + \dots + 6 \cdot \frac{1}{6}$$

$= 3{,}5$ (Theoretical value: when I roll the dice, the expected value is going to be "3,5" (Theoretical))

Can we solve it using the LLN?

LLN

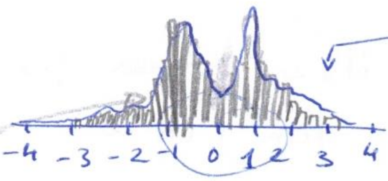


Questions WRT LLN:

1) how many number of samples are sufficient to estimate the mean
2) If I repeat the process, would I still get the same estimated value => what's the variance of the mean.
3) How do we ever sample from our desired distribution?

<u>let's explore 3)</u>

we generate rv in accordance to its distribution.



Consider this function. This function represents a distribution we want to sample from.

random # sample should be generated WRT to their plausibility.

→ More samples should come from here & less from the tails

Suppose I have an algorithm to generate samples from this distribution.
=> histogram would have looked like this (in black)
looks ~ like our target distribution.

However, I know that my computer (Algo) can ONLY GENERATE <u>UNIFORM</u> RV (equally likely) in the range I provided.
↳ ("Pseudo random numbers")

<u>Somehow these uniform random numbers are being transformed in such a way to make them appear in accordance to the desired distribution function.</u>

<u>let's see 3 such algorithms</u> (that transform uniform random numbers)

(A) Inverse CDF transformation:

Based on proba. integral transform theorem

for each continuous r.v, $\exists$ a transformation that gives a r.v w/ a $\underbrace{\text{standard}}_{\text{range } (0,1)}$ uniform distribution.

$\Rightarrow$ $Y = F_X(X)$

standard uniform distribution

CDF (cumulative distribution function)

That's the probability integral transform theorem.

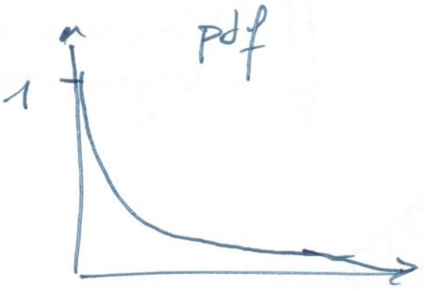
$\Rightarrow$ $X = F_X^{-1}(Y)$

$\hookrightarrow$ inverse of CDF

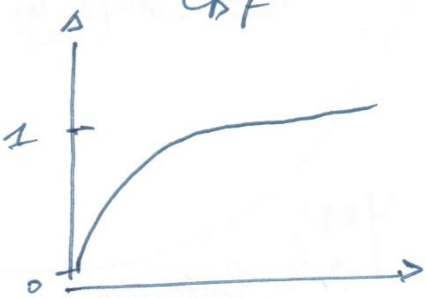
Key: I must have the analytical form of the CDF to be able to do it.

Example:

pdf



CDF



$f_X(x) = \begin{cases} 0 & x < 0 \\ e^{-x} & x \geq 0 \end{cases}$

$F_X(x) = P(X \leq x)$

$y = 1 - e^{-x}$

$\boxed{x = -\ln(1-y)}$

In this case we have the CDF & the inverse of the CDF. (PS: Some function don't have an inverse ex: ~~$\mathcal{N}$~~)

inverse of the CDF

$\Rightarrow$ We're good.

Sometimes CDF & inverse of CDF

PS: I derive y like this:

$F_X(x) = P(X \leq x) = P(-\infty \leq X \leq x)$

$= \int_{-\infty}^{x} f_X(t)\,dt = \int_{-\infty}^{x} f_X(t)\,dt$

$= \int_0^x f_X(t)\,dt = \int_0^x e^{-t}\,dt$

So the probability integral transforms theorem tells us the following:

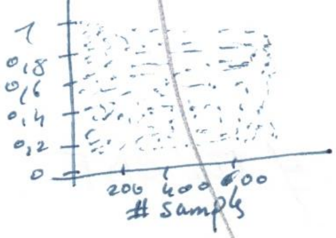
If I have a r.v w/ the pdf that has an CDF & inv(CDF) that have an analytic form:

$$F_X(x) = P(X \leq x)$$

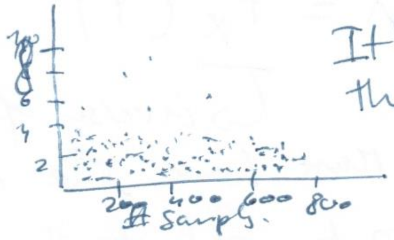
$$\text{CDF} \rightarrow y = 1 - e^{-x}$$

$$\text{CDF}^{-1} \rightarrow x = -\ln(1-y)$$

I can generate samples from the uniform distribution

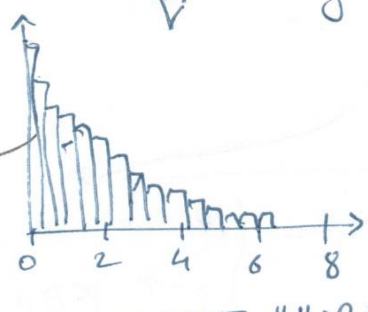


I then pass all the samples into the inverse CDF $X = -\ln(1-y)$



It's following the distribution of our pdf

let's draw a histogram to see it



Yes! it's following the pdf

But why/how does it work? IT WORKS THANKS TO

Proba. integral transform: $Y = F_X(X)$ (the CDF function transform the distribution of a rv to a uniform dist)

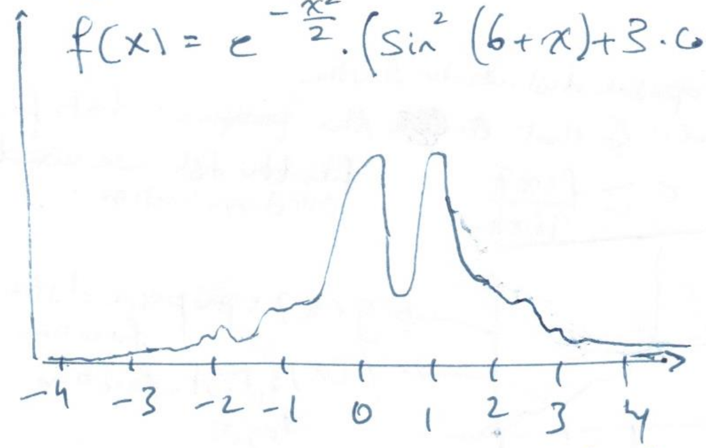
$\Longrightarrow \underline{X = F_X^{-1}(Y)}$

↳ inv. of CDF transforms uniform distribution to the distribution of the r.v

However this algo is very limited bcz we need to have a CDF for the distribution we want to sample from.
& in reality, not all distributions have CDFs that can be expressed in analytic form.

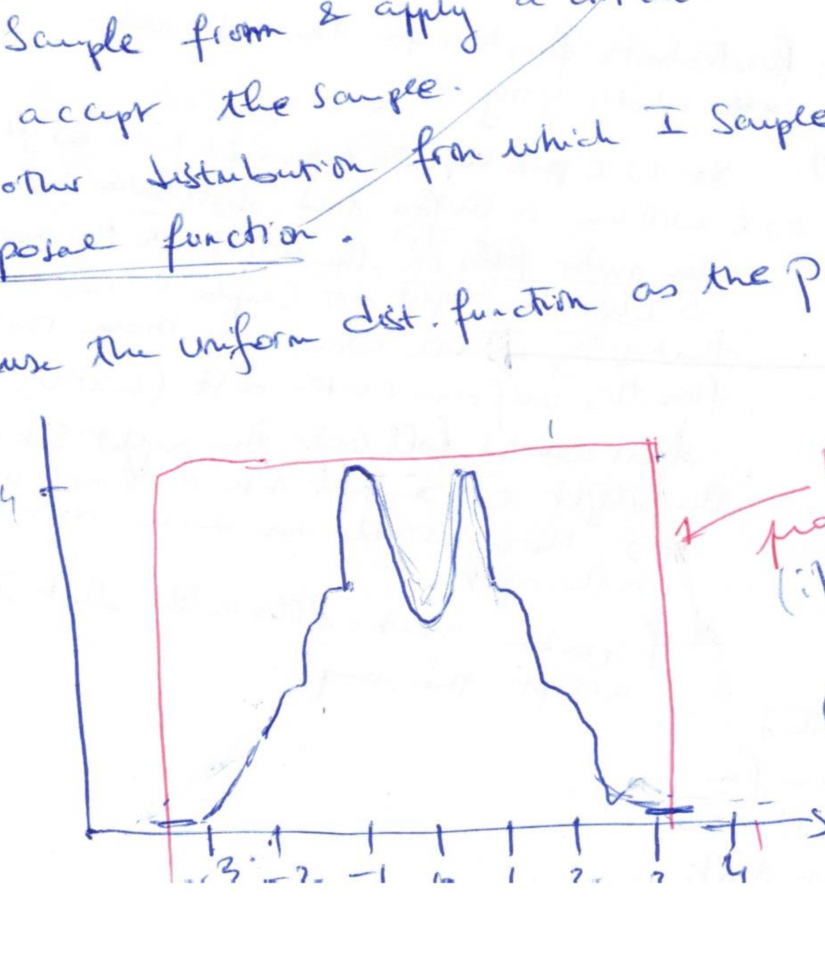
(B) 2nd Algo: Rejection Sampling:

$$f(x) = e^{-\frac{x^2}{2}} \cdot \left(\sin^2(6+x) + 3 \cdot \cos^2(x) \cdot \sin^2(4x) + 1\right)$$



Common Problem: If I cannot sample from target distribution function => I use another distribution function to sample from & apply a criteria to reject & accept the sample.
This other distribution from which I sample is called proposal function.

let's use the uniform dist. function as the proposal function

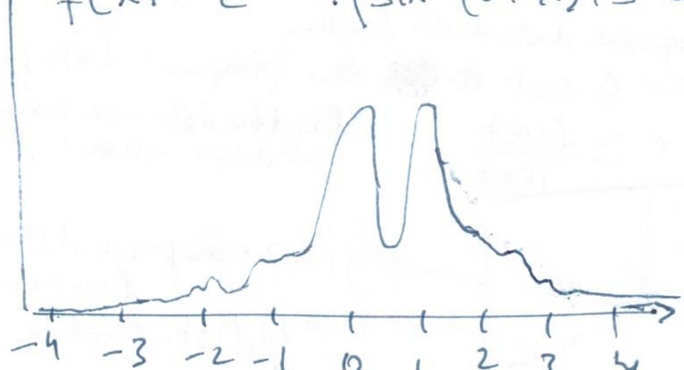


Here's the proposal function (it's a uniform dist function)
(scaled over all of my target dist)

However this algo is very limited bcz we need to have a CDF for the distribution we want to sample from.
& in reality, not all distributions have CDFs that can be expressed in analytic form.

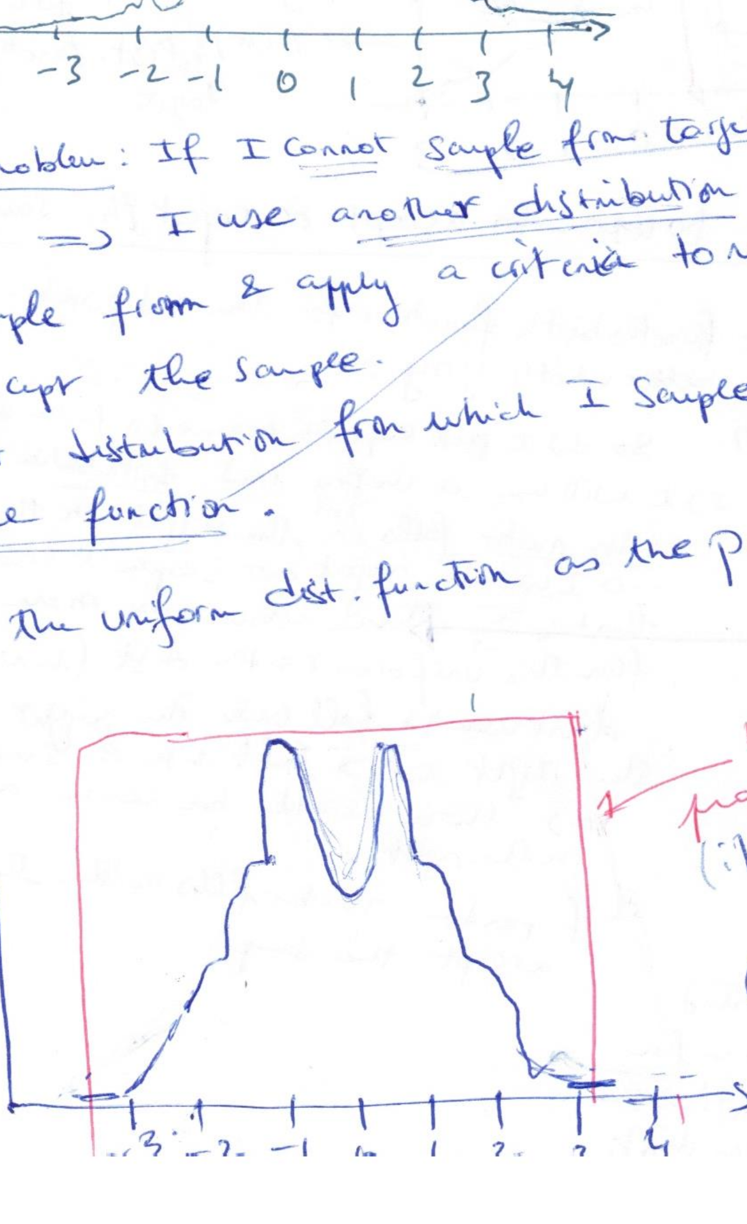
(B) 2^{nd} Algo: Rejection Sampling:

$$f(x) = e^{-\frac{x^2}{2}} \cdot \left(\sin^2(6+x) + 3 \cdot \cos^2(x) \cdot \sin^2(4x) + 1\right)$$



Common Problem: If I cannot sample from target distribution function => I use another distribution function to sample from & apply a criteria to reject & accept the sample.
This other distribution from which I sample is called proposal function.

let's use the uniform dist. function as the proposal function

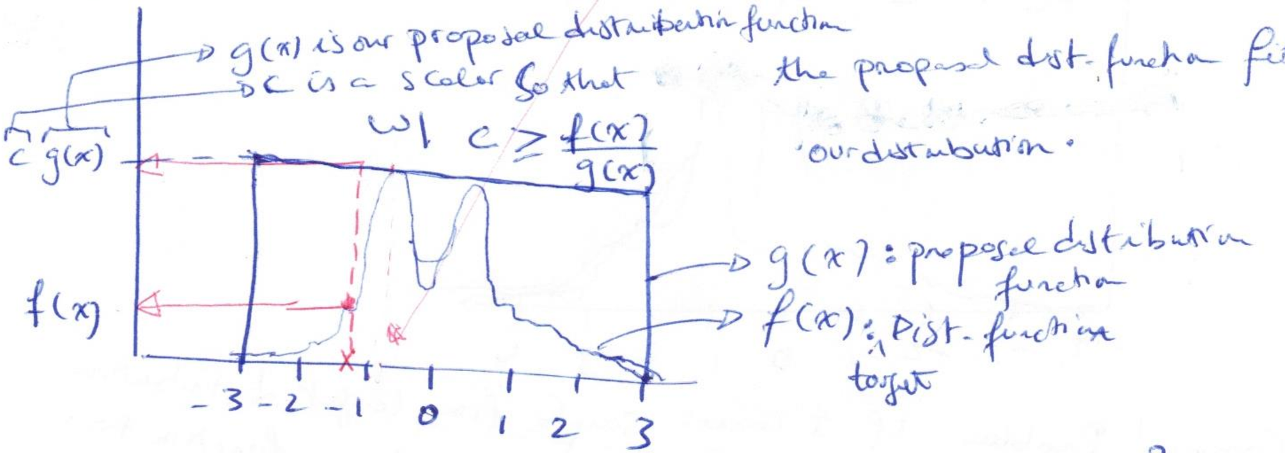


Here's the proposal function (it's a uniform dist function)

(scaled over all of my target dist)

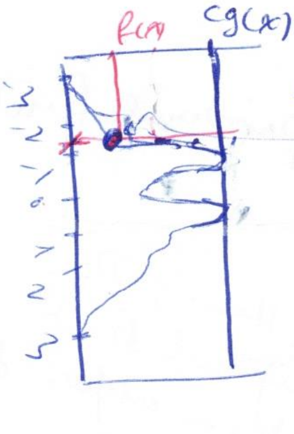
Now, I should sample from my proposal function (the uniform distribution function (this sample should fall in red between -3 & 3)

so let's say we do it & get the point in red. Our task now is to see whether to accept or reject it



What criteria to use to accept or reject the sample?

let's evaluate both functions for the red point.
I'll rotate to see better what's going on



1) I will sample from proposal dist. function & I get the red point
2) I will use a uniform rand. distribution : if random number falls on the right on the marker (●)
→ I would reject our sample x (red x)
that said, there would be more random numbers from this uniform random dist ($u \sim U(0, c \cdot g(x))$) that would fall under the right side since the right side > left side & since uniform
⇒ there would be more numbers falling on the right
& if random number falls on the left I would accept the sample.

Note: I have generated 2 random numbers: one from the proposal dist function & one from uniform dist.

=> u : sample we get from the 2nd uniform distribution

w/ $u \sim U(0, c \cdot g(x))$ (see rotated graph)

=> Acceptance Criteria:

$$u \leq f(x) \text{ for } u \sim U(0, c\,g(x))$$

It can be also written:

$$u \leq \frac{f(x)}{c\,g(x)} \text{ if } u \sim U(0,1)$$

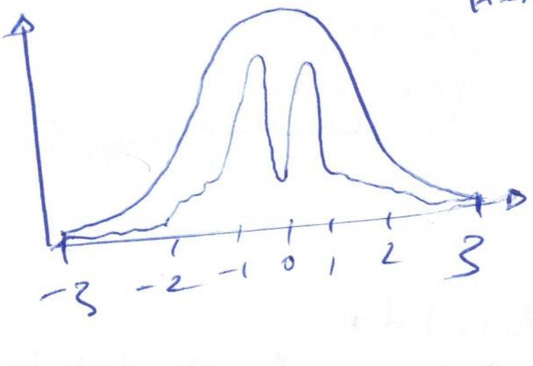
So far we have used uniform distribution as a proposal dist function : it works, but it has limitations:

for ex: in our example, for a uniform dist

the rejection region is very large.

=> Let's use another proposal function:

Here rejection region is smaller



If we compare acceptance rates for both models:

I get:

Uniform	Gaussian
23,5%	53,63%

(c) IMPORTANCE SAMPLING:

↳ is used to compute expected values rather than estimating densities

↳ Not a sampling algo, rather a trick to find expectations & provide solutions to intractable integrals.

Our main task in Bayesian inference is to compute expectations: Posterior predictive dist. is an expected value

- $E_{p_\theta}[h(x)] = \sum_i^{\infty} h(x_i)\, p_\theta(x_i)$

$E_{p_\theta}(h(x)) = \int_K h(x)\, p_\theta(x)\, dx$

Main challenge: Intractable integration

Solution: Law of large n

$\frac{1}{N}\sum_{i=1}^{N} h(x_i) \approx E_{p_\theta}[h($

But even here, the new challenge is: what should be the value of N? How many samples we should have?

3 main challenges from expected value computation

1 - Sampling aspect: We cannot sample from distribution p ⟹ can't apply LLN

2 - Inefficiency aspect: maybe we can sample, but it's inefficient (I need huge amount of data (samples) to reach convergence ⟹ more computation time).

Also, large amount of sample → Large variance

Also, we may not be able to capture events pro

Imp. Sampling provides a solution to this (variance reduction algo.

3 - p_θ is not normalized. $p_\theta = \frac{\tilde{p}_\theta(x)}{Z_p}$ → cannot compute

!! PS: A function is considered a pdf ONLY IF IT IS NORMALIZED (we use PDF to compute prob of a rv, prob is a number btwn 0 & 1) let's adress the first 2 challenges: $E_{P_\theta}[h(X)] = \int_R h(x) P_\theta(x) dx$

Scenario: while I cannot sample from r.v. distribution shown here as P, we can evaluate its density function

Monte Carlo approach: If I cannot sample from dist of r.v
$\Rightarrow$ I should use a proposal dist & then use an acceptance criteria.

let's bring the proposal dist into the expectation:

$$E_{P_\theta}[h(X)] = \int_R h(x) \frac{q_\phi(x)}{q_\phi(x)} P_\theta(x)\, dx$$

q → function of our sampling dist (proposal dist)

(switch p & q)

$$= \int_R h(x) \frac{P_\theta(x)}{q_\phi(x)} q_\phi(x)\, dx$$

right most place is reserved for proba dist.

I change subscript to q

$$E_{q_\phi}(h'(x)) = \int_R \underbrace{h(x) \frac{P_\theta(x)}{q_\phi(x)}}_{\text{& function of r.v has also changed} \to \text{we call it } h'(x)} q_\phi(x)\, dx$$

Clear Equation:

$$E_{q_\phi}[h'(x)] = \int_R h(x) \overbrace{\frac{P_\theta(x)}{q_\phi(x)}}^{\text{Importance weight}} \cdot q_\phi(x)\, dx.$$

$$\text{LLN} \approx \frac{1}{N} \sum_{i=1}^{N} h(x_i) \frac{P_\theta(x_i)}{q_\phi(x_i)}$$

How to choose q_ϕ?