Stochastic Simulation
29 October 2020 18:48

What is a stochastic model?

⇒ It consists basically of two key ingredients

1. Stochastic Input Variables
$\mathbb{X} = (X_1, \dots, X_p) \in \mathbb{R}^p$

2. A deterministic function $h(\mathbb{X})$ mapping
$h: \mathbb{X} \in \mathbb{R}^p \to Y \in \mathbb{R}$

and I is interested in finding the properties of the distribution of Y, i.e. of $Y \sim \Pi$

For instance the most basic approach to stochastic simulation would involve the following:

1. Draw a sample of size N of $\mathbf{X}$, i.e.,
$$\mathbf{X}_i = (X_{i1}, \dots, X_{ip}), \quad i = 1, \dots, N$$
2. Compute the corresponding output
$$\mathbf{Y}_i = h(X_{i1}, \dots, X_{ip}), \quad i = 1, \dots, N$$
3. The law of large numbers justifies the use of approximations like
$$\mathbb{E}(Y_1) \approx \frac{1}{N} \sum_{i=1}^{N} Y_{i1}$$
$$\mathbb{P}(Y_1 \le c) \approx \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{\{Y_{i1} \le c\}}$$
} get CDF by simulating Y_i

1) Bootstrap - Not known $\mathbb{X}$ distribution

A slightly more interesting case in comparison to the one above, happens when the distribution of $\mathbb{X}$ is unkown,

$$\mathbb{X} \sim ?, \text{ with } \mathbb{X} \in \mathbb{R}^p$$

In such a case a technique often used is the one of the bootstrap, which involves a two-fold data sampling from some observable sample,

which involves a "two-fold data" sampling from some observable sample, say

$x_1, \dots, x_p$ } 1 sample for X.

The idea is then the following:

Step 0: Take the sample $x \in \mathbb{R}^p$ as being distributed according to the X.

Step 1: Generate N samples for the distribution X by sampling with replacement from $x \in \mathbb{R}^p$, p-times.

Step 2: Take the distribution of the sample generated from step 1 as approximation for X.

You can now generate N samples of Y, by using the N-samples from step 1 and plugging them into $h(.)$.

Step 3: Leverage the LLN for computing the distribution of Y.

Notice that the general idea behind the bootstrap is the following:

$$\begin{array}{ccccc} F & \xrightarrow{\text{sample}} & F_n & \xrightarrow{\text{bootstrap}} & F_n^* \\ h(x)\downarrow & & h(x)\downarrow & & h(x)\downarrow \\ \theta & \xrightarrow{?} & \hat{\theta} & \xrightarrow{\text{bootstrap}} & \theta^{**} \end{array}$$

So we observe F_n in practice and do not know the distribution F.

So we observe F_n in practice and do not know the distribution F.

The question is now how far is $\hat{\theta}$ from the unknown θ? What is the distribution of θ?

Assuming in bootstrap that:

- the distribution F_n is representative of the distribution F.
- the map $h(\cdot)$ is smooth such that if an $x_i \in F_n$ is close to an $x_j \in F$ then so do $\hat{\theta}_i$ and θ_j respectively.

Then, given these conditions, you have that you can infer the distribution of $\hat{\theta}$ by observing the θ^* obtained by the bootstrap.

In the next sections we will introduce a couple of fields where stochastic simulation is especially prevalent and we will see some techniques used in the space.

(1) Bayesian Statistic

The entire concept of bayesian statistic in contrast to frequentist statistics is to model the parameter of interest as a random variable and try to model its distribution.

This in contrast to frequentist statistic where we treat the parameter(s) of choice as fix and unknown.

The basic idea of bayesian statistics involves then to model the distribution of the parameter in the

involves then to model the distribution of the parameter in the following way:

$$f(\theta|x) = \frac{p(x|\theta)\cdot\pi(\theta)}{\int_{-\infty}^{\infty} f(x|\theta')\pi(\theta')\,\theta}$$

$$\propto \underbrace{p(x|\theta)}_{\text{likelihood of the data given the parameter of interest } \theta} \cdot \underbrace{\pi(\theta)}_{\text{prior distribution for the parameter}}$$

You can then immediately see that in the above setting we have a known distribution

$$X \sim p(x|\theta)$$

a deterministic map $h(.)$,

$$h(x) = p(x|\theta)\cdot\pi(\theta)$$

and our output variable for which we want to get information on the distribution:

$$Y \sim \pi(\theta|x)$$

we will see now that in some simple case, when we work with conjugate priors, we might well be able to get as a posterior distribution a well known distribution but in other situation this does not have to be the case and we might be forced to use a simulation procedure involving:

1. generation of N samples
2. Computation of Y_N samples
3. Relying on LLN to infer the desired information of the

(3.) Relying on LLN to infer the desired information of the distribution.

Case 1: Conjugate Prior

- Assume $X_1, \dots, X_n$ i.i.d. $\sim N(\theta, \sigma^2)$

So that:

$$p(X|\theta) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\left(\sum X_i - \theta\right)^2\right)$$

- Assume θ is $N(\xi, K)$ distributed

$$\pi(\theta) = \frac{1}{\sqrt{2\pi K^2}} \exp\left(-\frac{1}{2K^2}(\theta-\xi)^2\right)$$

Then it is easy to see that the posterior $\pi(\theta|x)$ belongs to a well known distribution.

$$\pi(\theta|x) \propto p(\theta|x)\,\pi(\theta)$$

$$= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2}\left[\left(\sum X_i\right)^2 + \dots\right.\right.$$

(constant; not dependent on θ)

$$\left.\left. - 2\theta\sum X_i + n\theta^2\right]\right)$$

$$\cdot \frac{1}{\sqrt{2\pi}K} \exp\left(\frac{1}{2K^2}(\theta^2 - 2\theta\xi + \xi^2)\right)$$

(constant)

$$= \exp\left(-\frac{1}{2}\left(\frac{n\theta^2}{\sigma^2} + \frac{\theta^2}{K^2} - \frac{2\theta\xi}{K^2} + \dots\right.\right.$$

$$\left.\left. - \frac{2\theta\sum X_i}{\sigma^2}\right)\right)$$

$$= \exp\left(-\frac{1}{2}\cdot\frac{1}{\gamma^2}\left(\theta - \mu(x)\right)^2\right)$$

this follows noticing that

$$\frac{1}{\gamma^2} = \frac{n}{\sigma^2} + \frac{1}{K^2} = \frac{nK^2 + \sigma^2}{\sigma^2 K^2}$$

and

$$\mu(x) = \left(\frac{\xi}{K^2} + \sum X_i\right)\cdot \gamma^2$$

and

$$\mu(\bar{x}) = \left(\frac{\xi}{K^2} + \frac{\Sigma x_i}{\sigma^2} \right) \cdot \sigma^2$$

↳ idea when computing the square $\mu(\bar{x})^2$ falls away as it is in constant in θ Moreover due to the squared term you would get the expression above.

$$\Leftrightarrow \mu(\bar{x}) = \left(\frac{\xi}{K^2} + \frac{\Sigma x_i}{\sigma^2} \right) \frac{\sigma^2 K^2}{nK^2 + \sigma^2}$$

$$= \frac{\sigma^2}{nK^2 + \sigma^2} \xi + \frac{K^2 \cdot n}{nK^2 + \sigma^2} \frac{1}{n} \Sigma_i x_i$$

so that convex combination among sample mean and prior mean.

Notice finally that the above is a well known distribution, i.e. it is proportional to a normal distribution, so that it is easy to infer all of the desired properties of

$$Y_i \sim \pi(\theta | x)$$

analytically.

We might for instance be interested in computing:

- posterior mean:

$$E(g(\theta) | X = x) = \int_\Theta g(\theta) \cdot \pi(\theta | x) \, d\theta$$

- Credible Interval

$$P(g(\theta) \in I(x) \mid X = x) = 1 - \gamma$$

- posterior predictive distribution.

↳ this happens when for a RV y you have conditioning on θ the independence from X and the same distribution, so that

↳ this happens when for a RV Y you have conditioning on θ the independence from X and the same distribution, so that

$$\begin{aligned}\mathbb{P}(Y \in B \mid \mathbf{X} = \mathbf{x}) &= \int_\Theta \mathbb{P}(Y \in B \mid \theta)\pi(\theta \mid \mathbf{x})d\theta \\ &= \int_B \int_\Theta p_{Y|\theta}(y)\pi(\theta \mid \mathbf{x})d\theta d\mathbf{y}\end{aligned}$$

So far we treated the case of conjugate priors. The more interesting case in simulation is when the posterior distribution is not a well known distribution such that we will have to rely on simulation.

This for instance happens in the easy case when:

- **Likelihood**: $X_1, \ldots, X_n$ *i.i.d.* $\sim \mathcal{N}(\theta, \sigma^2)$

$$p_{\theta,\sigma^2}(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}} \frac{1}{\sigma^n} \exp\left(-\frac{1}{2\sigma^2}\left(ns^2 + n(\bar{x} - \theta)^2\right)\right)$$

where
  - $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$
  - $s^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2$

- **Prior**: $\theta \sim \mathcal{N}(\xi, \kappa^2)$, $\frac{1}{\sigma^2} \sim \text{Gamma}(\gamma, \lambda)$, independent

$$\pi(\sigma^2) = \frac{\lambda^\gamma}{\Gamma(\gamma)} \left(\frac{1}{\sigma^2}\right)^{\gamma+1} \exp\left(-\frac{\lambda}{\sigma^2}\right)$$

It is then possible to prove that:

- **Posterior**:

$$\begin{aligned}\pi(\theta, \sigma^2 | \mathbf{x}) &\propto p_{\theta,\sigma^2}(\mathbf{x})\pi(\theta, \sigma^2) \\ &\propto \left(\frac{1}{\sigma^2}\right)^{0.5n+\gamma+1} \exp\left(-\frac{1}{2\sigma^2}\left(ns^2 + n(\bar{x} - \theta)^2\right)\right) \\ &\quad \cdot \exp\left(-\frac{1}{2\kappa^2}(\theta - \xi)^2 - \frac{\lambda}{\sigma^2}\right)\end{aligned}$$

Which does not belong to any well known distribution.

It is then for instance obvious to see that if you might want to get a posterior predictive density, where conditioning on the parameters the Y does not depend on the observable data anymore, we have:

$$P(Y \leq b \mid X = x) = \int_\Theta P(Y \mid \theta, \sigma^2) \cdot \pi(\theta, \sigma^2 \mid x)\, d\theta d\sigma^2$$

$= E(Y)$ given kn of

$= E(Y)$ given kn of total expectation.

$\approx \frac{1}{N} \sum P(Y|\theta_i, \sigma_i^2)$

where you simulate $\theta_i, \sigma^2 \sim \pi(\theta, \sigma^2|X)$ derived before.

Another interesting case that will bridge directly to the next simulation technique is the case where we might not have a known distribution for:

$$\pi(\sigma^2, \theta|X) = ??$$

You might then decompose the above in the following

$$\pi(\sigma^2|\theta, X) = \frac{\pi(\sigma^2, \theta|X)}{\pi(\theta|X)} \quad \text{constant in } \theta, X$$

$$\propto \pi(\sigma^2, \theta|X)$$

equally

$$\pi(\theta|\sigma^2, X) \propto \pi(\sigma^2, \theta|X)$$

such that you can iteratively come to the desired distribution leveraging the so called Gibbs Algorithm that leverages the full conditional distributions above (recall that we assume here that conditional probabilities are known), in the sense that you iteratively converge to the desired distribution.

iteratively.

Algorithm (Gibbs sampler)

Choose / simulate initial values (θ_0, σ_0^2)

For $t = 1, 2, \ldots,$ *simulate*

1. $\theta_t \sim \pi(\theta|\sigma_{t-1}^2, \mathbf{x})$
2. $\sigma_t^2 \sim \pi(\sigma^2|\theta_t, \mathbf{x})$ → you hit them as known and fix!

Notice that convergence to the distribution of interest is guaranteed by some Markov Chain logic in the above. To see that read the notes on stochastic simulation!

by some Markov chain logic in the above. To see that read the notes on stochastic simulation!

This section brings us directly to a second important topic in the space of stochastic simulation - statistical mechanics. This is important as there it was in fact created the the Gibbs Algorithm and it is fundamental to the understanding of hamiltonian MCMC that we will take approach.

(2) Statistical Mechanics

One important application in statistical mechanics involves the modeling of the Gibbs Distribution / Boltzmann Distribution.

This models the probability of being in a state given the temperature and the energy state level.

It can be expressed as:

The **Gibbs distribution** on Ω is given by

$$\pi(\mathbf{x}) = \frac{1}{Z(\beta)} \exp\left(-\beta \sum_{i \neq k,\ i,k \in L} J_{ik} x_i x_k\right)$$

where
- $Z(\beta)$ is a normalizing constant } very difficult to compute
- β is the inverse temperature: $\beta = \frac{1}{k_b T}$ (k_b: Boltzmann's constant)
- $J_{ik} = J_{ki}$ denotes the interaction between particles at sites i and k

Notice that in the model above $\mathbf{x}$ is a particular state configuration that involves values for all of the particles x_i on a spin system lying on a lattice $L = \{1, 2, \ldots, n\}^d$.

Notice moreover that each state $\mathbf{x} \in \Omega = \{\pm 1\}^2$ so that each particle takes either the value of 1 or -1.

Further things to know are:

① $\sum_{\substack{i=1 \\ k \neq i,\ k=1}}^{k} J_{ik} x_i x_k$ is called the energy configuration of the system.

① $\sum_{i=1}^{k} \sum_{k \neq i, k=1} J_{ik} x_i x_k$ is called the energy configuration of the system.

↳ it follows immediately from the above that

$$\begin{cases} \text{if } x_i = x_k \Rightarrow \pi(x) = \frac{1}{Z(\beta)} \exp\left(-\beta \sum_i J_{id}\right) \\ \text{if } x_i \neq x_k \Rightarrow \pi(x) = \frac{1}{Z(\beta)} \exp\left(\beta \sum_i J_{ik}\right) \end{cases}$$

so that it follows:

- If $J_{ik} < 0$, there is a preference for equal spins at sites i and k
- If $J_{ik} > 0$, there is a preference for opposite spins at sites i and k
- The absolute value of J_{ik} expresses the strength of this preference

② The temperature has an important impact on the state probability function.

In particular:

If $\beta \to 0$ $(T \to \infty)$, then $\pi(x) \to c$, where c is a constant, such that all of the state configurations are equally likely.

If $\beta \to \infty$ $(T \to 0)$, then

$$\pi \to \begin{cases} 1/N & \text{for states in the minimum energy configuration} \\ 0 & \text{else} \end{cases}$$

To see that consider

$$x' \in \operatorname{argmin}_x \sum_{\substack{i,k \in I \\ k \neq i}} J_{ik} x_i x_k$$

$$x \notin \text{---------- " ----------}$$

then,

$$\frac{\pi(x')}{\pi(x)} = \exp\left(\beta \left(\sum J_{ik} x_i' x_k' - \sum J_{ik} x_i x_k\right)\right)$$

< 0 by def

$$\frac{\pi(x)}{\pi(x')} = \exp\left(\beta\left(\sum_{jk} J_{jk} x_j' x_k' - \sum_{jk} J_{jk} x_j x_k\right)\right)$$

$\sum_{jk} J_{jk} x_j' x_k' - \sum_{jk} J_{jk} x_j x_k < 0$ by def of minimum energy state

$$\rightarrow \lim_{\beta \to \infty} \frac{\pi(x)}{\pi(x')} = 0$$

Then given the fact that $\sum_x \pi(x) = 1$, it follows

$$\sum_{x \in \operatorname{argmin} \sum J_{jk} x_j x_k} \pi(x) + \sum_{x \notin \operatorname{argmin} \sum J_{jk} x_j x_k} \pi(x) = 1$$

$$\frac{1}{\pi(x')} \left(\sum_{x \in \operatorname{argmin} \sum J_{jk} x_j x_k} \frac{\pi(x)}{\pi(x')} + \sum_{x \notin \operatorname{argmin} \sum J_{jk} x_j x_k} \frac{\pi(x)}{\pi(x')} \right) = \frac{1}{\pi(x')}$$

$= 1$ as x from argmin

$= 0$ by previous prove

$$\Leftrightarrow \quad \underbrace{\left| x \in \operatorname*{argmin}_x \sum J_{jk} x_j x_k \right|}_{N} = \frac{1}{\pi(x')}$$

number of such occurrences

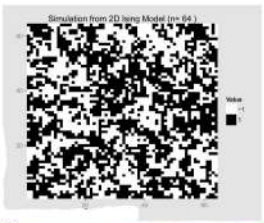
$$\Rightarrow \quad \pi(x) = \begin{cases} 1/N & \text{if } x \in \operatorname{argmin} \sum J_{jk} x_j x_k \\ 0 & \text{else} \end{cases}$$

□

We will now show the benefit of leveraging simulation in the case of a type of gibbs distribution modeling - the Ising Model.

In the **Ising model**, one assumes

$$J_{ik} = \begin{cases} 0 & \text{if } \|i-k\| \neq 1 \\ -1 & \text{if } \|i-k\| = 1 \end{cases}$$ No interaction for non-neighbouring particles.



We consider the distribution of the **average magnetization**:

$$M_n = \frac{1}{n^d} \sum_{i \in L} X_i$$

Recall here you tend to all of the x to be uniformly distributed, such that with the correct normalization (here n^d) M_n becomes binomial dist due to CLT.

- For large fixed n and $\beta \to 0$, $n^{d/2} M_n$ is approximately standard normal by the central limit theorem
- For $\beta \to \infty$ and for fixed n, $M_n = \pm 1$ with probability $\frac{1}{2}$ each
- For some values β in between, the distribution of M_n will thus be bimodal
- What happens if we keep β fixed and let n go to infinity: Will the two modes disappear, or will there be a bimodal and thus non-normal limit distribution?

give rise to CLT.

If two modes remain we talk about spontaneous magnetization.

In order to investigate if the two modes remain a simulation approach can be take leveraging again full conditional distributions and the Gibbs Algorithm.

It is namely possible in this sense to show that

$$P(X_i = 1 \mid X_{(i)}) = \frac{\exp(-\beta 2A_i)}{\exp(-\beta 2A_i) + \exp(\beta 2A_i)}$$

where $A_i = \sum_{k,l \in L, k \neq i, l \neq i} J_{ik} X_k$ } notice it does not depend on X_i the variable of interest.

Proof: to see this understand, that fixing X_i you can write for any $i \in L$

$$\sum_{k \neq l} J_{kl} X_k X_l = \underbrace{\sum_{k \neq i} J_{ik} X_i X_k}_{\text{fixed } X_i,\ := A_i} + \underbrace{\sum_{l \neq i, k \neq l, k \neq i} J_{kl} X_k X_l}_{:= B_i}$$

due to the simmetry you would also have $J_{ki} X_k X_i$ but this the same as $J_{ik} X_i X_k$

It follows now that

$$P(X_i = +1 \mid X_{(-i)}) = \frac{P_r(X_i, X_{(-i)})}{C}$$ } normalizing constant depending on $X_{(-i)}$

$$= \exp(-\beta(2A_i + B_i)) \cdot \frac{1}{C}$$

$$P(X_i = -1 \mid X_{(-i)}) = \frac{P_r(X_i, X_{(-i)})}{C}$$

$$= \exp(-\beta(\underbrace{-2A_i}_{\text{from inserting } X_i = -1 \text{ above}} + B_i))$$

It follows immediately that

$$P_r(X_i = 1 \mid X_{(-i)}) = \frac{P_r(X_i = 1 \mid X_{(-i)})}{\quad}$$

$$\Pr(X_i = 1 \mid X_{(-i)}) = \frac{\Pr(X_i = 1 \mid X_{(-i)})}{\Pr(X_i = 1 \mid X_{(-i)}) + \Pr(X_i = -1 \mid X_{(-i)})}$$

$$= \frac{\frac{1}{C} \cdot \exp(-\beta(2A_i + B_i))}{\frac{1}{C}\left[\exp(-\beta(2A_i + B_i)) + \exp(-\beta(-2A_i + B_i))\right]}$$

$$= \frac{\exp(-\beta(2A_i + B_i))}{\exp(-\beta(2A_i + B_i)) + \exp(-\beta(-2A_i + B_i))}$$

$$= \frac{\exp(-\beta 2A_i) \cdot \exp(-\beta B_i)}{\exp(-\beta 2A_i) \cdot \exp(-\beta B_i) \cdot (1 + \exp(4\beta A_i))}$$

Given the above straight forward derivation of the full conditional probability it is now possible to leverage the Gibbs Sampler to explore iteratively the state space and through the Markov Chain setting that we will explore later it is possible to see that the entire space is explored such that the algorithm will converge to the steady state distribution and will eventually sample from the distribution of interest.

The **Gibbs sampler**

1. Choose an arbitrary initial configuration $\mathbf{X}^0$
2. For $t = 1, \ldots, N$, do the following:
   - Set $\mathbf{X}^{t+1} = \mathbf{X}^t$
   - Simulate particle X_i^{t+1} according to $\mathbb{P}(X_i = +1 | X_k^{t+1}, k \neq i)$ and keep the other components constant. Update the corresponding entry in $\mathbf{X}^{t+1}$
   - Repeat this step for all components i

3. Simulation in Financial Mathematics

Simulation is heavily leveraged in mathematical finance, think for instance to the job of derivatives pricing.

Here we propose a simulation method to efficiently estimate the loss of a portfolio and the default probabilities.

The idea is to model the portfolio loss as follows:

- **Portfolio model**: notation and assumptions
  - ℓ_j denotes the loss if the j-th debtor defaults. Assume that ℓ_j is deterministic with integer values and bounded $\ell_j \leq n_j < \infty$
  - Y_j is the indicator for this default event. Assume that the Y_j's are stochastic with some joint distribution
  - The total loss is then

$$L = \sum_{j=1}^{J} Y_j \ell_j$$

- ℓ_j denotes the loss if the j-th debtor defaults. Assume that ℓ_j is deterministic with integer values and bounded $\ell_j \leq n_j < \infty$
- Y_j is the indicator for this default event. Assume that the Y_j's are stochastic with some joint distribution
- The total loss is then

$$L = \sum_{j=1}^{J} Y_j \ell_j$$

- The **goal** is to compute the distribution of L

Assume the following **model for the default indicators** $(Y_1, \ldots, Y_J)'$:

- There are latent variables $W = (W_1, \ldots, W_p)'$ such that given W, the Y_j's are conditionally independent with

$$\mathbb{P}(Y_j = 1 \mid W) = f_j(W)$$

$f_j(W) = \frac{1}{1+\exp[-\sum_{i=1}^{p} a_{ji} W_i]}$ or, alternatively, $f_j(W) = \Phi\left(\sum_{i=1}^{p} a_{ji} W_i\right)$

- The W_i's are assumed to be independent with a known distribution, e.g., a Gaussian, Gamma or Lognormal distribution

} key and strong assumption

The idea is then for the simulation to get a precise estimation for the distribution of the loss of the portfolio L. This can be done by generating

simulate $W_{ji} \sim$ dist of W_i

plug in $h(W_i) \sim$ samples for default Y_i indicators

use L.L.N to infer the distributions properties of the default L by plugging in the simulated indicator variables.

Hence the standard approach works fine once you define the distribution of your economic underlying factors and given the loadings a_{ij}. (these are either taken as given by some platform provider or estimated via historic data through the logit model).

Notice however that it is difficult to get reliable estimator for the loss in the tail (the most interesting case!!)

In such a case relying on standard simulation methods is dangerous as you hardly enter the tail of the distribution; i.e. you need extremly long simulations for estimating it well

↳ Two solutions for the case are:

- Importance Sampling
  ↳ will be discussed later in the course

- Importance Transform

in the course

- Leveraging Fourier Transform

The idea of the second method will be discussed shortly next.

The basic idea of this second method is to express the distribution of the losses X_L through its characteristic function.

You would then leverage the conditional independence of the defaults conditioning on the economic variables to sample from the loss distribution and leveraging then the fourier inverse transform you can get estimates for the probability density at any given value of interest.

Mathematically:

$$\chi_L(\lambda) = E(e^{i\lambda L}) = \sum_{n=1}^{N} e^{i\lambda n} \cdot P(L=n)$$

where by definition here you are assuming

$$\sum_{i=1}^{N} l_i \leq N$$ such that the loss is capped (for instance to your leveraged investment).

You can then immediately see that it is possible to simulate from the loss distribution by:

$$\chi_L(\lambda) = E(E(e^{i\lambda L} | W))$$

$$= E(E(e^{i\lambda \sum_j \gamma_j l_j} | W))$$

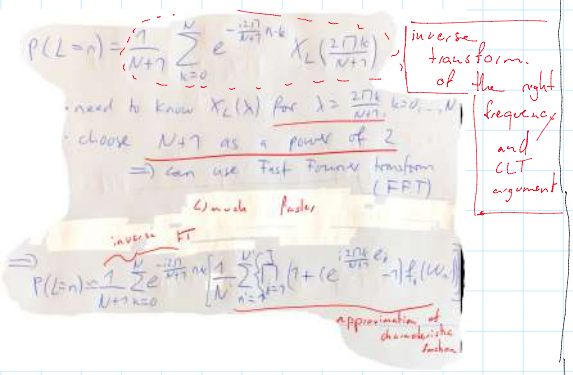
$$= E\left(\prod_{j=1}^{J} 1 - f(W) + f(W) \cdot e^{i\lambda \sum_j \gamma_j l_j}\right)$$ (here $l_j = 0$ s.t. $\gamma_j = 1$)

$$\overset{\text{via CLT}}{\approx} \frac{1}{N'} \sum_{n'=1}^{N'} \prod_{j=1}^{J} 1 + (e^{i\lambda \gamma_j l_j} - 1) f(W_{n'})$$

Then given the above you know you can obtain the probability density function for a particular value by sampling from the right frequency from the fourier transform of the PDF and

frequency from the fourier transform of the PDF and use the inverse fourier transform to go back to the PDF.

↳ From this it is then straightforward to see that you can sample from the tails by:



As a final step to this introductory chapter on simulation, we will explore Monte Carlo Methods.

↳ These are heavily used in simulation as they have the nice property of convergence to the desired statistic of interest in $\sqrt{N}$ time and independently from the distribution dimension.

↳ This is basically due to the LLN and CLT as we will see next.

↳ So far in these notes we basically relied on these methods without making the last step explicit, which will do now, also deriving some of the most fundamental convergence properties of the methods.

On Monte Carlo Methods

Recall that our goal was to derive some properties of the $h(X) = Y \in \mathbb{R}$ distribution.

distribution.

This sums up to the following setting:

$$h: A \subseteq \mathbb{R}^p \to \mathbb{R}$$

where we are then interested in

$$\int_A h(x)\,dx$$

Notice that by suitable transformation you can express the above w.l.o.g. on an integration on the unit cube $A \in [0,1]^p$ such that you can approximate the above by computing the necessary transformation to $h(.)$ such that $A \in [0,1]^p$ and then:

- Generate $N \times p$ uniform random variables U_{ij} on $[0,1]$
- Use the approximation

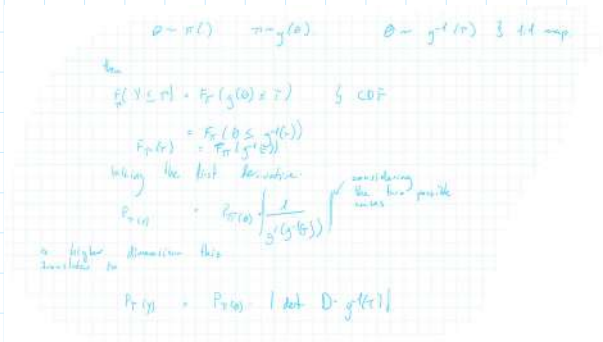
$$\frac{1}{N}\sum_{i=1}^{N} h(U_{i1}, \ldots, U_{ip})$$

Notice that using a transformation

$$T: A \in \mathbb{R}^p \to [0,1]^p$$

that is 1-to-1 it is immediate to see that $h(\underbrace{T(x)}_{=U})$ on the new unit cube can be easily found applying the change of variable formula

Idea:



Such That:

Such that:

$$\int_A h(\mathbf{x})d\mathbf{x} = \int_{T(A)} h\left(T^{-1}(\mathbf{u})\right) \left|\det\left(DT^{-1}\right)(\mathbf{u})\right| d\mathbf{u},$$
$$= \int_{[0,1]^p} h\left(T^{-1}(\mathbf{u})\right) \left|\det\left(DT^{-1}\right)(\mathbf{u})\right| 1_{T(A)}(\mathbf{u})d\mathbf{u},$$

Why to use Monte Carlo Methods and not numerical integration?

Recall: Numerical Integration

Numerical integration

- Choose N deterministic points $\mathbf{x}_i \in [0,1]^p$
- Use the approximation $\sum_{i=1}^N w_i h(\mathbf{x}_i)$ with given weights w_i
- In the simplest case, the $\mathbf{x}_i$ are on a cubic lattice

$$\mathbf{x}_i \in \left\{\frac{1}{2K}, \frac{3}{2K}, \ldots, \frac{2K-1}{2K}\right\}^p$$

with $N = K^p$ and the weights are constant

Reason

Convergence rates

- **Numerical integration**: For smooth functions h, one can show that in the simple lattice case, the **convergence rate is equal to** $N^{-1/p}$, which is very slow in high dimensions
- **Monte Carlo integration** has **convergence rate** $N^{-1/2}$ independent of the dimension p and without any smoothness assumptions

↳ CLT shrinks at this rate

The benefit is given by using a stochastic and not a deterministic method such that through the laws of statistics you get a rate which does not depend on the dimension of the problem.

Proof.

The proof of this is straightforward given the three classical mathematical statistics results:

Law of Large numbers (LLN)

$\hat{\theta}_N = \frac{1}{N}\sum_{i=1}^{N} h(X_i)$ $\quad X_i$ iid, $i = 1, \ldots N$

$E(|h(X)|) = \theta < \infty$

$\Rightarrow$ • $\hat{\theta}_N \xrightarrow{P} \theta$, i.e. $P(|\hat{\theta}_N - \theta| > \varepsilon) \xrightarrow{(N\to\infty)} 0, \forall \varepsilon$

(weak LLN) Convergence in probability

• $\hat{\theta}_N \xrightarrow{a.s.} \theta$ i.e. $P(\lim_{N\to\infty} \hat{\theta}_N = \theta) = 1$

$\Rightarrow$ • $\hat{\theta}_N \xrightarrow{P} \theta$, ie $P(|\hat{\theta}_N - \theta| > \varepsilon) \to 0, \forall \varepsilon$ (weak LLN) — convergence in probability

• $\hat{\theta}_N \xrightarrow{a.s.} \theta$, ie $P(\lim_{N\to\infty} \hat{\theta}_N = \theta) = 1$ (strong LLN) — almost sure convergence

Central Limit theorem (CLT)

$\hat{\theta}_N = \frac{1}{N}\sum_{i=1}^{N} h(X_i)$, X_i iid, $E(|h(X_i)|) = \theta$, $Var(h(X_i)) = \sigma^2 < \infty$

$\Rightarrow \sqrt{N}(\hat{\theta}_N - \theta) \xrightarrow{d} N(0, \sigma^2)$

ie $\lim_{N\to\infty} P(\sqrt{N}|\hat{\theta}_N - \theta| \le z) = \Phi(z/\sigma)$ — convergence in distribution

$\to$ CLT

Slutsky's theorem → could also be a.s. convergence

$Z_N \xrightarrow{d} Z$, $A_N \xrightarrow{P} a$ $(\in \mathbb{R}^1)$

$\Rightarrow A_N Z_N \xrightarrow{d} aZ$

In our case, $Z_N = \sqrt{N}(\hat{\theta}_N - \theta)$, $A_N = S_N^{-1}$ $(\xrightarrow{P} \sigma^{-1})$

$\Rightarrow \sqrt{N}\frac{(\hat{\theta}_N - \theta)}{S_N} \xrightarrow{d} N(0,1)$

Now given that the sample variance

$$S_N \xrightarrow{a.s.} \sigma^2$$

It follows immediately using Slutsky's Theorem and the CLT that given

$$P_{N\to\infty}\left(\sqrt{N}(\hat{\theta}_N - \theta) \le z\right) \sim N(0, \sigma^2)$$

$$P_{N\to\infty}\left(\frac{\sqrt{N}(\hat{\theta}_N - \theta)}{S_N} \le z\right) \sim N(0,1)$$

$\Leftarrow$ Slutsky's

such that:

$$P_{N\to\infty}\left(\hat{\theta}_N - \theta \le z_{\alpha\text{-quantile}}\right) - CI$$

$$= \hat{\theta}_N \pm \frac{S_N}{\sqrt{N}} \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$$

Notice also the interesting case of approximating not the first moment, as analyzed that far, but rather the cdf of a given event.

In such a case:

$\theta = \pi(A)$; ie probability of an

$$\theta = \pi(A) \quad ; \text{ i.e. probability of an event } A.$$

It follows immediately that this can be expressed as:

$$h(X) = \mathbb{1}_{\{X \in A\}}$$

so that with $X \sim \pi$ it follows that you can estimate the above by:

$$\hat{\theta}_N = \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}_{\{X_n \in A\}}$$

It is now interesting to see that θ is bernoulli distributed $B(\pi(A))$ so that its variance is:

$$\sigma^2 = \pi(A)\,(1 - \pi(A))$$

this is known here; we do not use the sample variance and slutsky's theorem

You can then immediately leverage the property to compute the minimum number of samples necessary to get to a given precision.

For instance, say that the error should be capped at 10% of the estimator value, then

$$P_r\left(\frac{\sqrt{N}(\hat{\theta}_N - \theta)}{\sigma} \leq z\right) \sim N(0,1)$$

so that

$$z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{N}} \leq 0{,}1\,\pi(A)$$

keep error bounded while keeping α-quantile confidence for the interval.

inserting the value for the variance

$$z \sqrt{\frac{\pi(A)\,(1-\pi(A))}{N}} \leq 0{,}1\,\pi(A)$$

Say that you want to keep the level of confidence at 95% then $z = 1{,}96$ and

$$1{,}96^2 \cdot \frac{\pi(A)\,(1-\pi(A))}{N} \leq \frac{1}{100}\,\pi(A)^2$$

It follows

$$385 \cdot \frac{\pi(A)\,(1-\pi(A))}{\pi(A)^2} \leq N$$

so that the smaller $\pi(A)$ the larger the

$\pi(A)^c$

so that the smaller $\pi(A)$ the larger the necessary sample will be. This is the exact issues that one faces when estimating tail probabilities.

A final interesting point in the case of Monte Carlo simulation is the one of getting an α-quantile CI for the q_α-quantile of a distribution.

↳ This method is interesting in the sense that it is non-parametric in the sense that it is unaffected by the underlying distribution of Y and this does not enter directly the calculations.

The idea to get the CI for the q_α-quantile is to estimate the ordered data samples for which we expect that the quantile will lie within the interval 95% of the times.

This works through the following theorem

Let
- $Y_i := h(\mathbf{X}_i)$,
- q_α be the α-quantile of Y_i: $q_\alpha = \inf\{y; \mathbb{P}(Y_i \le y) \ge \alpha\}$
- $Y_{(1)} \le Y_{(2)} \le \dots \le Y_{(N)}$ denote the ordered observations
- $\lfloor x \rfloor$ denote the integer part of $x \in \mathbb{R}$

Theorem
If $\mathbb{P}(Y_i \le q_\alpha) = \mathbb{P}(Y_i < q_\alpha) = \alpha$, then $[Y_{(k_1)}, Y_{(k_2)}]$ where

$$k_1 = \left\lfloor N\alpha + 0.5 - \sqrt{N\alpha(1-\alpha)}\,\Phi^{-1}\left(1-\frac{\gamma}{2}\right)\right\rfloor,$$
$$k_2 = \left\lfloor N\alpha + 0.5 + \sqrt{N\alpha(1-\alpha)}\,\Phi^{-1}\left(1-\frac{\gamma}{2}\right)\right\rfloor + 1,$$

is an approximate $(1-\gamma)$-confidence interval for q_α.

The idea is the following, no matter the actual distribution of the Y, the following holds true:

$$h(X_i) = \mathbb{1}_{\{Y_i \le q_\alpha\}}$$

so that you can estimate the ordered $Y_{(k)}$ for which the quantile inequality would hold true and given a large sample of $Y_{(i)}$ you know that the quantile is expected to lie on the

$$S_N = \sum_{i=1}^{N} h(X_i) = k$$

lie on the

$$S_N = \sum_{i=1}^{N} h(X_i) = k$$

observation due to the LLN, no matter the Y distribution.

The question is now to find upper and lower bounds for the ordered $Y_{(i)}$ such that the true quantile belongs to the interval with a given confidence.

It follows than that we are looking for:

① $P_r(Y_{(k_1)} > q_\alpha) \leq \alpha/2$ } upper bound

② $P_r(Y_{(k_2)} < q_\alpha) \leq \frac{\alpha}{2}$ } lower bound

Given the definition of what we search it is now straightforward to derive the above by noting

$$P_r(Y_{(k_1)} > q_\alpha) = P_r(\text{at most } k_1 - 1 \; Y_i \leq q_\alpha)$$

logic: should this not be the case then you would observe $Y_{(k_1)} \leq q_\alpha$ and this is not the desired case.

Given our simulation setting of S_N that approximates the quantile-distribution it follows immediately that:

$$P_r(Y_{(k_1)} > q_\alpha) \overset{\text{limit case}}{=} P_r(S_n \leq k_1 - 1)$$

hence we are searching the k_1 for which the above holds true with the desired confidence.

It follows now that given the CLT property

with the sample variance of $S_N[h(X)]$ $h(X)$ i.i.d bernoulli (α), we have:

$$\sigma_{N\{S_N\}} = N \cdot \alpha \cdot (1 - \alpha)$$

$$\mathbb{V}\{S_N\} = N \cdot \alpha \cdot (1-\alpha)$$

So that

$$\frac{S_N - N\alpha}{\sqrt{\mathbb{V}\{S_N\}}} \xrightarrow{d} N(0,1)$$

It hence follows

$$P_r(S_N \leq k_1 - 1) \approx \Phi\left(\frac{k_1 - 1 - N\alpha + 0{,}5}{\sqrt{N \cdot \alpha \cdot (1-\alpha)}} \leq 2\right) = \frac{\alpha}{2}$$

small sample correction term for $N \to \infty$

and for the CI

$$k_1 = N\alpha + 0{,}5 - \Phi^{-1}\left(z_{\alpha/2}\right) \cdot \sqrt{N \cdot \alpha \cdot (1-\alpha)}$$

and a similar result can be obtaind for the upper ordered observation for the quantile.

This concludes the first introductory chapter on the topic.

The next addresses the generation of uniform random numbers, a key topic in the case of simulation as many simulation methods will in fact leverage such uniform distributed RV to eventually sample from the distribution of interest.

Chapter 2 - Uniform RV Generation

- In practice, one uses **deterministic algorithms to generate i.i.d. $U(0,1)$** variables for stochastic simulation
  - Producing a deterministic sequence of values which imitates a sequence of i.i.d. uniform random variables is a somewhat contradictory requirement

So theoretically you should not do it, in practice:

- From a practical point of view, what matters are the **properties of the pseudo random numbers**
  1. **Statistical properties**
     - The pseudo-random numbers should behave in as many ways as possible like realizations of i.i.d. $U(0,1)$ random variables
     - There are deep mathematical theories developed by Kolmogorov and Martin-Löf which formalize the concept of "like realizations of i.i.d. uniform random variables"
  2. **Predictability** ⇒ you should not be able
  3. **Speed**

For instance this is key in

Martin-Löf which formalize the concept of "like realizations of i.i.d. uniform random variables"

For instance this is key in applications of cyber security

2. **Predictability** ⇒ you should not be able
3. **Speed** ⇒ should be fast.

- Different applications emphasize different properties
  - E.g., statistical properties (for Monte Carlo methods) vs. predictability (for gambling machines, cryptography)

we are rather concerned with the first one.

We will now try to expose techniques to generate pseudo-random numbers that fulfill at best the above properties. We will especially go through

1. Linear Congruential Methods
2. Other Methods (non-linear substructure)
3. Combination of Pseudo-Random Numbers Generators.

The algorithms we discuss generate pseudo random numbers (u_n) according to:

$$x_{n-1} = f(x_n), \quad u_n = h(x_n),$$

with given functions f and h and some starting or seed value x_0

A good generator should have the following properties:

Periodicity { 1. It has a **long period** length

Statistical properties { 2. The sequence of successive ***d*-tuples fills out the *d*-dimensional unit cube well**

Speed and Reproducibility as research { 3. It should be **efficiently computable** and **reproducible**

1. Linear Congruential Methods

Idea: leverage the following generator

A **linear congruential generator** has the form

$$u_n = x_n / M$$

and x_n satisfies the recursion

$$x_{n-1} = (ax_n + c) \bmod M$$

where $x_0, a, c, M \in \mathbb{N}$.

Notice on support $[0, M-1]$ so that dividing by M support $[0, 1)$.

- a, c, and M affect the quality of the generator (period length, even distribution of d-tuples)
- The period can never be larger than M

It can then be seen that for any M, there are multiple choices of a, c which can guarantee a sufficiently long period.

It can then be seen that for any M, there are multiple choices of a, c which can guarantee a sufficiently long period.

This is straightforward to see given the theorem

Theorem

1. *If $c \neq 0$, the period is equal to M for all x_0 iff c and M are relatively prime and if $a - 1 \mod p$ for all prime divisors p of M and also for $p - 4$ if M is a multiple of 4.*
2. *If $c = 0$, the period is equal to $M - 1$ for all $x_0 \neq 0$ iff M prim and $a^{(M-1)/p} \neq 1 \mod M$ for all prime divisors p of $M - 1$.*
3. *If $c - 0$ and $M - 2^k \geq 16$, then the period is $\frac{M}{4}$ iff x_0 is odd and $a \mod 8 \in \{3.5\}$.*
4. *If $c = 0$, $M = 2^k \geq 16$ and $a \mod 8 = 5$, then $x_n \mod 4$ is constant =: b, and if $b \subset \{1.3\}$, then $\frac{1}{4}(x_n - b)$ is identical to the sequence produced by the generator with $a' - a$, $c' - b\frac{a-1}{2}$, $M' = \frac{M}{2}$. (This means we should simply ignore the last two bits which are constant anyhow).*

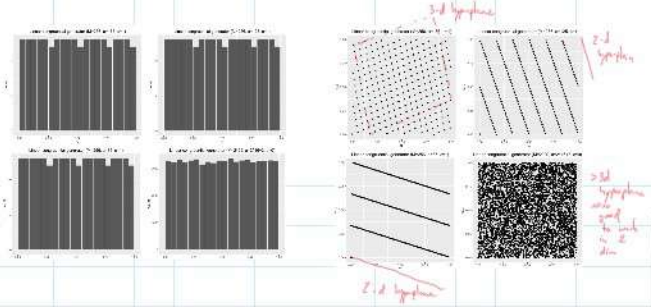
We will see next that despite with suitable choices of a and c the period can be long this does not always guarantee for a good periodicity for all of the applications.

Moreover, especially important for our analysis is the fact that it has limits to the extent it uniformly fits the unit cube.

As we will show next it is not difficult to show that for all of the linear congruential methods, all of the points will lie on equidistant parallel hyperplanes.

↳ This means that there is a natural restriction on how well you can fill the unit cube. You might well not be able to see this in lower dimensions but there is the natural risk in higher dimensions of using such poor generator.

One Example that makes it clear in lower dimensions is the following construct:



Two important solutions for working with linear congruential generators in higher

Two important solutions for working with linear congruential generators in higher dimensions are the following

1. Search a, c, M such that a long period is assured and the distance among hyperplanes is minimized.

One can show that the distance between these hyperplanes depends on the combinations of M and a (and not on c)

The goal is to choose a and M such that the **maximal distance between two neighboring hyperplanes is as small as possible**

In general, there is no analytic formula to calculate this. But there are algorithms to compute this, see Ripley (1987) or Knuth (1998)

The distance can increase drastically if d is increased by one $\Rightarrow$ choosing a is difficult

↳ dimension of hyperplanes

2. Always double check the statistical properties of the generated pseudo-random numbers via statistical tests and check that they pass them.

Proof - Linear Congruential Generated Numbers lie on parallel equidistance hyperplanes

We denote by Λ_d **the set of all d-tuples** produced by the generator

$\Lambda_d = \{(x_n, x_{n+1}, \ldots, x_{n+d-1}),\ 0 \le n < M\} \quad (\cup\{(0, \ldots, 0)\} \text{ if } c = 0)$

↳ the extra period made possible by the generator

↳ 1 tuple for any of Consecutively num tuples

- Λ_d contains for any d maximally M points from the M^d possible points of the set $\{0, \ldots, M-1\}^d \Rightarrow$ With increasing d, the d-tuples are necessarily farther apart from each other

↳ At any draw you could get one value among $\{0, \ldots, M-1\}^d$

- For a good generator, distances should increase in all directions equally

- A **lattice** $L \subset \mathbb{R}^d$ is the set of all integer linear combinations of a set of d linear independent vectors $g_i \in \mathbb{R}^d$:

$$L = \{x = t_1 g_1 + \ldots + t_d g_d;\ t_i \in \mathbb{Z}\}$$

The set of the g_i's are called a basis of L

- One can show that Λ_d consists of all points of a **lattice** shifted by a fixed vector which belong to the the integer cube $\{0, \ldots, M-1\}^d$

↳ This is exactly our theorem that we want to show

Recall support of linear congruential generator $\{0, \ldots, M-1\}$

↳ More formally:

Theorem

Let L_d denote the lattice with generating vectors

$$g_1 = (1, a, a^2, \ldots, a^{d-1})^T,$$
$$g_j = (0, \ldots, \underbrace{M}_{j}, \ldots, 0)^T \quad (j = 2, 3, \ldots, d).$$

If $c > 0$ and the period is equal to M or if $c = 0$ and the period is $M-1$, then

$\Lambda_d = \{c(0, 1, 1 + a, \ldots, (1 + a + \ldots + a^{d-2}))^T + L_d\} \cap \{0, \ldots, M-1\}^d$

↳ intersect ↳ recall that points

If $c > 0$ and the period is equal to M or if $c = 0$ and the period is $M - 1$, then

$$\Lambda_d = \left(c(0, 1, 1 + a, \dots, (1 + a + \dots + a^{d-2}))^T + L_d \right) \cap \{0, \dots, M-1\}^d.$$

shifting vector + lattice; intersect → recall that equals to this set

To prove this equality we will first show the $\subseteq$ direction and then the $\supseteq$ direction which just leaves "=" as the only possibility.

$\subseteq$ direction:

Idea from the very definition of linear congruential method we know

$$x_n = (a x_{n-1} + c) \bmod M$$

it follows automatically by induction that

$$x_{n+j} = \left\{ x_n a^j + c\left(1 + a + a^2 + \dots + a^{j-1}\right) + M \cdot \mathbb{Z} \right\} \cap \{0, \dots, M-1\}^d$$

— = straightforward to see that it implements the modulus operator with intersection, and $\mathbb{Z} \in \{1, \dots\}$

and it is also possible to immediately see that

$$x_{n+j+1} = \left\{ a^{j+1} x_n + c + c\left(a^j + \dots + a\right) + M \cdot \mathbb{Z} \right\} \cap \{0, \dots, M-1\}^d$$

So that it follows immediately given the above that you can rewrite

$$\begin{pmatrix} x_n \\ x_{n+1} \\ \vdots \\ x_{n+d-1} \end{pmatrix} = (a \cdot x_{n-1} + c) \underbrace{\begin{pmatrix} 1 \\ a \\ \vdots \\ a^{d-1} \end{pmatrix}}_{:= g_1} + \sum_{j=2}^{d} t_j \underbrace{\begin{pmatrix} 0 \\ \vdots \\ M_{j\text{-entry}} \\ \vdots \\ 0 \end{pmatrix}}_{:= g_j}, \quad t_j \in \mathbb{Z}$$

$$+\, c \begin{pmatrix} 0 \\ 1 \\ (1+a) \\ (1+a+a^2) \\ \vdots \\ (1+a+a^2+\dots+a^{d-2}) \end{pmatrix}$$

$$\cap \{0, \dots, M-1\}^d$$

|| = Lattice

|| = shifting vector

$\supseteq$ direction:

This follows immediately noticing that $|\Lambda_d| = M$ and that

$t_1 \in \{0, \dots, M-1\}$ } to see this notice that $t_1 = (a \cdot x_{n-1} + c)$ above

such that

$$t_1 g_1 + t_2 g_2 + \dots + t_d \cdot g_d \in \{0, 1, \dots, M-1\}^d +$$
$$+ c \cdot \begin{pmatrix} 0 \\ 1 \\ +a \\ \vdots \\ 1+a+\dots+a^{d-2} \end{pmatrix}$$

shifting vector

exist exactly l

$\Rightarrow \exists!$ $t_2, \dots, t_d$ given linear independent system

it follows immediately that the shifted lattice has maximum dimension M, with $|L_d + \text{shifting vector}| = M$ such that it must be a subset of the d-tuples Λ_d given that this spans the M dimension $\square$

Given such limits of linear congruential methods we now turn to non-linear congruential and other methods to free up from the lattice substructure and fill up the unit cube better.

2. Other Pseudo Random Numbers Generators

- Nonlinear congruential generators
- Shift register generators
- Lagged Fibonacci
- Multiplication with carry-over
- Mersenne-Twister of Matsumoto and Nishimura (1998)

- Mersenne-Twister of Matsumoto and Nishimura (1998)

Recall that again these are considered to be good if they respect:

1. Speed and Reproducibility
2. Statistical Properties
3. Hard Predictability.

Nonlinear Congruential Methods:

$$x_i = f(x_{i-1})$$

where $f : \{0, 1, \dots M-1\} \to \{0, 1, \dots M-1\}$ is nonlinear

- E.g., $f(x) = (ax^2 + bx + c) \mod M$
- These types of generators avoid the lattice structure of the d-tuples, but the computational effort is larger
- Maximal period length is still M
- Difficult to analyze theoretically

Shift Register Generators:

Consider binary sequence $b_n \in \{0, 1\}$ with

$$b_n = b_{n-p} + b_{n-q} \mod 2, \quad 1 \le q < p.$$

if both 1 or 0 $\Rightarrow b_n = 0$ else $b_n = 1$

and set

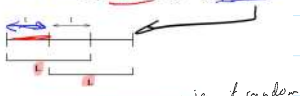
$$u_n = \sum_{j=1}^{L} b_{nt+j} 2^{-j}$$

L-bits in the end

multiple still

limit value range $[0, 1)$.

b_i if $d_k = 1 \Rightarrow$ shift of t



- The value of t controls the overlap of L-tuples
- The maximal period of (b_n) is $2^p - 1$, and there are choices of p and q such that this value of the period is attained

i.e. if random number influenced by the same bit.

Mersenne - Twister:

This is a method proposed by the authors in a paper and is today the standard in many statistical software packages.

- Consider

$$x_k = x_{k-227} + x_{k-623} \begin{pmatrix} 0 & 0 \\ 0 & I_{31} \end{pmatrix} A + x_{k-624} \begin{pmatrix} I_1 & 0 \\ 0 & 0 \end{pmatrix} A$$

where

- $x_k \in \{0, 1\}^{32}$ and $A \in \{0, 1\}^{32 \times 32}$
- All operations are modulo 2
- A can be chosen such that the period is $2^{19937} - 1$ for any non-zero starting value. I.e., the period is larger than the number of atoms in the universe

- Calculate $x'_n = Tx_n$, $T \in \{0, 1\}^{32 \times 32}$, and set u_n:

$$u_n = \sum_{j=1}^{32} (x'_n)_j 2^{-j}$$

- This is the default generator of R
- Not above all doubt as it fails some modern test, is relatively slow, and is relatively easily predictable

$u_n = \sum_{j=1}(x'_n)_j 2^{-j}$

- This is the default generator of R
- Not above all doubt as it fails some modern test, is relatively slow, and is relatively easily predictable

So far we analyzed single possibilities for generating pseudo random numbers. We will see next that through an interesting theorem that it is possible to see that

! Combining two pseudo-number generators we can achieve better statistical distribution properties for the combination.

Moreover notice a second benefit that follows immediately by noticing that when combining two generators the period of the new generator will be larger than the period of each individual generator. This due to the increased combinatorial properties.

Obviously the most basic combined estimator would be a shuffling method, where given two series you would combine the both obtaining a new sequence; the period would obviously be larger, however the extent to which the statistical distribution is affected is hard to compute.

Another more tractable way to combine generators is to use a combining function for the numbers in the two sequences

$X' =$ 1, 2, 3, 1, 2, 3, 1

$X'' =$ 1, 2, 1, 2, 1, 2, 1

$X^c_n = F(X'_n, X''_n)$ I can reason them in terms of X'_i & $X''_i \in [0,1]$

⇒ Immediately to see here that the period is less or equal to the common multiple.

period is less or equal to the common multiple.

We will show now that the combined generator obtained in such a way fills up the unit cube at least as well as each individual method when analyzed individually.

↳ Formally:

- **Distribution of d-tuples of a combined generator**
  - Assume that the seeds for the two generators with values in $\{0, 1, 2, \ldots, M-1\}$ are chosen uniformly and independently. This induces distributions p' and p'' on $W = \{0, \ldots, M-1\}^d$
  - The distribution of the combined generator is

$$p(w) = \sum_{F(w',w'')=w} p'(w')p''(w'')$$

summation over the two. = w

  - One can show that the distribution p of the combined generator is at least as uniform as p' and p'' of the two individual generators

Lemma

Let p' and p'' be two distributions on a finite set W and let F be a function from $W \times W$ to W such that $F(., w)$ and $F(w, .)$ are both bijections for any $w \in W$. Moreover, let p be the distribution on W defined above. Then it holds that

$$\sum_w \left| p(w) - \frac{1}{|W|} \right| \leq \min \left(\sum_w \left| p'(w) - \frac{1}{|W|} \right|, \sum_w \left| p''(w) - \frac{1}{|W|} \right| \right)$$

cardinality of the set W

↳ expresses uniformity

i.e. $F(w,x)$ map is a bijection such that only x such that $y = F(w,x)$ for all w

Proof:

Given the above it is immediate to see that for the transition kernel from state w' to w it holds

$$Q(w', w) := \sum_{w'' : F(w', w'') = w} p''(w'')$$

sum of probabilities for the states where you transition to the distribution of choice.

Given the bijective map $F(w', .)$ we know that just one such w'' exists.

$$\Rightarrow \quad \forall w' : \sum_w Q(w', w) = \sum_w \sum_{w'' : F(w', w'') = w} p''(w'')$$

as 1 for each w

$$= \sum_{w''} p''(w'') = 1 \qquad \text{(I)}$$

the same holds for

$$\forall w : \sum_{w'} Q(w', w) = \sum_{w'} \sum_{w'' : F(w', w'') = w} p'(w'')$$

$$= \sum_{w''} p'(w'') = 1 \qquad \text{(II)}$$

It finally follows that for the the probability of being into one state and transacting into another is given by:

(III) $\forall w: \quad \sum_{w'} p(w')\, Q(w',w) = \sum_{w'} p'(w') \cdot \sum_{w'':\, F(w,w'')=w} p''(w'')$ (Given bijection)

$$= p'(w') \cdot p''(w'') =: p(w)$$ (by def)

It now follows:

$$\sum_w \left| p(w) - \frac{1}{|W|} \right| \overset{II, III}{=} \sum_w \left| \sum_{w'} p'(w')\, Q(w',w) - \frac{1}{|W|} \right| \quad \left(\sum_{w'} Q(w',w) = 1 \text{ by II.}\right)$$

$$= \sum_w \left| \sum_{w'} \left(p'(w') - \frac{1}{|W|} \right) Q(w',w) \right|$$

$$\leq \sum_w \sum_{w'} \left| \left(p'(w') - \frac{1}{|W|} \right) \cdot Q(w',w) \right|$$

$$= \sum_w \sum_{w'} \left| p'(w') - \frac{1}{|W|} \right| \cdot Q(w',w)$$ ($Q(w',w)$ as always positive null)

$$\leq \sum_{w'} \left| p'(w') - \frac{1}{|W|} \right| \sum_w Q(w',w)$$ ($p'(w') - \frac{1}{|W|}$: constant in w; $\sum_w Q(w',w) = 1$)

doing all of the same steps above with $Q(w'',w)$ and $p''(w'')$ would then yield the corresponding inequality for the case, such that our proof is concluded. ∎

We conclude this section by noting that given a pseudo-random number generation you should check its statistical distribution properties through statistical test.

Statistical Tests for Uniform Distribution

- Any test for uniformity on $[0,1]^d$ can be used to test the quality of generators
- For $d = 1$, most generators pass such a test
- The situation is quite different for $d > 1$
- There are collections of tests called `Die Hard` (Marsaglia, 1996) and `TestU01` (L'Ecuyer and Simard, 2007)

Kolmogorov-Smirnov test ($d = 1$)

- The Kolmogorov-Smirnov test statistic is given by

Kolmogorov-Smirnov test ($d = 1$)

- The Kolmogorov-Smirnov test statistic is given by

$$D_n = \sup_x |F_n(x) - F(x)|,$$

where F_n is the empirical CDF of $x_1, \ldots, x_n$ and F the theoretical CDF

  - $F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{(-\infty, x]}(x_i)$
  - $F(x) = x$

- The asymptotic distribution of the test statistic is not a common distribution but it can be approximated and the quantiles have been tabulated

Chisquare test ($d \geq 1$)

- Partition $[0, 1]^d$ into K mutually disjoint classes $I_j, j = 1, \ldots, K$, and count how many d-tuples of consecutive values lies in each class
- The chisquare statistic is given by

$$\sum_{j=1}^{K} \frac{(O_j - E_j)^2}{E_j}$$

where

  - O_j is the number of observed values in I_j
  - E_j is the number of expected values in I_j, $E_j = n|I_j|$ where n is the number of d-tuples considered

- This statistic is approximately χ^2-distributed with $f = K - 1$ degrees of freedom provided the E_k are not "too small" and under the assumption of independence

what represents $|I_j|$?

not clear how much this is respected; I think for instance of the shifted random number generator.

Issues with the chisquare test

- Which partition should one use? Partitioning in subcubes with faces parallel to the coordinate axes quickly leads to too many classes
- Succesive d-tuples are overlapping and therefore not independent

One possible solution ("Monkey tests")

- Instead of counting how often each class occurs, simply count the number W of classes which never occur
- Assume W is approximately normally distributed
- Need to compute the expectation and variance of W (combinatorial problem)

idea: you should visit them all; then based on some normal ground mean of not visited classes given number of sample

↳ compute prob. of observed number of non-visited classes.

Given the proper sampling from uniform distribution we will now turn to techniques used for sampling from non-uniform distributions.

↳ As we will see next there the generation of uniform distributed samples might be of particular benefit.

Chapter 3 - Generating Non-Uniform Random Variables

In this chapter we will talk about some techniques for sampling from non-uniform distributed Random Variables.

In particular we will talk about the following methods

1. Inverse of CDF (generalized quantile function)
2. Relations among distributions (idea: sample from well known and leverage relations among distributions to get to the desired one).
3. Rejection Sampling
4. Importance Sampling
5. Simulation from SDE
6. Variance Reduction Techniques.
7. Quasi-Monte Carlo
8. Markov Chains and Markov Processes

We will explore these techniques in the following in turn:

1. Quantile Transformation

This is the most basic transformation.

The general idea leverages the fact that if the CDF of a distribution $[F(x)]$ is known, then it follows immediately that we can sample

$$X \sim F$$

$X \sim F$

via the following technique:

Step 1: Generate a series of $v_i \overset{i.i.d.}{\sim} U(0,1)$.

Step 2: Set x_i to the value corresponding to the generalized quantile function for the corresponding uniform distribution sample v_i.

I.e.:

$$F^{-1}(v) = \{\inf(x) \mid F(x) \geq v\}$$

Given the definition of this generalized quantile function we can see that the above function is defined even when the CDF function is not bijective. That is, it is easy to prove that:

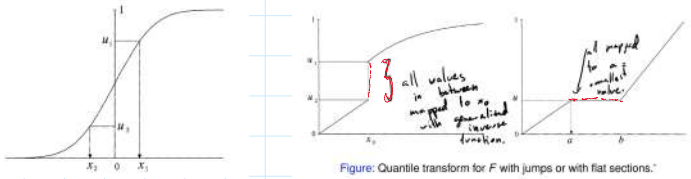


Figure: Quantile transform for F with jumps or with flat sections.[1]

Despite the above (especially for non-injective functions as in figure c) it is possible to prove that the following theorem holds:

Theorem

If $U \sim Uniform(0,1)$, then $X = F^{-1}(U) \sim F$.

Proof:

We know that due to the properties of the generalized quantile transform $F^{-1}(v)$, we have by definition that:

quantile transform $F^{-1}(\cdot)$, we have by definition that:

$$F(F^{-1}(v)) \geq v \qquad F^{-1}(F(x)) \leq x$$

It follows now that given that the:

$$\{x : F(x) \geq v\} \quad \text{and} \quad \{x : F^{-1}(v) \leq x\}$$

sets are equal, such that

$$\underbrace{Pr(X \leq x)}_{\substack{F_X(\cdot) \text{ by} \\ \text{def}}} = Pr(F^{-1}(U) \leq x)$$

$$= \underbrace{Pr(U \leq F(x))}_{\text{dis } F^{-1}(U) \sim F_X(\cdot)}$$

to see this jus chain F^{-1} and F respectively in turn and see how you go from one set to the other.

2. Rejection Sampling

In this section we explore another technique for sampling from non-uniform distributions.

Here in addition from assuming that you are able to generate i.i.d samples from the uniform distribution we also assume that we can sample from an envelope function.

Mathematically this means that given a distribution π (target distribution you aim to sample from) defined on a space $\mathbb{X}$ with a σ-algebra $\mathcal{F}$ with density $f(\cdot)$ w.r.t. to some reference measure μ, and a proposed distribution τ defined over the same space with a density $g(\cdot)$ given the reference measure μ, then you can sample from: i.i.d samples from

$$\boxed{Y \sim \tau}$$

} Need to be able to sample from these.

Theorem:

Theorem:

It is then possible to prove that given that there exist a constant $M < \infty$ such that the distribution $M \cdot g(\cdot)$ of the μ-measured σ-algebra is envelope to $f(\cdot)$ - i.e. $f(x) \leq M \cdot g(x)$ $\forall x$ - then it holds

$$a(x) := \frac{f(x)}{Mg(x)} \leq 1.$$

If Y and U are independent random variables with $Y \sim \tau$ and $U \sim Uniform(0,1)$, then the conditional distribution of Y given $U \leq a(Y)$ is π:

$$\mathbb{P}(Y \in A \mid U \leq a(Y)) = \pi(A) \quad \forall A \in \mathcal{F}.$$

We will prove the theorem next.

Proof:

(i.e. we integrate this's over measure)

Numerator: $\int_A \int_{[0,1]} \mathbb{1}_{\{u \leq a(x)\}} \, g(x) \, \mu(dx) \, du$

$= \int_A a(x) \, g(x) \, \mu(dx) = \frac{1}{M} \int_A \frac{f(x)}{g(x)} g(x) \, \mu(dx)$

$= \frac{1}{M} \int_A f(x) \, \mu(dx) = \frac{1}{M} \pi(x)$ by def.

Denominator: $\int_X \int_{[0,1]} \mathbb{1}_{\{u \leq a(x)\}} \, \mu(dx) \, du$

$= \int_X a(x) \, \mu(dx) \, du$

$= \frac{1}{M} \int \underbrace{\frac{f(x)}{g(x)}}_{=1} \mu(dx) \, du$

So that the ratio $= \pi(x)$ □

In practice it follows therefore:

Algorithm (acceptance-rejection algorithm)

1. *Generate (Y, U) independent with $Y \sim \tau$ and $U \sim Uniform(0,1)$.*
2. *If $U \leq a(Y)$, output $X = Y$, otherwise go back to 1.*

Comments

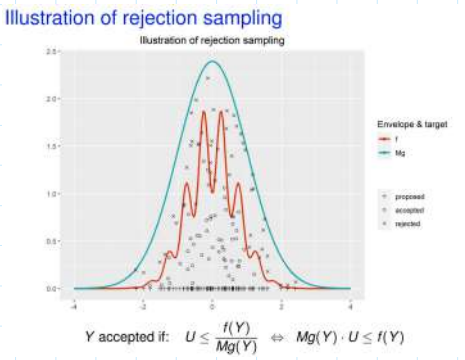
- $a(\cdot)$ is called the **acceptance function**
- $Mg(x)$ is called the **envelope**
- $supp(f) \subset supp(g)$ must hold true in order that there exists $M < \infty$ such that

$$\sup \frac{f(x)}{g(x)} \leq M$$

- $supp(f) \subset supp(g)$ must hold true in order that there exists $M < \infty$ such that

$$\sup \frac{f(x)}{g(x)} \leq M$$

Visual Understanding of Rejection Sampling:



⇒ Such that you would ultimately just sample from the target distribution.

! Important: Notice that the poore, the envelope the more rejections you will have and the higher your computational effort will be in order to obtain the same amount of data/samples.

Notice moreover the following straightforward lemma:

Lemma
Let T denote the number of pairs (Y, U) that have to be generated until $U \leq a(Y)$ for the first time. Then T is geometrically distributed with parameter $1/M$, in particular $\mathbb{E}(T) = M$.

- Geometric distribution: $P(T = k) = (1-p)^{k-1}p$. Probability that the first success (acceptance) occurs after k independent trials
- $p = P(T = 1) = P(U \leq a(Y)) = 1/M$

Idea prob of success after k (means exactly at k) indep trials

$= (1-p)^{k-1} \cdot p$ ($k-1$ failures, p success)

Then it is possible to prove that the above $E(T) = p$.

Then it is possible to prove that the above $E(T) = p$.

$$p = P(T = 1) = P(U \leq a(Y)) = \frac{1}{M} \quad \text{as shown before.}$$

Notice moreover the important property of rejection sampling where we have to know the target distribution just up to proportionality.

▶ If

$$f(x) = \frac{1}{Z} f_*(x) \quad \text{and} \quad f_*(x) \leq Mg(x)$$

and we do not know Z, then the acceptance function is

$$a(x) = \frac{f_*(x)}{Mg(x)}.$$

This is straightforward to see given the previous proof:

Numerator: $\frac{1}{M} \int_A f_*(x)\, \mu(dx)$

Denominator: $\frac{1}{M} \int \frac{f_*(x)}{g(x)}\, \mu(dx)$

Given that $f_*(x)$ is the unnormalized density it holds

$$f_*(x) = Z f(x)$$

Such that in the above

$$\frac{\text{Numerator}}{\text{Denominator}} = \frac{\frac{1}{M} \int f_*(x)\, \mu(dx)}{\frac{1}{M} \cdot Z} = \int f(x)\, \mu(dx)$$

$$= \pi(x) \qquad \square$$

Another interesting case for rejection sampling is the case in which it is difficult / impractical the case of finding a proposal with the above characteristics over the entire space.

↳ In such case it is possible to show that assuming we are able to define partitions of the space

In such case it is possible to show that assuming we are able to define partitions of the space for which it holds:

In many cases, we can find a proposal density by partitioning the space $\mathbb{X}$.

Consider

$$\mathbb{X} = \bigcup_{i=1}^{k} B_i, \quad B_i \cap B_j = \emptyset \ (i \neq j),$$

and suppose that for each B_i we have:

- A density g_i with support B_i
- $f_*(x) \leq M_i g_i(x) \ (x \in B_i)$

Then, we can use

- $g(x) = \sum_{i=1}^{k} \frac{M_i}{M_1 + \cdots + M_k} g_i(x) I_{B_i}(x)$
- $a(x) = \frac{f_*(x)}{\sum_{i=1}^{k} M_i g_i(x) I_{B_i}(x)}$

Notice that the proposal can be different over the different partitions $\Rightarrow g_i$

Proof of the Above:

$$g(x) = \sum_{i=1}^{k} \frac{M_i}{M_1 + \cdots + M_k} g_i(x) \, 1_{B_i}(x)$$

If $x \in B_i$:

$$\frac{f_*(x)}{g(x)} = \frac{f_*(x)}{\frac{M_i}{M_1 + \cdots + M_k} g_i(x)}$$

as disjoint partitions just 1 of these.

$$= \frac{f_*(x)}{M_i \, g_i(x)} \cdot \underbrace{(M_1 + \cdots + M_k)}_{:= M}$$

≤ 1 on each partition by definition

It follows immediately that

$$\frac{f_*}{M \, g_i(x)} \leq 1 \quad \forall B_i$$

□

Algorithmically this would then look as follows:

Algorithmically this would then look as follows:

$$g(x) = \sum_{i=1}^{k} \frac{M_i}{M_1 + \cdots + M_k} g_i(x) I_{B_i}(x)$$

is the density of a **mixture distribution**.

Simulation from this distribution can be done in two steps:

1. Simulate i with probability $\frac{M_i}{M_1 + \cdots + M_k}$
2. Simulate X from the distribution with density $g_i(x)$

for instance you could then compute the following

Example (beta distribution with $\alpha < 1$, $\beta < 1$)

$$f(x) \propto f_*(x) = x^{\alpha-1}(1-x)^{\beta-1} \quad (0 < x < 1).$$

We choose the partition with $B_1 = (0, 0.5)$ and $B_2 = [0.5, 1)$ and the densities

$$g_1(x) = 2^{\alpha}\alpha x^{\alpha-1} \ (x \in B_1), \quad g_2(x) = 2^{\beta}\beta(1-x)^{\beta-1} \ (x \in B_2).$$

Then $M_1 = ?$, $M_2 = ?$

then it is immediate to see that:

$$\frac{f_*(x)}{M_1 \cdot g_1(x)} = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{M_1 \cdot 2^{\alpha}\alpha x^{\alpha-1}} \leq 1$$

$$= \frac{(1-x)^{\beta-1}}{2^{\alpha}\alpha} \leq M_1$$

given that $x \in [0, 0,5]$

$$\frac{0,5^{\beta-1}}{2^{\alpha}\alpha} \leq M_1$$

A final particular important technique happens when the target function has the particular important property of being

log-concave

↳ Then given the fact that the tangent at any point of the function you can leverage such tangent property for approximating the log-concave function at hand.

for approximating the log-concave function at hand.

To see that consider the following:

If f is log-concave and differentiable, rejection sampling can be used to simulate from f as follows:

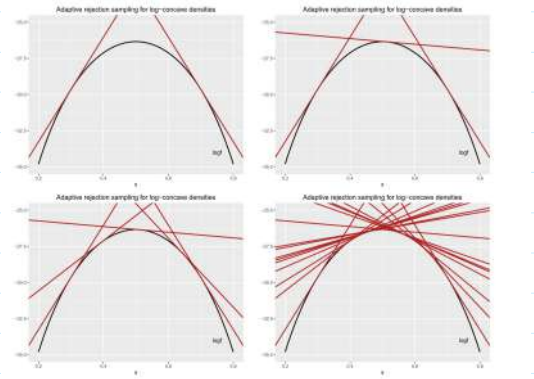
- Choose a partition of the support of f with $B_i = (c_i, c_{i+1}] \subset \mathbb{R}$
- Compute the tangent $a_i + b_i x$ to $\log f$ at some point in B_i ($\Rightarrow \log f(x) \leq a_i + b_i x$)
- Choose $g_i(x) = Z_i^{-1} \exp(b_i x)$ (normalizing constant)
  - We can simulate from g_i with the quantile transform
- Set $M_i = Z_i e^{a_i}$ such that $f(x) \leq M_i g_i(x)$

Such that the original relation holds!

If the intervals are small, this will give a good approximation for f and thus a high acceptance probability

So that in such a case you can easily find a proposal as this.

This is central, you would approximate it very well with smaller partitions, to understand that consider:



Notice that in practice as also computing numerical derivatives costs, you would actually use an adaptive algorithm where you would actually have sampled from a poorly approximated region, i.e.:

The idea can be used in an **adaptive way**:

- Start with 2 intervals
- Whenever a proposal is rejected, compute a new tangent at the proposed value and a corresponding finer partition

3. Relations among Distributions

In this section we will explore some relations among very well known

In this section we will explore some relations among very well known distributions such that it will be possible to sample from them using well known distributions

For instance it is a well know fact that

- If X and Y are independent, X has a standard normal distribution and Y is chi-squared distributed with k degrees of freedom, then $X\sqrt{\frac{k}{Y}}$ has a **t-distribution** with k degrees of freedom
- The **chi-squared distribution** with k degrees of freedom is the distribution of the sum $\sum_{i=1}^{k} Z_i^2$ of k independent squared

Then given that we are able to sample from standard normal distribution by generating i.i.d. uniform samples (as we will see next it is clear that we can generate from all of the above distributions by being able to sample $U_i \overset{i.i.d.}{\sim} U(\cdot,\cdot)$.

Simulating from Standard Normal:

Recall the polar coordinates, i.e.

Let (X, Y) be a two-dimensional random variable. Consider the polar coordinates:

$$R = \sqrt{X^2 + Y^2}, \quad \Phi = \arctan(Y/X)^*.$$

* Logical from the below that the sign will decide on which quadrant the vector will lie.

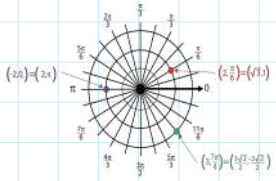
Recall the basic trigonometric properties

so that it is immediate to see that all the below equal

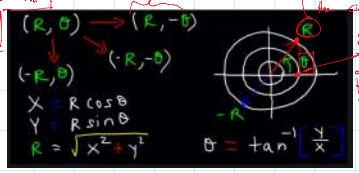
Recall that you should read V as $(r, 360°\theta)$

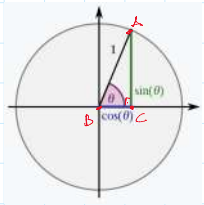
give the address of the circle

gives the angle at which the point lies.



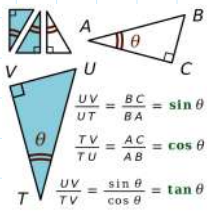
Calcworkshop.com





$\Rightarrow \cos(\theta) = \frac{BC}{BA} \overset{\text{long leg}}{\underset{\text{hypotenuse}}{\}}} \overset{\text{small leg}}{\} } = BC$

$\sin(\theta) = \frac{CA}{BA} \overset{\text{long leg}}{\underset{=1}{\}}} = CA$



$\frac{UV}{UT} = \frac{BC}{BA} = \sin\theta$

$\frac{TV}{TU} = \frac{AC}{AB} = \cos\theta$

$\frac{UV}{TV} = \frac{\sin\theta}{\cos\theta} = \tan\theta$

Immediate to see

$$\frac{Y}{X} = \frac{\sin\theta}{\cos(\theta)} = \tan(\theta)$$

$$\tan^{-1}\left(\frac{Y}{X}\right) = \theta$$

$$\frac{Y}{X} = \frac{\sin\Theta}{\cos(\Theta)} = \tan(\Theta)$$

$$\tan^{-1}\left(\frac{Y}{X}\right) = \Theta$$

Given this short refresher we will now state the following Lemma

Lemma
Let X, Y be i.i.d. $N(0,1)$ distributed. Then R and Φ are independent, Φ is uniform on $(-\pi, \pi)$ and R has the distribution function $1 - e^{-\frac{1}{2}r^2}$.

Such that with the two equations of the polar coordinates and the two unknowns you can sample

$$\Phi \sim U(-\pi, \pi)$$

} like this you sample effectively uniformly from 0 - 360.

and from:

$$R \sim 1 - e^{-1/2\, r^2}$$

and you can then convert them into i.i.d. standard normal samples.

The open question that is left is on how to sample from

$$R \sim 1 - e^{-1/2\, r^2}$$

In order to that we can leverage the Box-Muller Algorithm:

Box-Muller algorithm

- Simulate U, V i.i.d. $\sim$ Uniform$(0,1)$
- Then, $2\pi V - \pi \sim$ Uniform(π, π) and $\sqrt{-2\log(U)} \sim 1 - e^{-\frac{1}{2}r^2}$, independent
- The random variables

$$(X, Y) = \sqrt{-2\log(U)}(\cos(2\pi V), \sin(2\pi V))$$

are i.i.d. $\sim N(0,1)$

It is straightforward to see that

$$\sqrt{2\log(U)} \sim 1 - e^{-1/2\, r^2}$$

$\log(U)$ ... $1-x$

by the quantile transformation seen before

We search

$F(F^{-1}(v)) = x$

$F(x) = 1 - e^{-1/2 x^2}$

$1 - e^{-1/2 F^{-1}(v)^2}$

$\Rightarrow 1 - e^{-1/2 F^{-1}(v)^2} = v$

$1 - e^{-1/2 \sqrt{-2\log U + 2\log(1)}}$

$= 1 - e^{-1/2 \left(-2\log(U) + 2\log(1)\right)}$

$= 1 - e^{\log(U)} + 1 = v$

Notice now that $\log(1) = 0$ such that you can acually express the inverse function as

$F^{-1}(v) = \sqrt{-2\log(v)}$

We will now prove the Lemma above such that we will have a complete proof on how to go from uniform i.i.d. samples to standard normal i.i.d. samples:

I.e. in particular we show that:

Show that if Φ has a uniform distribution on $(-\pi, \pi)$ and R the cumulative distribution function $1 - e^{-\frac{1}{2}r^2}$, then X and Y are i.i.d. $\sim N(0,1)$, when $(X,Y) = \sqrt{-2\log(U)}\,(\cos 2\pi V, \sin 2\pi V)$

To see that understand that:

Φ R

$$P(X \le x', Y \le y') = \int_{-\pi}^{\pi}\int_{0}^{\infty} 1_{\{r\cos(2\pi\phi)\le x', r\cos(2\pi\phi)\le y'\}}(r,\phi)\frac{1}{2\pi} r e^{-\frac{1}{2}r^2} d\phi dr$$

den Φ den R ; immediate by taking derivative of $F(\cdot)$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} 1_{\{r\cos(2\pi\phi)\le x', r\cos(2\pi\phi)\le y'\}}(r,\phi)\frac{1}{2\pi} r e^{-\frac{1}{2}r^2}\frac{1}{r} dx dy$$

$$= \frac{1}{2\pi}\int_{-\infty}^{x'}\int_{-\infty}^{y'} e^{-\frac{1}{2}(x+y)^2} dx dy,$$ standard normal

Noticing then that given independence of

$2\pi V - \pi$ and $\sqrt{-2\log(U)}$ you can see that you can factor the first line

$2\pi U - \pi$ and $\pi - 2\log(U)$ you can see that you can factor the first line and consequently the last one coming to the desired distributions.

It is moreover straightforward to see that given standard normal variables it is possible to sample from a multivariate normal distribution:

Simulation of X according to $\mathcal{N}_p(\mu, \Sigma)$ can be done as follows:

1. Choose matrix A with $\Sigma = AA^T$ (e.g. using Cholesky decomposition)
2. Simulate $Y_1, \ldots, Y_p$ i.i.d. $\sim \mathcal{N}(0,1)$ and set $Y = (Y_1, \ldots, Y_p)^T$
3. Calculate $X = \mu + AY$

- If $\Sigma^{-1} = BB^T$ is given, solve $B^T X = Y$ by backward elimination and add μ

Finally notice that in a similar way as seen that far you can sample from a poisson distribution through the following Lemma and the following algorithm

Lemma

Let (X_i) be i.i.d. $\sim Exp(1)$ and $S_n = \sum_{i=1}^{n} X_i$ with $S_0 = 0$. Then

$$S_n \sim Gamma(n, 1)$$

number of events

gamma = prob k-events up until time t.

and

$$\mathbb{P}(S_n \le t < S_{n+1}) = e^{-t}\frac{t^n}{n!}$$

Interpretation

- Assume X_i is the time between arrivals of customers in a queueing system
  $\Rightarrow S_n$ is the arrival time of the n-th customer
- $S_n \le t < S_{n+1}$ means that the number of customers that have arrived up to time t is equal to n. This number has a Poisson distribution with parameter t

Algorithm:

1. Simulate U_i i.i.d. Uniform(0, 1)
   - Then $X_i = -\log(U_i) \sim \text{Exp}(1)$
2. Set

$$Y = \min\{k \mid \sum_{i=1}^{k}(-\log(U_i)) > t\} - 1$$
$$= \min\{k \mid U_1 \cdot U_2 \cdots U_k < e^{-t}\} - 1$$

- Y is Poisson(t) distributed since

$$\sum_{i=1}^{n} X_i \le t < \sum_{i=1}^{n+1} X_i \Leftrightarrow \min\{k \mid \sum_{i=1}^{k} X_i > t\} = n+1$$

such that you would have in fact have sampled from $S_n \le t < S_{n+1}$

We will prove the Lemma now:

We do this by induction; i.e.

we assume $S_n = Gamma(n, 1)$ and $X_{n+1} \sim exp(1)$

and we will prove that $S_{n+1} = Gamma(n+1, 1)$.

It follows then given the prove that

It follows then given the prove that this follows as $S_1 = Exp(1) = Gamma(1,1)$.

Proof.

$$
\begin{aligned}
f_{S_{n+1}}(z) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{x^{n-1}}{\Gamma(n)} e^{-x} e^{-y} 1_{\{x\geq 0\}} 1_{\{y\geq 0\}} 1_{\{x+y=z\}} dxdy \\
&= \int_{-\infty}^{\infty} \frac{x^{n-1}}{\Gamma(n)} e^{-x} e^{-(z-x)} 1_{\{x\geq 0\}} 1_{\{z-x\geq 0\}} dx \\
&= \frac{1}{\Gamma(n)} e^{-z} \int_0^z x^{n-1} dx \\
&= \frac{1}{\Gamma(n)} e^{-z} \frac{z^n}{n} = \frac{1}{\Gamma(n+1)} z^n e^{-z}
\end{aligned}
$$

□

Given the above proof & the fact that

$S_n \sim Gamma(n,1)$

$S_{n+1} \sim Gamma(n+1,1)$

We can now prove that:

$$P(S_n \leq t < S_{n+1}) = P(S_n \leq t) - P(S_{n+1} < t)$$

$$= \int_0^t e^{-x} \left(x^{n-1} \frac{1}{\Gamma(n)} - x^n \frac{1}{\Gamma(n+1)} \right)$$

$$= \left[e^{-x} \frac{x^n}{n!} \right]_0^t = e^{-t} \frac{t^n}{n!}$$?

This concludes the section how to sample from very well known distributions using Uniformly i.i.d. samples.

In the next section we turn to Importance Sampling

4. Importance Sampling

This is another important technique in

This is another important technique in the space of stochastic simulation.

It is similar to the rejection sampling technique in the sense that we leverage a proposal distribution (well known) in order to sample from the target distribution (difficult if not impossible to integrate over it)

↳ However in contrast to rejection sampling we do not reject samples but reweight each sample such that it would actually adjust to the targeted distribution.

- **Goal**: calculate

$$\theta = \mathbb{E}_\pi(h(X)) = \int_{\mathbb{X}} h(x)f(x)\mu(dx)$$

- Assume that τ is a distribution on $(\mathbb{X}, \mathcal{F})$ with density g, $supp(f \cdot h) \subset supp(g)$, and denote

$$w(x) = \frac{f(x)}{g(x)}$$

- Importance sampling is based on a similar idea as rejection sampling. Instead of rejecting some variables, we **weight the samples with a weighting function w**

should this not be the case if is straightforward to understand that you will never sample from some areas of the distribution → bad

Consider for instance the following:

$$\hat{\theta} = \frac{1}{N} \sum_i h(Y_i)\, w(Y_i)$$

with $Y \sim \tau$; i.e. distributed according to the proposal.

It is then straightforward to see that the above is unbiased, meaning:

$$E_\tau(h(Y)\, w(Y)) = \int h(y) \frac{f(y)}{g(y)} g(y)\, \mu(dy)$$

$$= \int h(y) f(y)\, \mu(dy)$$

$$= \int h(y)\, \pi(dy)$$

$$= E_\pi(h(Y))$$

and notice

$$= E_{\pi}(h(Y))$$

↳ and notice that as $Y \in \mathbb{X}$

It now follows that

$$E(\hat{\theta}) = E\left(\frac{1}{N}\sum h(Y_i)w(Y_i)\right)$$

$$= \frac{1}{N}\sum E_{\pi}(h(Y)) = \frac{1}{N} \cdot N \, E_{\pi}(h(Y))$$

Important: Notice that in contrast to rejection sampling, importance sampling does not require

$$w(x) = \frac{P(x)}{g(x)}$$

to be bounded!

↳ It is therefore more general in this sense

Notice moreover that a second important property is that the importance sample is not linear

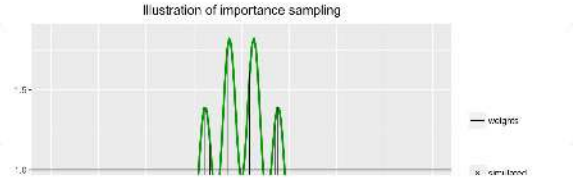
To see that consider for instance the estimator above:

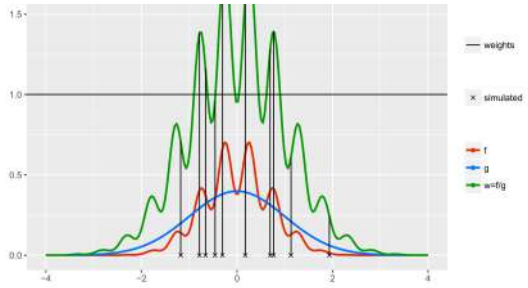
$$\hat{\theta} = \frac{1}{N}\sum (h(Y)+c)\, w(Y)$$

$$= \frac{1}{N}\sum h(Y)w(Y) + \frac{1}{N}\sum c\, w(Y)$$

$\neq c$ as $\frac{1}{N}\sum w(Y) \neq 1$

A graphical understanding of the distribution can be gained from the following:





Notice, moreover, the following important fact.

Even if you do not have to impose any conditions on $\frac{f(x)}{g(x)}$ in practice you would do so as otherwise for instance you have no guarantee that your importance sampling estimator fulfills common understanding properties.

This leads to the following:

- It is desirable that the estimator $\tilde{\theta}$ has **finite variance**. This leads to **conditions for f and g**
- For instance, in order to have **finite variance for all bounded functions h**,

$$\int f(x)^2/g(x)\mu(dx)$$

must be finite

- Weaker condition than $\frac{f(x)}{g(x)} \leq M < \infty$

so that you have to introduce some conditions in a similar setting to the one of rejection sampling

due to the squared factor.

The proof of this is no magic, it leverages standard variance properties.

↳ Proof:

$$\mathrm{Var}\left(\frac{1}{N}\sum h(Y)\,w(Y)\right) \overset{\text{due to independence}}{=} \frac{1}{N^2}\sum \mathrm{Var}\left(h(Y)\,w(Y)\right)$$

$$= \frac{1}{N^2}\, N\, \mathrm{Var}\left(h(Y)\,w(Y)\right)$$

through basic variance rules

$$= \frac{1}{N}\left(E\left(\left(h(Y)\,w(Y)\right)^2\right) - E\left(h(Y)\,w(Y)\right)^2\right)$$

$E\left(h(Y)\,w(Y)\right) = \theta^2$ as it is unbiased as previously shown

$\Rightarrow$ It follows now that in order to guarantee that the variance

⟹ It follows now that in order to guarantee that the variance is bounded we must guarantee that:

cas it is unbiased as previously shown

$$E\left[\left(h(Y)\, w(Y)\right)^2\right] < \infty$$

$$E\left[\left(h(Y)\, w(Y)\right)^2\right] = \int h(y)^2\, w(y)^2\, g(y)\, \mu(dy)$$

$$= \int h(y)^2 \frac{f(y)^2}{g(y)^2}\, g(y)\, \mu(dy)$$

It follows immediately that

$$\mathrm{Var}\left(\frac{1}{N}\sum h(Y)\, w(Y)\right) < \infty$$

- when $h(.)$ is bounded
- and $\frac{f(y)^2}{g(y)}\mu(dy) < \infty$

□

Should you not be satisfied with bounded variance the topic becomes complex.

We will see at a later stage that closer $g(.)$ to $f(.)$ the smaller the variance of the estimator will be.

⟺ In high dimensions it is difficult to assess and guarantee this property

↳ Use of importance sampling is limited in practice in high dimensions.

Notice now finally that a slightly modified importance sampling estimator for the mean of the targeted distribution is possible:

There is an **alternative version** of importance sampling:

$$\tilde{\theta}' = \frac{\frac{1}{N}\sum_{i=1}^{N} h(Y_i) w(Y_i)}{\frac{1}{N}\sum_{i=1}^{N} w(Y_i)}$$

Comments

- This estimator is linear, no longer unbiased, but consistent

Comments

- This estimator is linear, no longer unbiased, but consistent
- In contrast to the first version, this version can also be used if f is known only up to a normalizing constant

Giving this it follows quite straighly that in the case you would know the target up to proportionality and in the case a sufficiently large sample can be generated, then this measure is preferred.

Proof of the above Properties:

$$\frac{1}{N}\sum w(Y) \xrightarrow{a.s.} \int w(y)\, g(y)\, \mu(dy)$$

$f(y)$ or $f_*(y)$ if unnormalized target known

Continuing with $w(Y) = \frac{f_*(y)}{g(y)}$ with $f_*(y) = Z f(y)$

we see that the above:

$$\frac{1}{N}\sum w(Y) \xrightarrow{a.s.} Z$$

! Notice the interesting property here $\frac{1}{N}\sum w(Y)$ estimates for Z.

and that

$$\frac{1}{N}\sum w(Y)\, h(Y)\, g(y)\, \mu(dy) \xrightarrow{a.s.} Z \cdot 0$$

such that due to the continuous mapping theorem the

$$\tilde{\theta}' \xrightarrow{a.s.} 0$$

A final version of importance sampling does not work through reweighting directly sampled observations but rather you sample twice one from the original distribution and one time from the sample of it.

↳ Where this second time you sample according to the importance scemes:

→ Where this second time you sample according to the importance scemes: $\frac{w(Y_i)}{\sum w(Y_i)}$.

This gives finally rise to the following:

Sampling Importance Resampling

If we want an **unweighted sample** instead of a weighted one, we can use resampling:

1. Sample Y_i i.i.d. $\sim \tau$, $i = 1, \ldots, N$
2. We generate additional variables I_i which take values in $\{1, 2, \ldots, N\}$ with probabilities proportional to the weights $w(Y_j)$:

$$\mathbb{P}(I_i = j) = \frac{w(Y_j)}{\sum_{j'=1}^{N} w(Y_{j'})}$$

($\mathbb{P}(I_i = j)$: probability of selecting the jth; then resample based on this probability.)

3. Set $X_i = Y_{I_i}$.

→ We will show now that the samples generated in such a way are i.i.d., however it is important once more to understand that the above estimator has a higher variance in comparison to a plain vanilla IS, this due to the additional resampling step

Proof - i.i.d. π of Sampling Importance Resampling

First for the estimator $\tilde{\theta} = \frac{1}{N} \sum h(X) \xrightarrow{a.s.} \theta$

To see that observe that for SIR

$$P(I_i = j) = \frac{w(j)}{\sum_{k=1}^{N} w(k)}$$

Such that

$$P_r\left(\frac{1}{N} \sum h(X_i) \mid Y_1, \ldots, Y_N\right) = \frac{1}{N} \sum P_r\left(h(X_i) \mid Y_1, \ldots, Y_N\right)$$

Such that

$$P_r\left(h(X_i) \mid Y_1, \ldots, Y_N\right) = \sum_{j=1}^{N} P_r(I_i = j) \cdot h(X_i)$$

(as there might be more than 1 Y_k taking this value you need to take the sum.)

$$= \frac{\sum_{i=1}^{N} w(i) \cdot h(X_i)}{\sum_{k=1}^{N} w(k)}$$

$$\sum w(i)\, h(X_i) \quad a.s. \quad \theta$$

$$= \frac{\sum_{k=1}^{N} w(i)\, h(X_i)}{\sum_{k=1}^{N} w(k)} \xrightarrow{a.s.} \frac{0}{1}$$

to see that think that it converges to

$$\int \frac{f(x)}{g(x)} h(x) \cdot g(x)\, \mu(dx) = E_{\pi}(h(x))$$

It follows now that taking the sum over all $h(X_i)$ you get to the desired result considering the $\frac{1}{N}$ factor.

(5) Simulation from Stochastic Differential Equation.

A **stochastic differential equation** (SDE) arises from an ordinary differential equation by adding a stochastic noise:

$$\frac{dX_t}{dt} = f(X_t) + \sigma(X_t) N_t$$

where N_t is white noise: N_t and N_s are independent for $t \neq s$.

- **White noise N_t** cannot be defined as a regular stochastic process but only as a generalized process*

This can be seen as N_t violates the continuity property of stochastic processes.

↳ To see this consider the following:

N_t and $N_{t'}$ are independent white noise series $\forall\ t \neq t'$ with $E(N_t) = 0$ and variance c.

It follows now that

$$\mathrm{Var}\left(\int_t^{t+h} N_t\right) = c \cdot h \quad \left\} \text{ due to the independence}\right.$$

Given now that

$$\mathrm{Var}\left(\int_t^{t+h} N_t\right) = \int_t^{t+h} N_t^2\, dt + \cancel{E\left(\int_t^{t+h} N_t\right)^2}^{=0}$$

It now follows that if the

$$\mathrm{Var}\left(\int_t^{t+h} N_t\right) = c\,h$$

you must have some

$$\int_t^{t+h} N_t^2\, dt > ch$$

$$\int_t^{t+h} N_t^2\,dt > ch$$

$$\Longleftrightarrow$$

$$\int_t^{t+h} N_t\,dt > \underbrace{\sqrt{c}}_{:=\varepsilon}\sqrt{h}$$

It now follows that for a sequence $h_1, h_2, \ldots$ with $h_i = 0$

$$\int_t^{t+h_i} N_t\,dt > \varepsilon\sqrt{h_i}$$

$\cdot \frac{1}{h_i}$

$$\frac{1}{h_i}\int_t^{t+h_i} N_t\,dt > \frac{\varepsilon}{\sqrt{h_i}}$$

$$\frac{1}{h_i}\int_t^{t+h_i} N_t\,dt > \infty$$ ↯ violates mean value theorem, hence the process is nowhere continuous.

Mean value theorem

$$\frac{1}{h_i}\int_t^{t+h_i} N_t\,dt \to N_t$$

What we can do however with the above is to write it in the integral form:

$$X_t = X_0 + \int_0^t f(X_s)ds + \int_0^t \sigma(X_s)dB_s.$$

Where $B_t = \int_0^t N_t$ is a so called brownian motion / wiener process, with the following properties:

Brownian motion is a stochastic process with:

1. $B_0 = 0$ almost surely
2. For all $t_0 = 0 < t_1 < t_2 < \cdots < t_n$, the increments $B_{t_i} - B_{t_{i-1}}$ $(i = 1,\ldots n)$ are independent and $\mathcal{N}(0, t_i - t_{i-1})$-distributed

- Wiener has shown that there is such a process and it can be chosen so that the paths are almost surely continuous

} we do not prove the important result here

As a final step, in order to solve the above stochastic differential equation, we need to precisely define what $\int$ means to integrate a random variable over a brownian motion

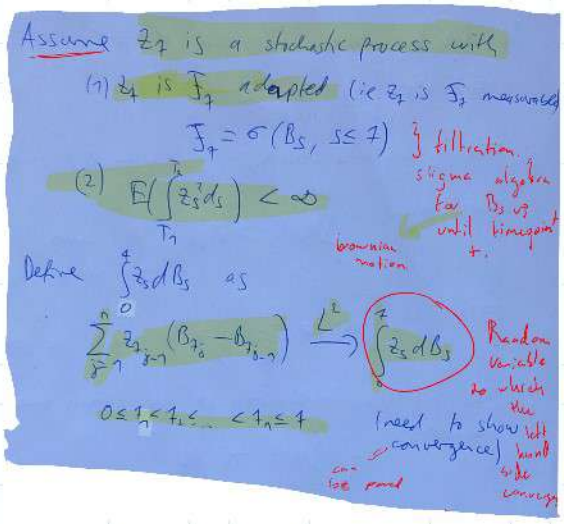
↳ One possible definition is using Ito's Integral:

- Wiener has shown that there is such a process and it can be chosen so that the paths are almost surely continuous

} not prove the important result here

As a final step, in order to solve the above stochastic differential equation, we need to precisely define what $\int$ means to integrate a random variable over a brownian motion

↳ One possible definition is using Ito's integral:



Given this definition it is clear that a finite sample approximation for the integral up to t is given by

$$\sum_{j=1}^{n} Z_{t_{j-1}} \left(B_{t_j} - B_{t_{j-1}} \right)$$

where $0 < t_1 < \dots < t_n < t$

We will use this for simulation.

Given this definition it then possible to rigorously define what a solution to the above SDE is.

For instance for finding a solution you would start with the above SDE in integral form and proceed iteratively

$$X_t^{(m)} = X_0 + \int_0^t f(X_s^{(m-1)})\, ds + \int_0^t \sigma(X_s^{(m-1)})\, dB_s$$

and you can then actually prove convergence when $m \to \infty$

Given all of these and the definition of Ito's integral you can actually use the following euler scheme for simulating

of Ito's integral you can actually use the following euler scheme for simulating from stochastic differential equations as the above:

Euler scheme

Generate approximate solution at time points $k\Delta$, $k = 1 \ldots, K$, $\Delta > 0$, according to

$$X_{(k-1)\Delta} = X_{k\Delta} + f(X_{k\Delta})\Delta + \sigma(X_{k\Delta})\underbrace{(B_{(k-1)\Delta} - B_{k\Delta})}_{\sim \mathcal{N}(0,\Delta)}.$$

Algorithm

1. Simulate Z_k i.i.d. $\sim \mathcal{N}(0, \Delta)$ and X_0
2. Calculate $X_{(k+1)\Delta} - X_{k\Delta} + f(X_{k\Delta})\Delta + \sigma(X_{k\Delta})Z_k, \quad k - 1 \ldots, K$

We conclude with the following Lemma that gives an intuition for the proof of the reason why

$$B_t = \int_0^t N_t$$

and shows how theoretically generate a brownian motion

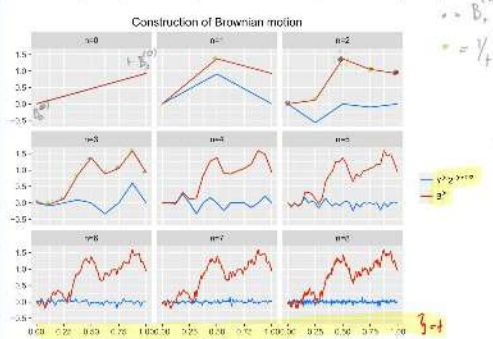
Lemma

Let $(B_1, X_{n,j}, n = 1, 2, \ldots, j = 1, 2, \ldots, 2^{n-1})$ be i.i.d. standard normal random variables and for each n let $(Y_i^{(n)})$ be the continuous process on $[0, 1]$ which is equal to $X_{n,j}$ for $t = (2j - 1)2^{-n}$, equal to zero for $t = 0$ and $t = 2j2^{-n}$, and which interpolates linearly in between. Then the sequence $B_t^{(0)} - tB_1$,

$$B_t^{(n)} = B_t^{(n-1)} + Y_t^{(n)} 2^{-(n+1)/2},$$

converges with probability one uniformly in t to a Brownian motion on $[0, 1]$ with continuous sample paths.

↳ Grafical rapresentation of using the above algo



Notice then that it follows immediately due to the independence of B_1 and $X_{n,j}$, that

$$B_t^{(n)} = \sum \text{i.i.d. } N(0,1)$$

such that it has the property of

D_t ~ ⟵ i.i.d. $N(0,1)$

such that it has the property of being $N(0, x)$ where x = # i.i.d. variables defined above, and moreover it holds

$$Y_t^{(n)} \; 2^{-\frac{(n+1)}{2}} = B_t^{(n)} - B_t^{(n-1)}$$

$B_t^{(n)} - B_t^{(n-1)} \sim N(0, x_n - x_{n-1})$ from brownian motion def.

then probably $2^{-(n+1)} = x_n - x_{n-1}$, would have to verify it though.

This section briefly introduced SDE and showed that after properly specifying them and analyzing them you might approximate them simulating from standard well known distributions.

↳ This is also interesting cause when studying SDE in finance for the understanding of them this is actually what you try to do: i.e. understand first some underlying atomic component of the system and its distribution and then you generalize it.

(6) Variance Reduction

So far we have seen how to compute i.i.d. samples of $h(X)$ and how to compute:

$$\hat{\sigma} = \frac{1}{N} \sum h(X_i)$$

This was basically the idea of rejection sampling, or sampling directly from the target distribution via quantile transformation or the relation to well known distributions.

↳ Given the Monte-Carlo properties we then derived at first the convergence properties for such

↳ Given the Monte-Carlo properties we then derived at first the convergence properties for such series.

↳ In this section we explore techniques through which it is possible to gain in precision (i.e. reduce variance keeping the estimator consistent) in comparison to standard Monte-Carlo Methods.

Notice, that we excluded importance sampling from the above, there we just proved directly that

$$\tilde{\theta} = \frac{1}{N} \sum h(Y) w(Y) \qquad Y \sim \tilde{\pi}$$

is an unbiased estimator for

$$E_{\pi}(h(X))$$

however the approach is fundamentally different, as we do not sample $h(X)$ directly at any point.

↳ We will see towards the end of the section some ideas on how to keep the variance of such estimator low.

We will explore now two techniques for reducing the variance in monte-carlo estimators.

1. Antithetic Variables
2. Control Variates

Antithetic Variables

The idea of this section is the following:

Antithetic Variables

The idea of this section is the following:

- We want to estimate the mean $\theta = E(X)$ of some random variable X with distribution π. Assume X_1 and X_2 have both the distribtion π
- If X_1 and X_2 are independent, then

$$\mathrm{Var}\left(\frac{1}{2}(X_1 + X_2)\right) = \frac{1}{2}\mathrm{Var}(X).$$

- If X_1 and X_2 are not independent, then

$$\mathrm{Var}\left(\frac{1}{2}(X_1 + X_2)\right) = \frac{1}{2}\left(\mathrm{Var}(X) + \mathrm{Cov}(X_1, X_2)\right).$$

$\Rightarrow$ **if X_1 and X_2 are negatively correlated, the variance of $\frac{1}{2}(X_1 + X_2)$ can be reduced.**

The question of the section is therefore on how to generate pairs of $X_{i,j}$ that are negatively correlated among themselves but i.i.d. across each other.

To understand how will get there consider the following definition and Lemma:

- We consider the situation

$$\theta = \int_0^1 h(x)dx = \mathbb{E}(h(U)), \; U \sim \text{Uniform (0,1)}$$

- Instead of $\hat{\theta}_N = \frac{1}{N}\sum_{i=1}^N h(U_i)$, U_i i.i.d. $\sim Unif(0,1)$, we use

$$\tilde{\theta}_N = \frac{1}{2N}\sum_{i=1}^N h(U_i) - h(1-U_i)$$

and obtain

$$\mathrm{Var}(\tilde{\theta}_N) = \frac{\mathrm{Var}(h(U_i)) + \mathrm{Cov}(h(U_i), h(1-U_i))}{2N}$$

this follows immediately by noting $\mathrm{Var}(h(U_i)) = \mathrm{Var}(h(1-U_i))$ since $1-U_i \sim U(0,1)$

- If $\mathrm{Cov}(h(U_i), h(1-U_i)) < 0$, $\tilde{\theta}_N$ has a smaller variance than $\hat{\theta}_{2N}$.

↳ It follows immediately by noting that:

$$\mathrm{Var}\left(\hat{\theta}_{2N}\right) = \frac{1}{4N^2}\sum \mathrm{Var}\left(h(U_i)\right)$$

$$= \frac{2N\,\mathrm{Var}(h(U_i))}{4N^2}$$

So we have to prove that:

$$\frac{\mathrm{Var}(h(U_i))}{2N} > \frac{\mathrm{Var}(h(U_i)) + \mathrm{Cov}(h(U_i), 1-h(U_i))}{2N}$$

$2N$ $2N$

$$0 > \mathrm{Cov}(h(U_i), h(1-U_i))$$

It follows now that

Lemma
If the function h is monotone, then $Cov(h(U), h(1-U)) < 0$, *unless h is constant on* $(0,1)$.

Proof:

To prove the above you want to get an expression for the covariance $\mathrm{Cov}(h(U), h(1-U))$ that is just dependent on monotonic terms.

↳ This can be obtained as follows:

Think of two variables U_1 and U_2 both $\sim \mathrm{Unif}(0,1)$, with h holds $E(h(U)) = \theta$, then it holds

$$E\left[\left((h(U_1)-\theta) + (h(U_2)-\theta)\right)\left((h(1-U_1)-\theta) - (h(1-U_2)-\theta)\right)\right]$$

$$= E\left[(h(U_1)-\theta)\cdot(h(1-U_1)-\theta)\right] - E(h(U_1)-\theta)\,E(h(1-U_2)-\theta)$$

$$- E(h(U_2)-\theta)\,E(h(1-U_1)-\theta) + E\left((h(U_2)-\theta)(h(1-U_2)-\theta)\right)$$

($E(h(U_1)-\theta) \neq 0$, $E(h(1-U_2)-\theta) \neq 0$, $E(h(U_2)-\theta) = 0$, $E(h(1-U_1)-\theta) = 0$)

Note: due to independence you can take the expectation for each term individually

$$= 2\,\mathrm{Cov}(h(U), h(1-U))$$

Notice therefore that with the yellow equation

$$\mathrm{Cov}(h(U), h(1-U)) = \frac{1}{2} E\left[\left(h(U_1) - h(U_2)\right)\left(h(1-U_1) - h(1-U_2)\right)\right]$$

Assume h monotone as per Lemma then two cases:

(i) $U_1 > U_2$ $\quad h(U_1) - h(U_2) > 0$ $\quad h(1-U_1) - h(1-U_2) < 0$

(ii) $U_2 > U_1$ $\quad h(U_1) - h(U_2) < 0$ $\quad h(1-U_1) - h(1-U_2) > 0$

So you see that in both cases the Cov is negative. □

So you see that in both cases the Cov is negative. □

From this it follows the following simple mechanism to reduce variance the most:

If ***h* is monotone** and we can simulate from F with the quantile transformation, then the following algorithm can be used:

1. Simulate U_i
2. Approximate $\int h(x)F(dx)$ by

$$\frac{1}{2N}\sum_{i=1}^{N} h(F^{-1}(U_i)) + h(F^{-1}(1-U_i))$$

- $h(F^{-1})$ is monotone since both h and F^{-1} are monotone

⇒ Notice that the above technique can be used directly for the quantile transformation method.

↳ For rejection sampling, you could first compute the i.i.d. X_i samples and then compute $\hat{F}_n$ from it and apply the above method with it.

↳ The extent to which the variance is affected by the resampling step was not treated in the lecture and it is up to you to investigate it further if you have time.

Control Variates

This section relies on the identification of a function $r(X_i)$, with the property that $E(r(X_i))$ is known, such that w.l.o.g. we can assume $E(r(X_i)) = 0$.

Moreover, to obtain a solid reduction in variance it must hold:

$$\boxed{\text{corr}(r(X_i), h(X_i)) > 0}$$

Then it is possible to construct a new estimator which is unbiased via

Then it is possible to construct a new estimator which is unbiased via
↳ trivial to see due LLN.

$$\tilde{\theta} = \frac{1}{N}\sum h(X_i) - c \cdot r(X_i)$$

It follows that

$$Var(\tilde{\theta}) = \frac{1}{N^2} N \left(Var(X_i) - 2Cov(h(X_i), r(X_i)) + c^2 Var(r(X_i)) \right)$$

it follows that choosing an optimal scaling factor c, you would have:

$$\frac{\partial Var(\tilde{\theta})}{\partial c} = 2c\, Var(r(X_i)) - 2Cov(h(X_i), r(X_i)) = 0$$

$$\Leftrightarrow \quad c_{opt} = \frac{Cov(h(X_i), r(X_i))}{Var(r(X_i))}$$

i.e.

$$Var(\tilde{\theta}) = \frac{1}{N}\left(Var(h(X_i)) - \frac{Cov(h(X_i), r(X_i))^2}{Var(h(X_i))} \right)$$

$$= \frac{1}{N}\left(Var(h(X_i))\left(1 - Corr(h(X_i), r(X_i))^2\right) \right)$$

such that the variance could be reduced by a $Corr(h(X_i), r(X_i))^2$ term, i.e. by a positive factor if $h(X_i)$ and $r(X_i)$ have a linear relation

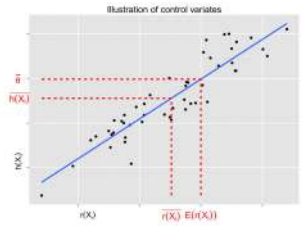
⇒ Important is however to notice that c_{opt} depends on RV and it is not known a priori.

↳ Notice however that we can estimate it consistently via the usual linear model:

$$\hat{c}_{opt} = \frac{\sum_{i=1}^{N}(h(X_i) - \hat{\theta}_N) r(X_i)}{\sum_{i=1}^{N} r(X_i)^2}.$$

- **Interpretation**: $\hat{c}_{opt}$ is the slope of the least-squares regression line passing through the points $(r(X_i), h(X_i))$

Illustration of control variates



- In the example, the sample mean $\overline{r(X_i)} < 0$ underestimates $\mathbb{E}(r(X_i)) = 0$
- Since $r(X_i)$ and $h(X_i)$ are correlated, this suggests that the sample mean $\overline{h(X_i)} = \hat{\theta}_N$ will likely also underestimate $\theta = \mathbb{E}(h(X_i))$
- The sample mean $\overline{h(X_i)} = \hat{\theta}_N$ is thus adjusted upward to obtain $\bar{\theta}_N$

Notice that, if all of the $r(X_i) > 0$, then you might also use the following estimator

one scales r such that

$$\mathbb{E}(r(X_i)) = 1$$

and can use the following **multiplicative correction**:

$$\bar{\theta}_N = \frac{\frac{1}{N}\sum_{i=1}^{N} h(X_i)}{\frac{1}{N}\sum_{i=1}^{N} r(X_i)}$$

- One cannot calculate the expected value and variance of $\bar{\theta}_N$ exactly
- However, the following holds true:
  - $\bar{\theta}_N$ converges almost surely to θ
  - $\sqrt{N}(\bar{\theta}_N - \theta)$ is asymptotically normal with mean zero and variance $\mathrm{Var}(h) - 2\theta\mathrm{Cov}(h, r) + \theta^2\mathrm{Var}(r)$

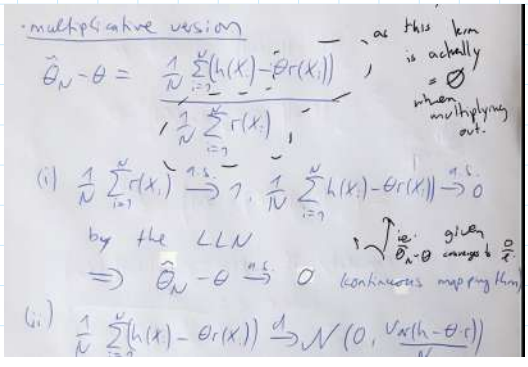
Recall that you can rewrite this as:

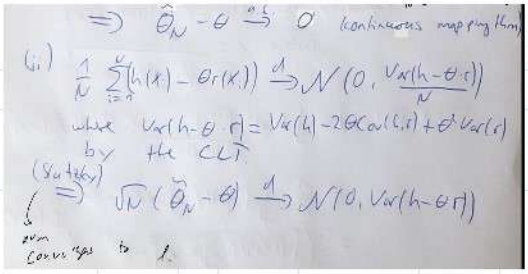
$$\mathrm{Var}(h(X_i)) + \mathrm{Var}(r(X_i))\left(\theta^2 - 2\theta\,\mathrm{Corr}(h, r)\right)$$

$$= \mathrm{Var}(h(X_i)) + \mathrm{Var}(r(X_i))\theta\left(\theta - 2\,\mathrm{Corr}(h, r)\right)$$

such that if < 0 then the variance is reduced.

Proof:





We now turn to methods for reducing the variance of Importance Sampling as mentioned before.

Importance Sampling & Variance Reduction

As a review recall:

- **Goal**: estimate
$$\theta = \int h(x) f(x) \mu(dx),$$
where f is the density of the target distribution π
- **Approach of IS**:
  1. Simulate Y_i i.i.d. from a "wrong" distribution τ with density g
  2. Correct for this by computing the weighted average
$$\tilde{\theta}_N = \frac{1}{N}\sum_{i=1}^{N} h(Y_i) w(Y_i),$$
where
$$w(x) = \frac{f(x)}{g(x)}$$
- IS cannot only be used in situations where it is impossible or difficult to simulate from π
- IS can also be used to obtain an estimate with **lower variance** than
$$\hat{\theta}_N = \frac{1}{N}\sum_{i=1}^{N} h(X_i), \quad X_i \text{ i.i.d. } \sim \pi$$
} standard approach, sampling directly from the target dist.

The idea to lower the variance is the following, first noticing that both estimators are unbiased, it holds that

$$E(\tilde{\theta}_N) = E(\hat{\theta}_N)$$

then given that it is possible to rewrite the variance of the two estimators as follows

rewrite the variance of the two estimators as follows

$$\mathrm{Var}\left(\frac{1}{N}\sum h(x_i)\right) = \frac{1}{N^2} \cdot N \,\mathrm{Var}(h(x))$$

$$N\,\mathrm{Var}(\hat{\theta}_N) = \int h(x)^2 \pi(dx) - \theta^2$$

And for the estimator $\tilde{\theta}_N$ it holds

$$\mathrm{Var}\left(\frac{1}{N}\sum h(x_i)\, w(x_i)\right) = \frac{1}{N^2} N \,\mathrm{Var}(h(x)\, w(x))$$

$$N\,\mathrm{Var}(\tilde{\theta}_N) = \int h(x)^2 w(x)^2 \pi(dx) - \theta^2$$

It now follows for the two the following using Jensen's inequality:

$$\int h(x)^2 \pi(dx) \geq \left(\int |h(x)| \pi(dx)\right)^2$$

$$E(x^2) \geq E(x)^2$$

And in the same spirit

$$\int h(x)^2 w(x)^2 \pi(dx) \geq \left(\int |h(x)|\, w(x)\, \pi(dx)\right)^2$$

$$= \left(\int |h(x)| \pi(dx)\right)^2$$

Such that it follows immediately that the only way to reach equality - i.e. the lower bound - above is to have a linear function in λ.

↳ It is then clear that with the density function $g(x) = \frac{|h(x)| f(x)}{\int |h(x)| f(x) \mu(dx)}$ we would reach such a lower bound:

$$\int \frac{|h(x)|^2 \cdot f(x)^2}{\frac{|h(x)| f(x)}{\left(\int |h(x)| f(x) \mu(dx)\right)^2}} \pi(dx) = \int \mathrm{const}^2 \pi(dx)$$

$$= \mathrm{const}^2$$

and this equals

$$\left(\int |h(x)| \pi(dx)\right)^2$$

and this equals

$$(\text{const})^2.$$

$\Rightarrow$ In such a way we would actually have derived a new way to compute an unbiased estimator for which we would have minimal variance.

↳ Notice moreover that in the case $h(x) > 0 \ \forall x$, such that above the absolute value becomes obsolete we would actually have; with the above $w(x)$:

$$N \operatorname{Var}(\hat{\theta}_N) = \underbrace{\left(\int h(x)\,\pi(dx)\right)^2}_{=\theta^2 \text{ by def.}} - \theta^2$$

$$= 0.$$

▽ Important is however to understand how the above stays at the theoretical level as should we know how to integrate over

$$\int |h(x)|\, f(x)\, \mu(dx)$$

then we could compute our parameter directly by integration and would not rely on simulation at all in the first place.

$\Rightarrow$ What such an analysis tells us concretely is however that choosing $g(x)$ as close as possible to $|h(x)|f(x)$ helps extensively in keeping a low variance for our Importance Sampler Estimator

↳ To make this more explicit understand how for instance in the case of an estimation of a mean value of Some rare event the Importance Sampler might be a particularly useful technique to keep the variance low:

useful technique to keep the variance low.

- For instance, let $h(x) - \mathbf{1}_A(x)$ where A is a rare event under the distribution π
- Often, $\hat{\vartheta}_N$ needs many replicates for reasonable accuracy
- By the above result, the optimal π is simply the conditional distribution of $X \sim \pi$ given that $X \in A$
- In practice, choose a density g which is small outside of A and approximately proportional to f in A

check at the exercises there some empirical results where you can get a better taste of the results.

We conclude this section by noting that the multiplicative version of the Importance Sampler is nothing else than the multiplicative control variate.

↳ Such that it follows that we have convergence to the true parameter θ and potentially a low variance if the correlation among $h(x)$ and $w(x)$ is high.

- Which of the two IS versions is more precise, depends on h
- For estimating the probability of rare events, the multiplicative version is typically less precise

As you still sample from the entire distribution and do not focus on the event of interest with the proper adjustment

7) Quasi-Montecarlo

So far we have focused on reducing variance of monte-carlo

↳ Another possibility concern in methods reducing the rate of convergence

This is the task of Quasi-Monte Carlo Methods:

- **Goal:** calculate an approximation of the form

$$\int_{[0,1]^d} h(x)dx \approx \frac{1}{N}\sum_{i=1}^{N} h(u_i)$$

We have seen several methods for reducing the variance

Quasi-Monte Carlo (QMC) construct points $(u_i : 1 \leq i \leq N)$ that are more regular than uniformly distributed points on $[0,1]^d$, but less regular than points from a regular grid

Quasi-Monte Carlo (QMC) construct points $(u_i; 1 \le i \le N)$ that are more regular than uniformly distributed points on $[0, 1]^d$, but less regular than points from a regular grid

One way to construct such points is the Halton Sequence:

Halton sequence for $d = 1$

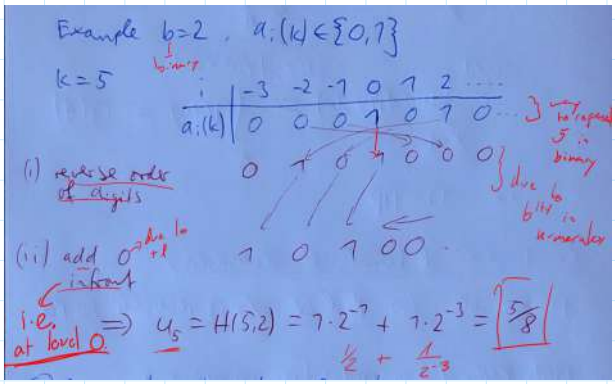
- Choose a natural number $b \ge 2$
- Represent the natural numbers k in the basis b:

$$k = \sum_{i=0}^{\infty} a_i(k) b^i \quad (a_i(k) \in \{0, 1, \ldots, b-1\})$$

- The k-th element of the Halton sequence is then

$$u_k = H(k, b) = \sum_{i=0}^{\infty} \frac{a_i(k)}{b^{i+1}} \in (0, 1)$$

So it is a pretty basic algo, go through the following example to make it even more explicit:



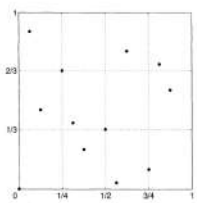
Given the understanding on the generation of such univariate halton sequences, it is possible to combine multiple univariate sequences to obtain a multivariate sequence

Halton sequence for general d

- Use for the j-th component the Halton sequence with basis b_j where b_j is the j-the prime number:

$$u_k = (H(k, 2), H(k, 3), H(k, 5), \ldots, H(k, b_d))^T.$$

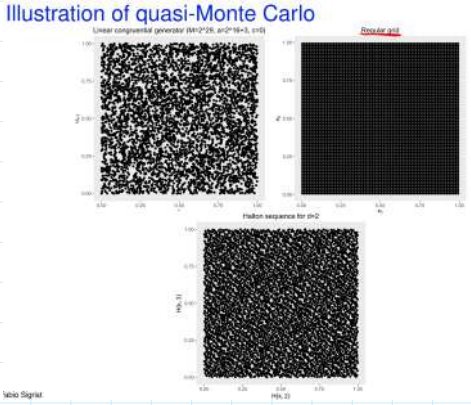
k	$H(k,2)$	$H(k,3)$
0	0	0
1	1/2	1/3
2	1/4	2/3
3	3/4	1/9
4	1/8	4/9
5	5/8	7/9
6	3/8	2/9
7	7/8	5/9
8	1/16	8/9
9	9/16	1/27
10	5/16	10/27
11	13/16	19/27



You will then obtain your sequence that is

You will then obtain your sequence that is less regular than a regular grid:

Illustration of quasi-Monte Carlo



More complex ways to generate such points with better properties exists.

More advanced Quasi-Monte Carlo methods use so-called **(t, m)-nets** in base $b \geq 2$

- These are sets of b^m points $\{u_0, u_1, \ldots, u_{b^m-1}\} \subset [0,1)^d$ such that any "elementary cube"

$$\prod_{i=1}^{d} \left[\frac{a_i}{b^{c_i}}, \frac{a_i+1}{b^{c_i}} \right)$$

with $c_1 + c_2 + \ldots + c_d = m - t \geq 0$ and $0 \leq a_i < b^{c_i}$ $(c_i, a_i, t, m \in \mathbb{N})$ contains exactly b^t points

- Such a cube has the volume b^{t-m}
- Therefore, such cubes contain exactly the expected number of points

- It can be shown now that for such points Quasi-Monte-Carlo will achieve the following:

One can show that quasi-Monte Carlo methods allow to approximate an integral

$$\int_{[0,1]^d} h(x)dx$$

where h has bounded variation with an error of the order

$$\frac{(\log N)^{d-1}}{N}$$

- In contrast to classical MC approximations, there are deterministic bounds for QMC
- In contrast to classical MC approximations, the constants for the error bounds are generally difficult to compute. When they are known, the bounds are often rather inaccurate. I.e., there is limited practical use for such error bounds

Recall in classical MC you have a convergence rate for $P(\hat{\theta}_n - \theta > \varepsilon)$, here for $\hat{\theta}_n - \theta$, i.e. deterministic convergence rate

Notice this is much lower than the $\frac{1}{\sqrt{N}}$ bound for large N. in comparison to classical Monte-Carlo.

That is for classical markov chain we know that the constant for the prob. error bound is the standard deviation of the parameter.

> ⚠ As such constant is not known it is very difficult for finite samples to judge which estimator performs better

QMC are appropriate for problems of moderately high dimension.

Chapter 4 - Markov Chain Monte Carlo

This chapter will introduce the following topics.

1. Markov Chains - Properties and Theorems
2. Metropolis - Hastings Algorithm
3. MCMC using Hamiltonian Dynamics
4. Reversible Jump MCMC
5. Accuracy of MCMC

In many cases, especially in high dimensions, there are no good methods to simulate from a general target distribution π

- The rejection algorithm fails because it almost always rejects (the bound for the ratio of the densities is too large)
- Importance sampling fails because the variance of the weights is too large

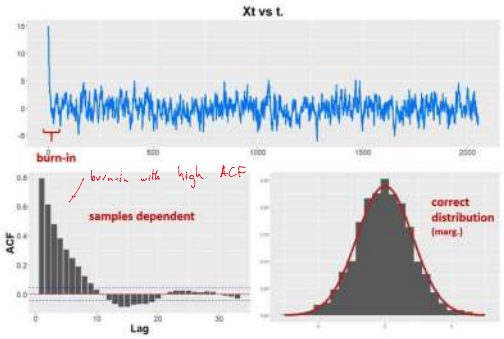
The current standard method for the simulation of distributions in high dimensions is called **Markov chain Monte Carlo (MCMC)**

Through these it is possible to properly sample from a distribution by creating a Markov chain which equilibrium distribution is the target distribution.

↳ We will then see that if the Markov chain satisfies a couple of properties no matter the initial distribution, by running the chain sufficiently often you will converge to the equilibrium distribution.

i.e.

Approach of MCMC: Instead of generating independent samples $X_t \sim \pi$, we generate dependent samples X_t such that for large t, X_t has (approximately) the correct distribution π



The general approach of MCMC can hence be represented as follows:

1. Generate a sequence $(X_0, X_1, \ldots)$ **recursively** such that for large t, X_t has approximately the desired distribution π ← depends on previous X_{t-1}

2. Use the approximation

$$\left(\int h(x)\pi(dx) =\right) \quad \mathbb{E}_\pi(h(X)) \approx \frac{1}{N-r+1}\sum_{t=r}^{N} h(X_t)$$

- The first r simulations are discarded until we reach the target π. r is the so-called **burn-in period**

Markov

$$\left(\int h(x)\pi(dx) =\right) \quad \mathbb{E}_\pi(h(X)) \approx \frac{1}{N-r+1}\sum_{t=r} h(X_t)$$

- The first r simulations are discarded until we reach the target π. r is the so-called **burn-in period**
- In contrast to previous methods, the **X_t's are not independent**

Markov property of lag 1.

We will now see the theory specify a transition rule $F(X_t, U_t)$ with the desired properties, i.e. convergence of $X_t \sim \pi$ for large t by going through the Markov Chains basics theorems.

We will see that for ensuring

How to specify a transition rule

$$X_{t+1} = F(X_t, U_t)$$

for the Markov chain such that the approximation

$$\mathbb{E}_\pi(h(X)) \approx \frac{1}{N-r+1}\sum_{t=r}^{N} h(X_t)$$

We have basically to ensure two properties:

1. If the starting value X_0 **has already the desired distribution** π, then all the following variables $X_1, X_2, \ldots$ **should also have the distribution** π
   - Such a π is called an **invariant or stationary** distribution of the Markov chain
2. The distribution of X_t should **converge to** π **for any arbitrary initial distribution** of X_0

We will see how to construct transition rules that meet these two requirements

1. Basics of MCMC

- Let $\mathbb{X}$ be a space with a σ-algebra $\mathcal{F}$
- A Markov chain describes a discrete time stochastic process on $\mathbb{X}$ with a simple dependence structure: the next state depends on the present, but not on past states
- A **Markov chain** is characterized by:
  1. The probabilities
  $$\mathbb{P}(X_{t+1} \in A | X_t = x) = P(x, A)$$
  for all $x \in \mathbb{X}$ and all $A \in \mathcal{F}$. P is a so-called **transition kernel**
  2. An **initial distribution** ν_0
  $$\mathbb{P}(X_0 \in A) = \nu_0(A)$$

Definition (transition kernel)

A transition kernel P of $(\mathbb{X}, \mathcal{F})$ is a mapping from $\mathbb{X} \times \mathcal{F}$ to $[0,1]$ such that

- *$P(x,.)$ is a probability on $(\mathbb{X}, \mathcal{F})$ for every $x \in X$*
- *$P(.,A)$ is a measurable function for every $A \in \mathcal{F}$*

Definition (Markov chain)

A (time-homogeneous) Markov chain on $(\mathbb{X}, \mathcal{F})$ with initial distribution ν_0 and transition kernel P is a sequence $(X_0, X_1, X_2, ...)$ of random variables with values in $\mathbb{X}$ such that

$$\mathbb{P}(X_0 \in A) = \nu_0(A)$$

and

$$\mathbb{P}(X_{t-1} \in A | X_t - x_t, ... X_0 - x_0) - \mathbb{P}(X_{t+1} \in A | X_t - x_t) - P(x_t, A)$$

- The **joint distribution** of $(X_0, X_1, ..., X_t)$ can be written as

$$\nu_0(dx_0) \prod_{s=1}^{t} P(x_{s-1}, dx_s)$$

We will introduce next some notation that we will use for the rest of the chapter:

- A kernel defines a **mapping of the set of measurable and bounded functions on $(\mathbb{X}, \mathcal{F})$ into itself** by

$$Pf(x) - \int P(x,dy) f(y)$$

- A kernel defines a **mapping of the set of probability distributions on $(\mathbb{X}, \mathcal{F})$** by

$$\nu P(A) - \int \nu(dx) P(x,A)$$

"Probability of ending in A when starting with initial distribution ν and applying P $P(X_1 \subset A | X_0 \sim \nu)$"

- One can **compose two kernels** to form a new kernel by

$$PQ(x,A) = \int P(x,dy) Q(y,A)$$

"Probability of ending in A when starting in x and applying P and then Q"
By P^k we denote the kernel P composed k times with itself

Given that it follows

For a Markov chain with initial distribution ν_0 and transition kernel P, it follows that for any $k > 0$

$$\begin{aligned} \mathbb{P}(X_k \subset A) &= \nu_0 P^k(A), \\ \mathbb{P}(X_{t-k} \in A | X_t = x_t, X_{t-1} = x_{t-1}, ..., X_0 = x_0) &= P^k(x_t, A), \\ \mathbb{E}(f(X_{t+k}) | X_t = x_t) &= P^k f(x_t). \end{aligned}$$

We will now define the concepts of irreducibility, reversibility, through which it will be easy to generate

Definition (invariant, stationary)

A probability distribution π on $(\mathbb{X}, \mathcal{F})$ is called invariant or stationary for a transition kernel P if

$$\pi P = \pi \quad \left(\int \pi(dx) P(x, A) = \pi(A), \ \forall A \in \mathcal{F} \right)$$

Interpretation: if one chooses π as the initial distribution, then X_1 and consequently all X_t have the distribution π

Definition (reversible)

A probability distribution π on $(\mathbb{X}, \mathcal{F})$ is called reversible for a transition kernel P if

$$\pi(dx) P(x, dy) = \pi(dy) P(y, dx)$$

Interpretation: with the initial distribution π, (X_0, X_1) and (X_1, X_0) have the same distribution

$$P(X_0 \in A, X_1 \in B) = \int_A \int_B \pi(dx) P(x, dy) = \int_A \int_B \pi(dy) P(y, dx)$$
$$= P(X_1 \in A, X_0 \in B)$$

- A reversible probability distribution is always invariant

↳ Proof:

$$\pi P(A) = \int_{\mathbb{X}} \pi(dx) P(x, A) = \int_{\mathbb{X}} \int_A \pi(dx) P(x, dy)$$

$$= \int_{\mathbb{X}} \int_A \pi(dy) P(y, dx) = \int_A \pi(dy) = \pi(A)$$

definition of probability, integration over all space $= 1$

So it follows from the above

Reversibility $\Rightarrow$ invariance

notice however that the opposite is not strictly necessary.

Definition (irreducible)

A transition kernel is called irreducible if a probability distribution ψ on $(\mathbb{X}, \mathcal{F})$ exists such that

$$\sum_{k=1}^{\infty} P^k(x, A) > 0$$

for all $A \in \mathcal{F}$ with $\psi(A) > 0$ and for all $x \in \mathbb{X}$

Interpretation: the chain can reach with positive probability all states for any initial distribution

Notice: it is important to understand that combining reducible kernels it is possible to obtain new irreducible kernels.

This idea will be later used. Two such ways to combine kernels are:

$k=1$

for all $A \in \mathcal{F}$ *with* $\psi(A) > 0$ *and for all* $x \in \mathbb{X}$

Interpretation: the chain can reach with positive probability all states for any initial distribution

Notice: it is important to understand that combining reducible kernels it is possible to obtain new irreducible kernels.

This idea will be later used. Two such ways to combine kernels are:

One can **combine reducible kernels** $P_i (i = 1, ..., k)$ so that the resulting kernel is irreducible:

1. Sequential composition in the order $(i(1), i(2), ..., i(k))$

$$P = P_{i(1)} P_{i(2)} \cdots P_{i(k)}$$

2. Random selection among the k possible transitions

$$P = \frac{1}{k}(P_1 + P_2 + \cdots + P_k)$$

Notice moreover the following properties:

1. For Sequential Composition:

- if all of the kernels are invariant so does the composed kernel.

$$\pi P(A) = \int_{\mathbb{X}} \int_A \pi(dx) \cdot (P_1(x, dz) \cdot P_2(z, dy))$$

(given invariance)

$$= \int \int \pi(dz) \cdot P_2(z, dy)$$

(invariance)

$$= \int \pi(dy) = \pi(A) \qquad \square$$

! - If all of the transition kernels are reversible then the sequential composition kernel does not have to be reversible.

$$\pi(dx) \cdot P(x, dy) = \int_{\mathbb{X}} \int_A \pi(dx) \left(P_1(x, dz) \cdot P_2(z, dy)\right)$$

(reversibility kernel P_1)

$$= \int_{\mathbb{X}} \int_A \pi(dz) \cdot P_1(z, dx) \cdot P_2(z, dy) \qquad \text{if } P_1, P_2$$

$$\neq \int_{\mathbb{X}} \int_A \pi(dx) \cdot P_2(x, dy) \cdot P_1(y, dz) \qquad P_2 \neq P_1$$

$$\neq \int_{\mathbb{X}} \int_A \pi(dy) \cdot P_1(y, dz) \cdot P_2(z, dx) = \pi(dy) P(y, dx)$$

2. For Random Selection among k kernels:

- If the random kernels are invariant, so does the resulting kernel

- If the random kernels are invariant, so does the resulting kernel.

$$\int_{x\in A} \pi(dx)\, P(x, dy) = \int_{x\in A} \pi(dx)\, \tfrac{1}{2}\left(P_1(x, dy) + P_2(x, dy)\right) \quad \text{given kernels}$$

$$= \int_A \pi(dy)\, \tfrac{1}{2} + \pi(dy)\, \tfrac{1}{2}$$

$$= \pi(A) \qquad \square$$

- If the kernels are reversible so does the combined kernel

$$\int_{x\in A} \pi(dx)\, P(x, dy) = \int_{x\in A} \pi(dx)\, \tfrac{1}{2}\left(P_1(x, dy) + P_2(x, dy)\right)$$

$$= \int_{x\in A} \pi(dy)\cdot \tfrac{1}{2}\left(P_1(y, dx) + P_2(y, dx)\right) \quad (I)$$

P_1, P_2 def. prob. $= 1$

$$= \int_A \pi(dy) = \pi(A)$$

Moreover it is easy to see that (I):

$$= \int_{x\in A} \pi(dx)\, P(dy, x) \quad \text{which proves reversibility.} \qquad \square$$

With the above definitions it follows now the theorem for Convergence and LLN:

Theorem

Let P be an irreducible transition kernel with a stationary distribution π. Then π is the only stationary distribution, and the following statements are true

- *For all $x \in \mathbb{X}$ and all $A \in \mathcal{F}$ with $\pi(A) > 0$*

$$\mathbb{P}(X_t \in A \text{ infinitely often} \mid X_0 = x) > 0$$

Null-set available

- *For π-almost all $x \in \mathbb{X}$ and all $A \in \mathcal{F}$ with $\pi(A) > 0$*

$$\mathbb{P}(X_t \in A \text{ infinitely often} \mid X_0 \quad x) \quad 1$$

finite mass

- *For π-almost all $x \in \mathbb{X}$ and all h with $\int |h(x)|\pi(dx) < \infty$*

$$\mathbb{P}\left(\frac{1}{n+1}\sum_{t=0}^{n} h(X_t) \to \int h(x)\pi(dx) \mid X_0 = x\right) = 1$$

Convergence

Important: We will not prove the above theorem however it is very powerful as it assures that we will converge to the desired distribution for π-almost all x, given that the transition kernel satisfies the above properties

↳ the issue is now:

What if we start with an x on the Null-set?

↳ We are not guaranteed to converge.

In practice we do not know if we are on such a Null-set, so you must be careful.

↳ Notice moreover that there are sufficient conditions for which the two statements above are valid for all x. Check the literature should you be interested in these.

Given this it is straightforward to see that we aim to construct a MCMC with a transition kernel with the properties of the theorem above so that:

$$\mathrm{E}_{\pi}(\mathbf{h}(\mathbf{X})) \approx \frac{1}{\mathbf{N}-\mathbf{r}+\mathbf{1}} \sum_{\mathbf{t}=\mathbf{r}}^{\mathbf{N}} \mathbf{h}(\mathbf{X}_{\mathbf{t}})$$

For this purpose, we have to **choose a transition kernel satisfying the following three conditions**:

1. P is irreducible
2. π is stationary or reversible for P
3. Simulation from $P(x, .)$ should be "easy" for all x

One such way to construct such a kernel is the Metropolis Hasting algorithm that we will explore next.

(2) Metropolis - Hasting Algorithm

Recall: We have a distribution $\pi(x)$, which we aim to sample from.

↳ So it is straightforward to see that the first step that we are gonna do is to assure that such target distribution is the invariant distribution of our MCMC.

Besides we will have to insure the desired properties of the transition kernel as above it follows:

Discrete case: our target disitribution $\pi(i)$ is a discrete ditribution

(Too) simple idea:

- For each pair of $i < j$, choose (arbitrarily) either $P(i,j)$ or $P(j,i)$
- Choose the other value such that $\pi(i)P(i,j) - \pi(j)P(j,i)$ is fulfilled

  E.g., $P(j,i) - \frac{\pi(i)P(i,j)}{\pi(j)}$

Problem with this approach: no guarantee that $\sum_j P(i,j) - 1$ is satisfied

(Choose a lower / upper triangular matrix)

(rev is satisfied.)

(prob. density in the second arg.)

⇒ So in general more thought is required, it is then possible to see that constructing a reversible chain for the stationary distribution, will lead to the following necessary conditions

Three cases for the condition $\pi(i)P(i,j) = \pi(j)P(j,i)$

1. If $\pi(i) = 0$ and $\pi(j) = 0$, the condition is satisfied automatically
2. If $\pi(i) = 0$ and $\pi(j) \neq 0$, then $P(j,i)$ must be zero
3. If both $\pi(i) \neq 0$ and $\pi(j) \neq 0$, then we must have $P(i,j) > 0 \Leftrightarrow P(j,i) > 0$

Interpretation

2. We must not go to a state which has probability zero from a state with positive probability
3. If a transition between two states which both have positive probability is possible, then also the reverse transition must be possible

- In summary, reversibility implies $\pi(i)P(i,j) > 0 \Leftrightarrow \pi(j)P(j,i) > 0$

The construction of the desired kernel is then guaranteed by the following mechanism

Construction for $P(i,j)$ that satisfies both conditions

1. Choose an arbitrary, irreducible transition matrix $Q(i,j)$ such that

$$\pi(i)Q(i,j) > 0 \Leftrightarrow \pi(j)Q(j,i) > 0$$

2. Set
   - $P(i,j) = Q(i,j)a(i,j)$ with $a(i,j) = \min\left(1, \frac{\pi(j)Q(j,i)}{\pi(i)Q(i,j)}\right)$, for $i \neq j$ (acceptance probability; ↳ Denominator $\neq 0$ if Numerator $\neq 0$)
   - $P(i,i) = 1 - \sum_{j:j\neq i} P(i,j)$

Rest of the mass on diagonal so that basic probability theory applies.

↳ Notice that the rest of the mass occurs and $\sum_{j:j\neq i} P(i,j) \leq 1$ as :

Case 1: $\pi(i)Q(i,j) > \pi(j)Q(j,i)$ (I)

Case 1: $\pi(i)\, Q(i,j) > \pi(j)\, Q(j,i)$ (I)

$$P(i,j) = Q(i,j) \cdot \min\left(1, \frac{\pi(j)\, Q(j,i)}{\pi(i)\, Q(i,j)}\right)$$

$$= \frac{\pi(j)\, Q(j,i)}{\pi(i)}, \text{ given (I)} \leq Q(i,j)$$

$$P(j,i) = Q(j,i) \cdot \min\left(1, \frac{\pi(i)\, Q(i,j)}{\pi(j)\, Q(j,i)}\right) = Q(j,i)$$

Case 2: $\pi(j)\, Q(j,i) \geq \pi(i)\, Q(i,j)$ (II)

$$P(i,j) = Q(i,j)$$

$$P(j,i) = Q(j,i) \cdot \min\left(1, \frac{\pi(i)\, Q(i,j)}{\pi(j)\, Q(j,i)}\right)$$

$$= \frac{\pi(i)\, Q(i,j)}{\pi(j)}, \text{ Given (II)} \leq Q(j,i)$$

It follows immediately

$$P(j,i) \leq Q(j,i) \quad \text{and} \quad P(i,j) \leq Q(i,j)$$

It is then clear that

$$\sum_{j;j\neq i} P(i,j) \leq \sum_{j;j\neq i} Q(i,j) \leq 1$$

Given the above it is now clear that we know how to construct the necessary transition kernel for our target distribution.

Metropolis-Hastings Algorithm for the discrete case simply translates the above into an algorithm formula.

Algorithm (discrete Metropolis-Hastings algorithm)

Simulate $X_0 \sim \nu_0$.
For $t = 1, 2, \ldots$

1. *Simulate* $Y_t \sim Q(i, .)$ $(x_{t-1} = i)$ *and* $U \sim Uniform(0,1)$
2. *If* $U \leq a(i, Y_t)$, *then set* $X_t = Y_t$, *otherwise* $X_t = i$

Comments

- Similar to rejection sampling, but when a proposed value Y_t is rejected, we keep the current value in accordance with $P(i,i) = 1 - \sum_{j\neq i} P(i,j)$ ↳ Markov Chain; status quo → same state.
- Q is called the **proposal distribution** and a is called the **acceptance probability**

Notice that the above corresponds to the more general case previously discussed:

$\boxed{j \neq i}$

$$P(X_t = j \mid X_{t-1} = i) = Pr(Y_t = j \cap U_t \leq a(i,j) \mid X_{t-1} = i)$$
$$= Q(i,j)\, a(i,j)$$
$$= P(i,j)$$

$\boxed{j = i}$

rejection

$$Pr(X_t = i \mid X_{t-1} = i) = Pr(Y_t = i \mid X_{t-1} = i) + Pr(Y_t = j \cap U_t > a(i,j)) \mid X_{t-1} = i)$$
$$= Q(i,i) + \sum_{j \neq i} Q(i,j)\,(1 - a(i,j))$$
$$= 1 - \sum_{j \neq i} Q(i,j)\, a(i,j)$$
$$= 1 - \sum P(i,j)$$

This concludes the Metropolis Hastings Algorithm for the discrete case. Next we will show the result for the continuous case, where more mathematical machinary will be necessary.

MCMC - Cont. Metropolis Hasting and Rest of Class

21 November 2020 12:53

2.2 Metropolis Hasting - Continuous Case

Here the approach is essentially unchanged in contrast to the discrete case.

- Conceptually, the construction of P is similar
- The continuous case analog of

$$\pi(i)Q(i,j) > 0 \Leftrightarrow \pi(j)Q(j,i) > 0$$

is expressed by the concept of absolute continuity

Absolute continuity

Assume that μ and ν are two measures on $(\mathbb{X}, \mathcal{F})$

- μ is called **absolutely continuous** with respect to ν, $\mu << \nu$, if

$$\nu(A) = 0 \Rightarrow \mu(A) = 0$$

} if Null Set for ν then also for μ

- If $\mu << \nu$, then there exists a ν-measurable function f denoted by $f = d\mu/d\nu$ such that

$$\mu(A) = \int_A f(x) d\nu(x)$$

f is called the **Radon-Nikodym** density } go from one measure to the other

- μ and ν are called **equivalent** if $\mu << \nu$ and $\nu << \mu$, i.e.,

$$\nu(A) = 0 \Leftrightarrow \mu(A) = 0$$

Given such notion it is then possible to understand the following Theorem, that we are going to prove next:

We are going to prove next:

Theorem (4.2)

Let π be a probability on $(\mathbb{X}, \mathcal{F})$ and Q be a kernel in the same space such that the two probabilities $\pi(dx)Q(x, dy)$ and $\pi(dy)Q(y, dx)$ on $(\mathbb{X}, \mathcal{F}) \times (\mathbb{X}, \mathcal{F})$ are equivalent. The following hold true.

- *The Radon-Nikodym density of $\pi(dy)Q(y, dx)$ with respect to $\pi(dx)Q(x, dy)$ exists. We denote it by $r(y, x)$.* (i)
- *The following kernel is reversible regarding π:* (ii)

$$P(x, A) = \int_A a(x, y)Q(x, dy) + \mathbf{1}_A(x) \cdot \left(1 - \int_X a(x, y)Q(x, dy)\right),$$

where

$$a(x, y) = \min(1, r(y, x)).$$

Proof:

As for point (i) this follows immediately from measure theory.

We will focus on point (ii)

Notice that once the proof is complete it is immediate to see that the general framework of the Metropolis Hasting is satisfied with an irreducible kernel $Q(\cdot, \cdot)$

It is namely straightforward to see:

$\int_A a(x, y)\, Q(x, dy)$ $\Rightarrow$ Prob of accepting the new proposals and landing

new proposals and landing in A / Prob of $y \in A$ and accepted.

$\mathbb{1}_A \left(1 - \int_X a(x,y)\, Q(x, dy)\right)$ → Prob of staying at the same value.

Proving the above (ii) goes now as follows,

First notice that the Radon-Nikodym density is given by:

$$r(x,y) = \frac{\pi(dx)\, Q(x, dy)}{\pi(dy)\, Q(y, dx)}$$

It now follows due to equivalence that

$$r(x,y) = \frac{d\mu}{d\nu} = \frac{\pi(dx)\, Q(x, dy)}{\pi(dy)\, Q(y, dx)}$$

and

$$r(y,x) = \frac{d\nu}{d\mu} = \frac{\pi(dy)\, Q(y, dx)}{\pi(dx)\, Q(x, dy)}$$

such that

$$\boxed{r(x,y)\, r(y,x) = 1}$$

It now follows using this property that

It now follows using this property that

- if $r(y,x) \leq 1$

$$\Rightarrow \begin{cases} a(x,y) = \min\left(1, r(y,x)\right) = r(y,x) \\ a(y,x) = \min\left(1, r(x,y)\right) = 1 \end{cases}$$

so that "1 in such case

$$\Rightarrow \quad a(x,y) = r(y,x) \cdot a(y,x)$$

- if $r(y,x) > 1$

$$\Rightarrow \begin{cases} a(x,y) = \min\left(1, r(y,x)\right) = 1 \\ a(y,x) = \min\left(1, r(x,y)\right) = r(x,y) \end{cases}$$

$$\Rightarrow \quad a(x,y) = 1 = r(x,y)\, r(y,x) = a(y,x) \cdot r(y,x)$$

So in both of the exhaustive cases the relation holds such that we can generally write

$$a(x,y) = r(y,x)\, a(y,x)$$

It follows now that using the above it is straightforward to show the reversibility property:

$$\iint_{A\,B} \pi(dy)\, P(y,dx)$$

inserting our adjusted kernel $Q(\cdot,\cdot)$

$$= \iint_{A\,B} \pi(dy)\, Q(y,dx)\cdot a(y,x) \qquad \text{(I)}$$

$$+ \int_{A} \pi(dy)\, 1_B \cdot \Big(1 - \int_{X} Q(y,dx')\, a(y,x')\Big) \qquad \text{(II)}$$

you can then rewrite (I):

$$\iint_{A\,B} \pi(dy)\, Q(y,dx)\, a(y,x)$$

given Radon Nikodym property

$$= \iint \pi(dx)\, Q(x,dy)\cdot r(y,x)\cdot a(y,x)$$

$$= \iint \pi(dx)\, Q(x,dy)\, a(x,y)$$

and for the (II):

as dy is just a variable you can substitute it with dx as long as you integrate over the same space

as dy is just a variable you can substitute it with dx, as long as you integrate over the same space there is no prob. in it:

$$\int_A \pi(dx)\, \mathbb{1}_B \left(1 - \int_X a(x,x')\, Q(x, dx')\right)$$

such that putting the two together you have:

$$= \int\int_{AB} \pi(dx)\, Q(x, dy)\, a(x,y) + \int_A \pi(dx)\, \mathbb{1}_B \left(1 - \int_X a(x,x')\, Q(x,dx')\right)$$

$$= \int\int_{A\,B} \pi(dx)\, P(x, dy) \qquad \square$$

Given the above it is now clear that upon having the Radon-Nikodym derivative it is easy to obtain the Metropolis-Hasting version in the continuous case

- How can we verify that $\pi(dx)Q(x, dy)$ and $\pi(dy)Q(y, dx)$ are equivalent?
- How do we calculate the Radon-Nikodym density $r(y, x)$?

Lemma (4.1)

Let P_1 and P_2 be two probabilities on $(\mathbb{X}, \mathcal{F})$.

1. *If P_1 and P_2 have densities p_1 and p_2 w.r. to a σ-finite measure μ, then P_1 is absolutely continuous with respect to P_2 iff*

Let P_1 and P_2 be two probabilities on $(\mathbb{X}, \mathcal{F})$.

1. *If P_1 and P_2 have densities p_1 and p_2 w.r. to a σ-finite measure μ, then P_1 is absolutely continuous with respect to P_2 iff $\{x|p_2(x) = 0, p_1(x) > 0\}$ is a null set with respect to μ. The Radon-Nikodym density of P_1 with respect to P_2 is then $p_1(x)/p_2(x)$, independent of the choice of μ.*

So in practice verify the above and set Radon-Nikodym accordingly.

Metropolis-Hastings algorithm: the continuous case

- Assume that both $\pi(dx)$ and the proposal $Q(x, dy)$ have densities $\pi(x)$ and $q(x, y)$ with respect to the Lebesgue measure (continuous case) or the counting measure (discrete case)

- Then $\pi(dx)Q(x, dy)$ and $\pi(dy)Q(y, dx)$ are equivalent if for all pairs (x, y)

$$\pi(x)q(x, y) > 0 \Leftrightarrow \pi(y)q(y, x) > 0$$

  I.e.,
  - $q(x, y) = 0$ if $\pi(x) > 0$ and $\pi(y) = 0$

    and
  - $q(x, y) > 0 \Leftrightarrow q(y, x) > 0$ if $\pi(x) > 0$ and $\pi(y) > 0$

- It follows that

$$r(y, x) = \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)} \Rightarrow a(x, y) = \min\left(1, \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}\right)$$

- Since only the ratio $\pi(y)/\pi(x)$ appears in the acceptance ratio, **it is sufficient to know π up to a normalizing constant**

- **Irreducibility** of Q is transferred to P. In particular, $q(x, y) > 0$ for all (x, y) is sufficient (but not necessary) in order that P is irreducible

Metropolis-Hastings algorithm: the continuous case

Metropolis-Hastings algorithm: the continuous case

Simulating from a Markov chain with transition kernel P can be done as follows:

Algorithm (Metropolis-Hastings algorithm)

Simulate $X_0 \sim \nu_0$
For $t = 1, 2, \ldots$

1. *Simulate* $Y_t \sim q(X_{t-1}, \cdot)$ *and* $U \sim Uniform(0,1)$
2. *If* $U \leq a(X_{t-1}, Y_t)$, *then set* $X_t = Y_t$, *otherwise* $X_t = X_{t-1}$

We will now look at two particular types of the Metropolis - Hasting algorithm.
Recall that the transition kernel can be chosen arbitrarily as long it is irreducible.

↳ We will now see two possible ways to specify it, moreover we will go into the componentwise modification that clarifies why the Gibbs Sampler works.

(2.3) Two special Cases of the Metropolis Hasting

1. Random Walk Metropolis

→ Here the idea is to specify the $Q(x, dy)$ as some multivariate- normal distribution in $\mathbb{R}^p$. wrt the Lebesgue measure on the space $(\mathbb{X}, \mathcal{F})$.

$(\mathbb{X}, \mathcal{F})$.

↳ The idea is hence:

$$X = x \qquad Y = X + \varepsilon \qquad \varepsilon \sim q(z)dz \in \mathbb{R}^p$$

and indep of x.

> ! Hence the idea is that we stay in expectation at the current value and we walk randomly in $\mathbb{R}^p$.

Notice now that as for the multivariate-normal is symmetric it holds:

$$q(x) = q(-x)$$

moreover it holds that the probability of the events x, y is just dependent on the distance among the two (recall that $Y = X + \varepsilon$) such that $Y - X = \varepsilon$.

$$q(x, y) = q(y - x) \leftarrow$$

Given such property it follows immediately:

$$\alpha(x, y) = \min\left(1, \frac{\pi(y)\ q(y, x)}{\pi(x)\ q(x, y)}\right)$$

$$\alpha(x,y) = \min\left(1, \frac{\pi(y)\, q(x,y)}{\pi(x)\, q(x,y)}\right)$$

$$= \min\left(1, \frac{\pi(y)\, q(x-y)}{\pi(x)\, q(y-x)}\right)$$

$$= \min\left(1, \frac{\pi(y)}{\pi(x)} \cdot \frac{q(y-x)}{q(y-x)}\right)$$ } symmetry

Such that overall $\boxed{a(x,y) = \min\left(1, \frac{\pi(y)}{\pi(x)}\right)}$

Notice, moreover, that the above holds for also more general $q(z)\,dz$ where the symmetric property holds, and that manage to sample the entire $\mathbb{R}^p$ space with positive probability.

As a final note, understand, that Σ of the multivariate normal distribution is a tuning parameter. If it is too little you will hardly explore the entire space. The effective sample will be small.

↳ If it is too large you will possibly lower the chances of accepting a new proposal as it is too random. You must go with something in between. We skip rigorous optimality derivations here.

derivations here.

As a final note, as $a(x,y)$ just depends on $\frac{\pi(y)}{\pi(x)}$ it follows immediately that we always accept, when we move to a point with higher mass in the target density.

As a second method worth to consider is the

- Independence Sampler

↳ This will be useful in a moment when we will refer to the conditional independence of the Gibbs Sampler.

The idea, is here that if you have a density function

$$q(x,y) = q(y)$$

i.e. which just depends on the proposed value, then the acceptance probability simplifies to

$$a(x,y) = \min\left(1, \frac{\pi(y)}{\pi(x)} \cdot \frac{q(y)}{q(x)}\right)$$

We will now introduce the Componentwise Modification which proves why the Gibbs Sampler works.

2.3 Componentwise Modification

In some cases choosing a proposal distribution $Q(x, dy)$ with a density does not lead to an efficient algorithm. With such a choice, the proposed value can be anywhere in the space $\mathbb{X}$. If the current value x is plausible for π and $\mathbb{X}$ is high-dimensional, then the proposed value will often be less plausible than the current one, hence rejection is very likely. In such cases, it can be advantageous if the proposed value differs from the current one in only a few components.

We will focus now on the two-dimensional case, however this can be easily generalized.

- Let $\mathbb{X} = \mathbb{X}_1 \times \mathbb{X}_2$, i.e., $x \in \mathbb{X}$ has the form (x_1, x_2) with $x_k \in \mathbb{X}_k$
- Assume that π is absolutely continuous with respect to the product measure $\mu_1(dx_1)\mu_2(dx_2)$ and we denote the density also by π

} So the density is defined both on the full space as well on the subspace with product measure $\mu_1(dx_1)\mu_2(dx_2)$

The above idea is then to identify different proposal for each different component.

↳ You do not move in $\mathbb{R}^p$

You do not move in $\mathbb{R}^1$ dimension, you move in $\mathbb{R}^p$ componentwise each time moving in $\mathbb{R}$ for one variable

- Consider **two proposal distributions** Q_1 and Q_2:
  - Q_1 modifies only the first component with an absolutely continuous distribution while the second component remains the same

$$Q_1(x, A_1 \times A_2) = \int_{A_1} q_1(x, y_1)\mu_1(dy_1) \cdot \mathbf{1}_{A_2}(x_2)$$

second unchanged

  - Q_2 modifies only the second component with an absolutely continuous distribution while the first component remains the same

$$Q_2(x, A_1 \times A_2) = \int_{A_2} q_2(x, y_2)\mu_2(dy_2) \cdot \mathbf{1}_{A_1}(x_1)$$

first unchanged

Notice now that given our previous theorems we know that we can construct a transition kernel with the desired properties for our MCMC, if

- $\pi(dx)\, Q_k(x, dy)$ equivalent to $\pi(dy)\, Q_k(y, dx)$
- We can calculate the Radon-Nikodym density of $Q_k(x, dy)$ and $Q_k(y, dx)$ respectively.

We will now prove the above two points such that it becomes trivial to see how to leverage componentwise modification

such that it becomes trivial to see how to leverage componentwise modification in order to specify proper transition kernel for which the MCMC will converge to the target distribution.

We can do the above using the following Lemma from measure theory:

Lemma (4.1)

Let P_1 and P_2 be two probabilities on $(\mathbb{X}, \mathcal{F})$.

1. *If P_1 and P_2 have densities p_1 and p_2 w.r. to a σ-finite measure μ, then P_1 is absolutely continuous with respect to P_2 iff $\{x|p_2(x) = 0, p_1(x) > 0\}$ is a null set with respect to μ. The Radon-Nikodym density of P_1 with respect to P_2 is then $p_1(x)/p_2(x)$, independent of the choice of μ.*
2. *Let ϕ be a measurable injective mapping from $(\mathbb{X}, \mathcal{F})$ to $(\mathbb{Y}, \mathcal{G})$ and let P_i' denote the distribution $P_i \circ \phi^{-1}$ (i.e. $\phi(X) \sim P_i'$ if $X \sim P_i$). If P_1 has the density r with respect to P_2, then P_1' has the density $r(\phi^{-1}(y))$ with respect to P_2'.*

Using the first Lemma it is easy to see that

$$\pi(dx) \cdot Q_1(x, dy) \quad \text{equivalent to} \quad \pi(dy) Q_1(y, dx)$$

as sharing the same Null-Set for $\mu_1 \mu_2$ and having the same support $A_1 \times A_2$.

The question is then on how to obtain the Radon-Nikodym derivative.

This follows immediately from the second point of the above Lemma.

Consider in fact the following map $\phi: \mathbb{R}^3 \to \mathbb{R}^4$

$$\phi(X_1, X_2, Y_1) \to (X_1, X_2, Y_1, X_2)$$

Then such map is injective cause each element of the domain of $\mathbb{R}^4$ is repesented by one distinct element of $\mathbb{R}^3$.

We can then see that

$$\phi^{-1}(X_1, X_2, Y_1, Y_2) = (X_1, X_2, Y_1) \cdot \mathbb{1}_{X_2 = Y_2}$$

It now follows that as

$$\pi(dx)\ Q((x_1, x_2), y_1) \quad \text{has density}$$

$$\pi(x)\ q((x_1, x_2), y_1) \quad = r(y)$$

Then

$$\pi(dx)\ Q((x_1, x_2), (y_1, y_2)) \quad \text{has density}$$

$$\pi(\cdot)\ (\cdot)\ \mathbb{1} \quad = r(\phi^{-1}(\cdot))$$

$$\pi(x)\; q((x_1,x_2),y_1)\; \mathbb{1}_{x_2=y_2} = r(\phi^{-1}(y))'$$

The same holds for

$\pi(dy)\; Q((y_1,y_2),(x_1,x_2))$ has density

$$\pi(x)\; q((y_1,y_2),x_1)\; \mathbb{1}_{x_2=y_2}$$

Plugging this in, it follows that the Radon-Nikodym density is

$$r_1(y,x) = \frac{\pi(y_1,y_2)\, q^{\overset{x_2}{\mathbb{1}}}((y_1,y_2),x_1)}{\pi(x_1,y_2)\, q^{\overset{x_2}{\mathbb{1}}}((x_1,x_2),y_1)} \cdot \mathbb{1}_{x_2=y_2}$$

(cond. prob. rule)

$$= \frac{\pi_{1|2}(y_1|y_2)\cdot \cancel{\pi(x_2)}}{\pi_{1|2}(x_1|x_2)\; \cancel{\pi(x_2)}} \; \frac{q((y_1,y_2),x_1)}{q((x_1,x_2),y_1)}$$

It holds now that:

- $r_1(x,y)$ for $x_2 \neq y_2$ can be defined arbitrarily since all the mass is on $x_2 = y_2$.

arbitrarily since all the mass is on $x_2 = y_2$.

- if $q((x_1, x_2), y_1) = \pi_{1|2}(y_1 \mid x_2)$ ⇒ so does not depend on the first value (x_1) conditioning on the second one

then it is immediate to see

$$r_1(y, x) = \frac{\pi_{1|2}(y_1 \mid x_2)}{\pi_{1|2}(x_1 \mid x_2)} \cdot \frac{q_1((y_1, x_2), x_1) = \pi_{1|2}(x_1 \mid x_2)}{q_1((x_1, x_2), y_1) = \pi_{1|2}(y_1 \mid x_2)}$$

so you see the idea of independence sampling here.

$$= 1$$

Such that if the density has the form of the conditional density above you would always accept.

The same holds for $r_2(y, x)$ with $q_2((x_1, x_2), y_2) = \pi_{2|1}(x_2 \mid x_1)$ when doing the calculations.

So that you see that when using componant wise modification it is easy to construct the transition kernel with the desired properties.

It now follows that

assuming the two irreducible

- If one applies both kernel $\bar{P}_1$ and $\bar{P}_2$ in an alternating sequence,

- If one applies both kernel P_1 and P_2 in an alternating sequence, the resulting kernel P is irreducible
- The combination of P_1 and P_2 can be made according to a fixed or random order. For a fixed order, the reversibility is usually lost, but the stationary distribution does not change

Check the above derivation of $P_1 \cdot P_2$ vs random $\frac{1}{2}(P_1 + P_2)$.

It is now clear that the Gibbs Sampler is just but a special case of the componentwise modification of above.

↳ i.e. particularly it holds for the Gibb Sampler

$$q_1(\vec{x}, y_1) = \pi(y_1 \mid x_2)$$

$$q_2(\vec{x}, y_2) = \pi(x_2 \mid x_1)$$

such that the above lemma with radon-nikodym density being equal to 1 holds and hence the acceptance probability always equals 1.

↳ This leads to the following for the Gibbs Sampler:

The Gibbs sampler for k components

The Gibbs sampler for k components

- We partition $\mathbb{X}$ into k components
- We denote by $\pi_i(x_i|x_{-i})$ the conditional density of the i-th component x_i of x given all the other components of x, $x_{-i} = (x_j)_{j \neq i}$
- The $\pi_i(x_i|x_{-i})$ are called full-conditionals
- $\pi_i(x_i|x_{-i}) \propto \pi(x)$. I.e., we can identify $\pi_i(x_i|x_{-i})$ by inspecting $\pi(x)$

  ↳ and ignore all terms depending on x_i

Algorithm (Gibbs sampler)

Simulate $X_0 = (X_{10}, X_{20}, \ldots, X_{k0}) \sim \nu_0$

For $t = 1, 2, \ldots,$ *simulate*

1. $X_{1t} \sim \pi_1(x_1|x_{2t-1}, \ldots, x_{kt-1})$
2. $X_{2t} \sim \pi_2(x_2|x_{1t}, x_{3t-1}, \ldots, x_{kt-1})$

   ...

   i. $X_{it} \sim \pi_i(x_i|x_{1t}, \ldots, x_{i-1t}, x_{i+1t-1}, \ldots, x_{kt-1})$

   ...

   k. $X_{kt} \sim \pi_k(x_k|x_{1t}, \ldots, x_{k-1t})$

- In each step i, we update only the component x_i and leave all other components x_{-i} of x unchanged. The component x_i is updated according to the full conditional distribution π_i of our target π

Another option is to use the componentwise modification together with the Random walk concept.

- Another special case of the Metropolis-Hastings algorithm with componentwise modification is the **componentwise random walk Metropolis algorithm**
- Use as proposal densities

  indep. on 1st component

  $q_1(x, y_1) = q_1(y_1 - x_1)$? then immediately ... $V_1 = X_1$ and

- Use as proposal densities

$$q_1(x, y_1) = q_1(y_1 - x_1)$$
$$q_2(x, y_2) = q_2(y_2 - x_2)$$

- If the random walk is symmetric, the acceptance probabilities are

$$a_1(x, y_1) = \min\left(1, \frac{\pi_{1|2}(y_1|x_2)}{\pi_{1|2}(x_1|x_2)}\right)$$
$$a_2(x, y_2) = \min\left(1, \frac{\pi_{2|1}(y_2|x_1)}{\pi_{2|1}(x_2|x_1)}\right)$$

Notice finally that you can combine multiple methods and use different methods for different components:

- One can also **combine different proposal densities for different components**. E.g., one can combine Gibbs steps with random walk Metropolis steps

Hamiltonian MCMC

30 November 2020 11:54

In some situations, commonly used Metropolis-Hastings versions such as the Gibbs sampler or the random walk Metropolis algorithm explore the target density π only slowly

↳ This happens especially when there is high correlation among the variables.

Consider for instance the following

For illustration, consider the case where we want to simulate from a bivariate normal distribution

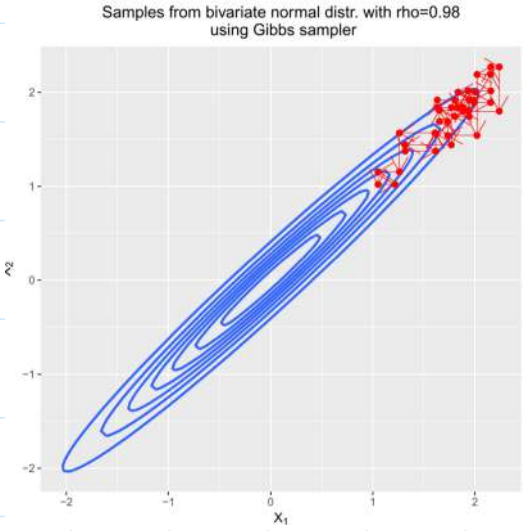
note: these can be also pooled any distances

$$(X_1, X_2) \sim N\left(0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right), \quad -1 < \rho < 1,$$

with high correlation ρ, e.g., $\rho = 0.98$

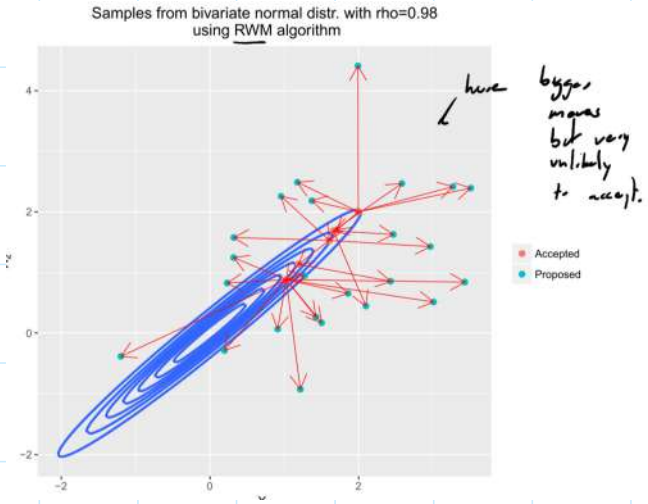
Then:

The Gibbs sampler makes only small moves and the random walk Metropolis (RWM) algorithm makes either small moves or has a low acceptance probability for proposals with big moves



very high autocorrelation

And:



have bigger moves but very unlikely to accept.

stucked at a particular value a lot of time before accepting a new prop.

We will introduce next Hamiltonian MCMC that tries to address the issues above.

- **Idea:** the Hamiltonian Monte Carlo (HMC) algorithm allows for making big moves that are still accepted with high probability

- **Assumptions:**
  - $X \in \mathbb{R}^p$ and X has an absolutely continuous distribution with density π with respect to the Lebesgue measure
  - We are able to evaluate the gradient of $\log \pi$ efficiently

these are requirements that must hold.

The idea in order to achieve that is to use hamiltonian dynamics in order

to use hamiltonian dynamics in order to sample from a new target $\tilde{\pi}(x, u)$ with the following property:

- Consider a new target $\tilde{\pi}$ on a space with doubled dimension

$$\tilde{\pi}(x, u) \propto \pi(x) \exp\left(-\frac{1}{2} u' M^{-1} u\right)$$

  - M is a symmetric, positive-definite "mass" matrix, which is often diagonal or simply a scalar multiple of the identity matrix
  - In the following, we assume that $M = \text{diag}(m_i)$

It follows immediately that such density is strictly related to Hamiltonian dynamics as:

- The **Hamiltonian $H(x, u)$** is defined as

$$H(x, u) = -\log \pi(x) + \sum_{i=1}^{p} \frac{u_i^2}{2m_i}$$

- Physical interpretation
  - x is the position (of particle)
  - u is the momentum
  - $-\log \pi(x)$ is the potential energy
  - $\sum_{i=1}^{p} \frac{u_i^2}{2m_i}$ the kinetic energy
  - $H(x, u)$ is the total energy in the system

As it follows therefore that

$$\tilde{\pi}(x,v) = e^{H(x,v)} \cdot e^{-1/2} = \left(\pi(x) \cdot \exp(v' M^{-1} v)\right) \cdot e^{-1/2}$$

$$\tilde{\pi}(x,v) \propto e^{H(x,v)}$$

Moreover, due to the interesting factorization property of the target distribution, we have:

$$\tilde{\pi}(x,v) = \pi(x) \cdot \exp(v' M^{-1} v)$$

∇_v Independence of $x \wedge v$

Given that the idea will be the one of generating new proposals x^*, v^* using Hamiltonian Mechanics i.e. using a map $G(x,v)$ derived from Hamiltonian Mechanics, such

from Hamiltonian Mechanics, such that:

$$\tilde{\pi}(x, v) = \tilde{\pi}(G(x, v)), \text{ i.e.}$$

the distribution is invariant and use then the doubled dimension and the kinetic energy V; to allow big moves in exploring the density $\tilde{\pi}$.

↳ Given the independence property of X and V it will in fact be possible to make big jumps in the $\tilde{\pi}(\cdot, \cdot)$ while keeping the same distribution. Moreover due to independence it is clear that just observing the x^*, we have $x^* \sim \pi(x^*)$ by def of $\tilde{\pi}(x, v)$.

$x^* \sim \pi(x^*)$ by dint of $\tilde{\pi}(x, u)$.

The question is then on how to generate the map $G(x, u)$.

↳ We will pose a second restriction to it:

- HMC is based on a deterministic, invertible map $G(x, u)$ that is volume preserving and keeps $\tilde{\pi}$ invariant

$$\left|\det \frac{\partial G(x,u)}{\partial x \partial u}\right| = 1, \quad \tilde{\pi}(G(x,u)) = \tilde{\pi}(x,u), \forall x, u$$

↳ i.e. the volume preserving property.

↳ this will be helpful as with it there will be no need to calculate the Jacobian of G which would be necessary in the process as we will see and would be a computationally expensive calculation.

calculation.

This will ultimately lead to the following:

- The **transformation $G(x, u)$** is given by the solution of the ordinary differential equation

$$\frac{dx_i}{dt'} = \frac{\partial H(x,u)}{\partial u_i} = \frac{u_i}{m_i}$$
$$\frac{du_i}{dt'} = -\frac{\partial H(x,u)}{\partial x_i} = \frac{\partial \log \pi(x)}{\partial x_i}, \qquad 0 \le t' \le T,$$

with initial condition (x, u)

- $G(x, u)$ is volume preserving and keeps $\tilde{\pi}$ invariant

↳ we will prove this next, before however understand that the volume preserving property is necessary as with it, it follows that for the Radon-Nikodym derivative of

$$\pi(d(x^*, v^*))\, Q((x^*, v^*), d(x, v))$$

w.r.t.

$$\pi(d(x, v))\, Q((x, v), d(x^*, v^*))$$

that is dist after transition

that is

$$r(y,x) = \frac{\pi(x,v)}{\pi(G(x,v))}$$

dist after transition / dist after transition.

, which by change of variable formula in nom and denom. yields:

$$r(y,x) = \frac{\pi(x,v)}{\pi(G(x,v))}$$

here $H(x) = y = x, v$, $x = G(x,v)$

$$= \frac{\pi(\overset{\text{now}}{\curvearrowright} y)}{\pi(G(x,v))} = \frac{\pi(H^{-1}(y)) \cdot \det\left|\frac{\partial H^{-1}(y)}{\partial y}\right|}{\pi(H^{-1}(x)) \cdot \det\left|\frac{\partial H^{-1}(x)}{\partial x}\right|}$$

make this change of variable what before y now x and vice versa.

now $\hookrightarrow x$

here now

$$H(y) = x = G(x,v)$$

$$y = (x,v) = x$$

mmm... so smth still wrong above... anyway it should theoretically be possible then to prove:

you see from here that it = 1 more easy to compute

$$r(y,x) = e^{-H(x^*,v^*) + H(x,v)} \det\left|\frac{\partial G(x,v)}{\partial(x,v)}\right|$$

$r(y,x) = e$ ... $\left| \det \frac{\partial(\cdot,\cdot)}{\partial(x,v)} \right|$

via change-of-variable formula.

I.e. with the volume-preserving property, the Radon-Nikodym derivative is easy to compute, and that will be helpful when correcting and keep the invariance property via a Metropolis-Hasting acceptance step, that is lost when solving the ODE.

We will now show the desired properties of the map $G_t(x,v)$ constructed as a solution of the ODE above.

(1.) $G(t,v)$ is invariant:

This is straightforward to see given that

$$(I)\quad \frac{d}{dt'} H(x(t'), v(t')) = \sum_{i=1}^{\text{all variables}} \frac{\partial H(x(t'), v(t'))}{\partial x_i} \cdot \frac{\partial x_i}{\partial t'} + \frac{\partial H(x,v)}{\partial v_i} \frac{\partial v_i}{\partial t'}$$

chain rule

total differential = sum of partial differentials

total differential = sum of partial differentials

from ODE definition

$$\frac{\partial v}{\partial t'} = \frac{\partial \log \pi(x)}{\partial x_i} = -\frac{\partial H(x,v)}{\partial x_i}$$

$$\frac{\partial x_i}{\partial t'} = \frac{\partial H(x,v)}{\partial v_i} \quad \rightsquigarrow \text{(I)}$$

it follows

$$\frac{d}{dt'} H(x(t'), v(t')) = \sum \frac{\partial H}{\partial x_i}\frac{\partial H}{\partial v_i} - \frac{\partial H}{\partial v_i}\frac{\partial H}{\partial x_i}$$

$$= 0$$

so that $H(x,v)$ is invariant under $G(x,v)$ and so must also be

$$\tilde{\pi} \propto e^{-H(x,v)}$$

as a straightforward consequence.

(2) $G(x,v)$ is volume preserving.

(2) $G(x,u)$ is volume preserving, so that

$$\det \left| \frac{\partial G(x,u)}{\partial(x,u)} \right| = 1$$

- First, note that

divergence (taking partial derivative w.r.t. every component and summing it up)

$$\nabla \circ \left(\frac{\partial H}{\partial u_i}, -\frac{\partial H}{\partial x_i} \right) = \frac{\partial^2 H}{\partial x_i \partial u_i} - \frac{\partial^2 H}{\partial u_i \partial x_i} = 0$$

(ode)

$$\Rightarrow \nabla \circ F = \nabla \circ \left(\frac{\partial H}{\partial u}^T, -\frac{\partial H}{\partial x}^T \right)^T = 0$$

(every direction not indexed)

$$\Rightarrow \left(\left| \frac{\partial G}{\partial(x,u)}(x,u) \right| = \right) \det \frac{\partial G}{\partial(x,u)}(x,u) = 1$$

→ det of Jacobian = 1, volume preserving (Liouville's theorem)

- (write (x,u) instead of $(x(t'),u(t'))$)

$$\frac{d}{dt'} \frac{\partial G(x(t'),u(t'))}{\partial(x,u)} = \frac{\partial}{\partial(x,u)} \frac{d}{dt'} G(x,u)$$

(→ def of ODE, so that)

derivative w.r.t. time of the ODE

(definition of G & F)

$$= \frac{\partial}{\partial(x,u)} \left(F(G(x,u)) \right)$$

(→ ODE equation)

(chain rule)

$$= \left(\frac{\partial F}{\partial(x,u)} \right) (G(x,u)) \cdot \frac{\partial}{\partial(x,u)} G(x,u)$$

(one can show that this Jacobian is invertible - here skipped.)

$$\Rightarrow \left(\frac{\partial G(x,u)}{\partial(x,u)} \right)^{-1} \frac{d}{dt'} \frac{\partial G}{\partial(x,u)}(x,u) \overset{(*)}{=} \left(\frac{\partial F}{\partial(x,u)} \right) (G(x,u))$$

(can show that $\frac{\partial G}{\partial(x,u)}$ is invertible)

one can show that this Jacobian is invertible - here skipped.

one can show that this Jacobian is invertible - here skipped.

$$\Rightarrow \left(\frac{\partial G(x,u)}{\partial(x,u)}\right)^{-1} \frac{d}{dt'} \frac{\partial G}{\partial(x,u)}(x,u) \overset{(*)}{=} \left(\frac{\partial F}{\partial(x,u)}\right)(G(x,u))$$

(can show that $\frac{\partial G}{\partial(x,u)}$ is invertible)

Use Jacobi's formula:

$$\frac{d}{dt'} \det \frac{\partial G(x,u)}{\partial(x,u)} = \det \frac{\partial G(x,u)}{\partial(x,u)} \cdot \operatorname{tr}\left(\left(\frac{\partial G(x,u)}{\partial(x,u)}\right)^{-1} \frac{d}{dt'} \frac{\partial G(x,u)}{\partial(x,u)}\right)$$

$$\overset{(*)}{=} \operatorname{tr}\left(\left(\frac{\partial F}{\partial(x,u)}\right)(G(x,u))\right) = (\nabla \cdot F)(G(x,u)) = 0$$

as we showed previously that the divergence is 0 no matter the values of F

(3)

- Since $\frac{\partial}{\partial(x,u)} G(x(0),u(0)) = I$ (initial condition)

i.e. melt x, v initial with it you will stay at identity

follows $\Rightarrow$ $\left|\frac{\partial}{\partial(x,u)} G(x,u)\right| = 1$

$\Rightarrow$ as now it was previously shown that it does not change in time it stays at 1.

$\left(\frac{d}{dt'} \det \frac{\partial G}{\partial(x,u)}(x,u) = 0\right)$

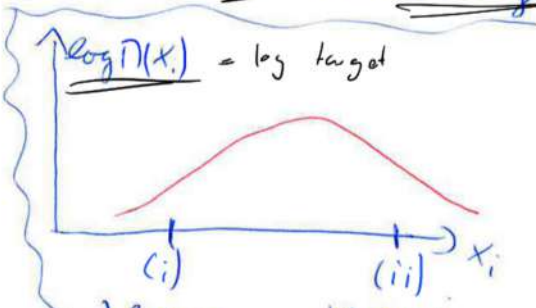
As a final step understand that Hamiltonian MCMC works as:

Hamiltonian MCMC works as:

Intuition for HMC → this in general intuition of why Hamiltonian Monte Carlo Works well and good acceptance.

Assume at T' $u_i > 0$, i.e., $x_i(T')$ is increasing

solved ODE



Case (i): if $\frac{\partial \log \pi(x_i)}{\partial x_i} > 0$: $\frac{\partial \log \pi(x_i)}{\partial x_i} > 0$, $\frac{\partial \log \pi(x_i)}{\partial x_i} < 0$

$\Rightarrow$ $x_i(T')$ moves into a direction of higher mass & by (**) u_i increases
↳ def of ODE
↳ so move even faster.

Case (ii): if $\frac{\partial \log \pi(x_i)}{\partial x_i} < 0$:

$\Rightarrow$ $x_i(T')$ moves into a lower-density region & by (**) u_i is decreased and eventually will be negative → as v_i will decrease.

Where,

(**) = ODE definition.

Notice that this concludes the general understanding of the Hamiltonian MCMC, two small adjustments are still needed in order to have a

still needed in order to have a theoretically sound algorithm:

1. We need to solve the ODE to get proposals for our distribution $\tilde{\pi}(x, u)$.

- In practice, we need to solve the differential equation by some discretization procedure
- The so-called **leap frog method** induces only small changes to $\tilde{\pi}$, it preserves volume exactly and it is time-reversible, i.e., its implied mapping G is invertible
- The exact **invariance of $\tilde{\pi}$ is restored by a Metropolis-Hastings step at the end** with acceptance ratio

$$a((x, u), (x^*, u^*)) = \min\left(1, \exp\left(-H(x^*, u^*) + H(x, u)\right)\right),$$

where (x^*, u^*) is the newly proposed value and (x, u) the current one

as argued above.

2. We showed that exact invariance holds, meaning that $\tilde{\pi}(x^*, u^*) = \pi(x, u)$ holds $\forall\, x^*, u^*$, cause in particular $H(x, u) = H(x^*, u^*)$.

↳ But then:

If $C = H(x,u) = H(x^*,u^*) = -\log \Pi(x^*) + \sum_{i=1}^{D} \frac{u_i^{*2}}{2m_i}$

↳ invariance, hamiltonian

$\Rightarrow -\log \Pi(x^*) = C - \sum_{i=1}^{D} \frac{u_i^{*2}}{2m_i}$ remains constant under map G.

$\leq C$ ← always neg $\forall x^*$ → no matter the tuning parameter

$\Rightarrow$ Cannot reach every state of Π

↳ irreducibility broken

So that we will not sample from all of our density and the irreducibility necessary property of our chain is lost.

- **Solution:** First, simulate an independent new component u and then apply the mapping G
  - If $(X, U) \sim \tilde{\pi}$ and $U' \sim N(0, \text{diag}(m_i))$ is independent of (X, U), then also $(X, U') \sim \tilde{\pi}$

↳ Then you can easily see that sampling u from their marginal - a multivariate normal distribution - you will be able to sample from the entire density $\pi(x)$ and given independence $x^* \sim \pi(x)$.

given independence $x^* \sim \pi(x)$.

Summing it up, empirically you would do the following!

Algorithm (Hamiltonian Monte Carlo)

Choose $(X_0, U_0) = (x_0, u_0)$
For $t = 1, 2, \ldots$

1a. *Simulate* $U' \sim N(0, diag(m_i))$

1b. *Use the leap frog method (or any other method that results into an invertible G that is volume preserving) to generate a proposal* $(X^*, U^*) = G(x_{t-1}, U')$

2. *Simulate* $V \sim uniform(0, 1)$. *If* $V \leq a((x_{t-1}, U'), (X^*, U^*))$ *set* $X_t = X^*$, *otherwise* $X_t = x_{t-1}$

adjust via Metropolis-Hasting to keep invariance.

Notice finally that:

- T, the m_i's (mass), and the step size ε of the discretization in the leap frog method are tuning parameters

how long the ODE runs.

Notice that if you run the ODE too long a U-turn might occur (think of theoretical intuition of why it works!).

it works!

↳ There are extension of the above with just a little more computational effort that avoids such a U-turns.

↳ Stan, a probabilistic programming framework allows to work with Hamiltonian MCMC, without the need of specifying these hyperparameters and guaranteeing no U-turn will occur.

Reversible Jump MCMC

08 December 2020 12:31

So far we have seen how to simulate and approximate a distribution assuming that the parameter space, i.e. the dimensionality of the parameters, is known.

This section introduces a powerful technique for generalizing the above, and sampling from different parameters spaces.

Notice that

- This setting is related to the issue of **model selection**

Examples of this situation include:

- Mixture models where the number components k is not known
- Non-parametric regression models (*see below*)
- AR(p) models where the order p is not known
- Variable selection for regression models

Frequentist approach:

- Determining k: model selection problem } BIC, AIC, cross-validation LASSO etc...
- Determining x_k given k: estimation problem
  - E.g., maximize likelihood $p(y|k, x_k)$

The reversible jump algorithm discussed in this section is slightly different in that it rather takes a bayesian approach

Bayesian approach:

- Possible to do joint inference on k and x_k by considering the posterior $p(k, x_k|y)$

I.e. in this section we will specify the posterior in the following way and based on the observed samples and the priors we will then compute it:

- For a **Bayesian variable dimension model**, we specify a **prior distribution** on k and x_k:

$$p(k, x_k) = p(k)p(x_k|k)$$

i.e. assume that this is specified in such an hierarchical way.

  - Assume that $p(x_k|k)$ has a strictly positive density with respect to the n_k-dimensional Lebesgue measure on $\mathbb{X}_k$

- **Posterior** is given by $p(k, x_k|y) \propto p(y|k, x_k)p(k)p(x_k|k)$

- **Posterior** is given by $p(k, x_k|y) \propto p(y|k, x_k)p(k)p(x_k|k)$

The posterior $p(k, x_k|y)$ is a distribution on the union of the subspaces of different dimensions

$$\mathbb{X} = \bigcup_{k \in \mathcal{K}} \{k\} \times \mathbb{X}_k$$

We use the following **notation:**

- The posterior distribution $p(k, x_k|y)$ on $\mathbb{X}$ is denoted shortly by $\pi = p(k, x_k|y)$
- The restriction of π to $\mathbb{X}_k$ is denoted by $\pi_k = p(x_k|y, k)$. The following holds true

restriction

$$\pi_k(x_k) \propto \pi \propto p(y|k, x_k)p(k)p(x_k|k)$$

Reversible jump MCMC allows for sampling from the posterior $\pi = p(k, x_k|y)$

so from one state to the other we might jump in dimension

- The states of the Markov chain are of the form (k, x_k)
- The dimension of the of the states can vary
- From the output of a single Markov chain sampler, we can obtain both:
  - Posterior probability of each model $p(k|y)$
  - Posterior distributions of the individual models $p(x_k|y, k)$ } conditional posteriors.

The idea of the reversible jump MCMC is the usual:

How can we use the Metropolis-Hastings algorithm to sample from $\pi = p(k, x_k|y)$?

1. Propose transitions according to a kernel Q
2. Use an appropriate acceptance probability such that the target

2. Use an appropriate acceptance probability such that the target distribution π is **reversible**

↳ so here the idea is to transform the kernel Q to a reversible kernel P. Metropolis Hasting Step.

Notice moreover that to insure the desired properties for your chain such that it will converge to the equilibrium distribution you need irreducibility.

This can be guaranteed by:

- For **irreducibility**: need to propose transitions from $\{k\} \times \mathbb{X}_k$ to $\{l\} \times \mathbb{X}_j$ with $l \neq k$ (so called "jumps")

↳ and these should be proposed by the kernel Q.

We will now check how to construct the above:

In particular we will choose the transition kernel Q as follows:

Assume that the current state is (k, x) where $x = x_k \in \mathbb{X}_k$. A proposal $l, z)$ is then constructed as follows:

1. First, **choose the dimension index *l*** of the proposal according to a **stochastic matrix (β_{kl})**, where β_{kl} denotes the probability that l is selected
2. Then, consider transitions from $\mathbb{X}_k$ to $\mathbb{X}_l$ of the type

$$x \to \mathbf{z} = \mathbf{z}(\mathbf{x}, \mathbf{U_{kl}}),$$

where

- $\boldsymbol{U_{kl}}$ is a d_{kl} dimensional random variable with strictly positive **density f_{kl}**
- The **relationship $z = z(x, u_{kl})$** is assumed to be **deterministic**

So U_{kl} is a stochastic RV that augments/shrinks the parametric dimension to the proposal l.

Notice now that the following holds true

Let $\boldsymbol{Q_{kl}(x, dz)}$ **denote the corresponding transition kernel.** The following holds true:

- The distribution $\pi_k(dx)Q_{kl}(x, dz)$ is concentrated on a $(n_k + d_{kl})$-dimensional surface in $\mathbb{R}^{n_k+n_l}$ of the form $(x, z(x, u_{kl}))$
- The density of (x, u_{kl}) is equal to $\pi_k(x)f_{kl}(u_{kl})$

Notice: concentrated as you might well augment in 1 dimension of $\mathbb{R}^{n_k+n_l}$ keeping all of the others dimensions in the space

dimensions in the space constant.

As our goal is now to construct a reversible chain such that we know that the chain will converge to its steady-state equilibrium distribution, we need the following:

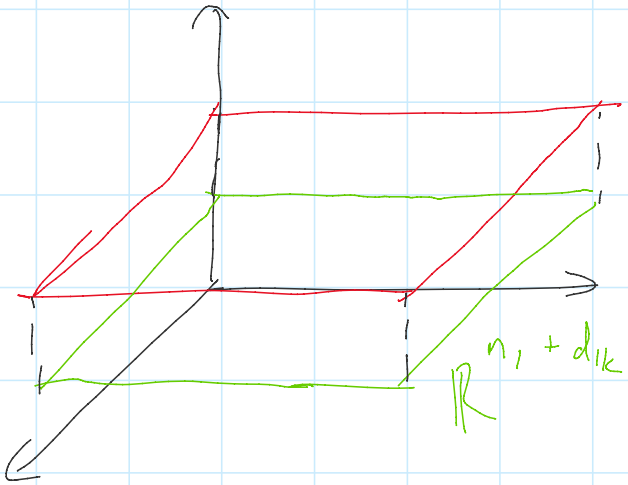
In order to guarantee **reversibility** (Theorem 4.2), we must also allow a transition $Q_{lk}(z, dx)$ from $\mathbb{X}_l$ to $\mathbb{X}_k$ with the property that **$\pi_l(dz)Q_{lk}(z, dx)$ is concentrated on the same surface** as $\pi_k(dx)Q_{kl}(x, dz)$.

Notice that the concentration on the same surface is quite intuitive.

⇒ If this is not the case you will never reach equivalence between $\pi_k(dx)\,Q(x, dz)$ and $\pi_l(dz)\,Q(z, dx)$ so that you cannot construct a

so that you cannot construct a reversible kernel. This due to the fact that they will not share the same Null-Set

↳ Consider for instance



$\mathbb{R}^{n_k + n_1} = \mathbb{R}^3$ in this case

complete different concentration and Null-set.

If however, the above holds true, then we immediately know from Theorem 4.2:

Theorem (4.2)

Let π be a probability on $(\mathbb{X}, \mathcal{F})$ and Q be a kernel in the same space such that the ***two probabilities*** $\pi(dx)Q(x, dy)$ ***and*** $\pi(dy)Q(y, dx)$ ***on*** $(\mathbb{X}, \mathcal{F}) \times (\mathbb{X}, \mathcal{F})$ ***are equivalent*** *in the sense of measure theory, i.e., the two probabilities have the same null sets. The following hold true.*

- *The Radon-Nikodym density of $\pi(dy)Q(y, dx)$ with respect to $\pi(dx)Q(x, dy)$ exists. We denote it by $r(y, x)$.*
- *The following kernel is reversible regarding π:*

$$P(x, A) = \int a(x, y)Q(x, dy) + \mathbf{1}_A(x) \cdot \left(1 - \int a(x, y)Q(x, dy)\right),$$

- The following kernel is reversible regarding π:

$$P(x, A) = \int_A a(x, y)Q(x, dy) + \mathbf{1}_A(x) \cdot \left(1 - \int_X a(x, y)Q(x, dy)\right),$$

where

$$a(x, y) = \min(1, r(y, x)).$$

Such that we immediately know how to compute the reversible kernel via the Metropolis-Hasting step.

All that is left is computing the Radon-Nikodym Derivative.

We note now that in order to have equivalence the following must hold.

We assume that $Q_{lk}(z, dx)$ is constructed in the same way as $Q_{kl}(x, dz)$:

$$z \to x = x(z, U_{lk}),$$

where

- U_{lk} is a d_{lk} dimensional random variable with strictly positive density f_{lk}
- the relationship $x = x(z, U_{lk})$ is assumed to be deterministic

It follows that

- The distribution $\pi_l(dz)Q_{lk}(z, dx)$ is concentrated on a $(n_l + d_{lk})$-dimensional surface in $\mathbb{R}^{n_k+n_l}$ of the form $(z, x(z, u_{lk}))$

Given this and the definition of Q_{kl} it follows:

In order that $(x, z(x, u_{kl}))$ and $(z, x(z, u_{lk}))$ are on the same surface and that the Radon-Nikodym density of $\pi_l(dz)Q_{lk}(z, dx)$ w.r.t. $\pi_k(dx)Q_{kl}(x, dz)$ exists, the following two conditions need to hold true:

1. The dimensions of the two surfaces must match

$$n_k + d_{kl} = n_l + d_{lk}$$

As if one in $\mathbb{R}^2$ and other in $\mathbb{R}^3$ straightforward to see that Will-set will be different.

2. There is a diffeomorphism* T_{kl} between (x, u_{kl}) and (z, u_{lk}):

$$(z, u_{lk}) = T_{kl}(x, u_{kl})$$

i.e. the subspaces.

↳ notice a diffeomorphisms is defined as a function for which both itself and its inverse are differentiable.

↳ Given this it is straightforward to see why this is necessary as if if such function is not differentiable in any of the two directions then there are jumps and the surface might not

jumps and the surface might not be the same.

Given this, and the construction of kernels Q_{kl} and Q_{lk} such that this is satisfied it is now possible to derive the Radon-Nikodym density of $\pi(dz)\, Q_{kl}(x, dz)$ w.r.t. $\pi(dz)\, Q(z, dx)$. as follows through the change-of-variables formula:

Assume a diffeomorphism

$$T : \mathbb{R}^n \to \mathbb{R}^n$$

i.e. $$x \longrightarrow T(x) = y$$

then

change of variables

$$\int h(y)\, f_Y(y)\, dy = \int h(T(x)) \cdot f_Y(T(x)) \cdot \left| \det \frac{\partial T(x)}{\partial x} \right| dx$$

det of Jacobi matrix !

$$= \int h(T(x)) \cdot \frac{f_Y(T(x))}{f_X(x)} \left| \det \frac{\partial T(x)}{\partial x} \right| f_X(x)\,dx$$

↳ any bounded function $h(\cdot)$, might also be an indicator function.

Radon - Nikodym density; transforms $f_Y(y)dy$ into $f_X(x)dx$ density.

It is then straightforward to see that inserting our variables of interest, i.e.

$$T(x) = T_{kl}(x, z(x, v_{kl}))$$

$$x = (x, v_{kl})$$

$$y = (z, v_{lk})$$

$$f_X = \pi_k(x) \cdot f_{kl}(v_{kl})$$

$$f_Y = \pi_l(z) \quad f_{lk}(v_{lk})$$

we get:

$$\pi_l(z) \cdot f_{kl}(v_{kl}) \cdot \left| \det \frac{\partial T_{kl}(x, v_{kl})}{\partial x} \right|$$

$$r(x,y) = \frac{\pi_l(z)}{\pi_k(x)} \cdot \frac{f_{kl}(u_{kl})}{f_{lk}(u_{lk})} \left| \det \frac{\partial T_{kl}(x, u_{kl})}{\partial (x, u_{kl})} \right|$$

So notice that in summary:

- In summary, we obtain the following proposal transition kernel from $\mathbb{X}$ into itself:

$$Q((k,x), d(l,z)) = \sum_{j=0}^{\infty} \beta_{kj} Q_{kj}(x, dz) \mathbf{1}_j(l)$$

- The acceptance probabilities are then

$$a((k,x),(l,z)) = \min\left(1, \frac{\pi_l(z)\beta_{lk} f_{lk}(u_{lk})}{\pi_k(x)\beta_{kl} f_{kl}(u_{kl})} \left| \frac{\partial T_{kl}(x, u_{kl})}{\partial (x, u_{kl})} \right| \right)$$

{ so same as derived augmented with the transition probabilities β_{lk} an β_{kl} respectively.

So generally it is possible to sample both k and x_k concurrently converging to the true distribution $\underline{k}$ with the following reversible jump algorithm

Algorithm (reversible jump algorithm)

Choose an initial value (k_0, x_0).
For $t = 1.2\ldots$.

Algorithm (reversible jump algorithm)

Choose an initial value (k_0, x_0).
For $t = 1, 2, \ldots$

- *Denote the current value by* $(k, x) = (k_{t-1}, x_{t-1})$.

1. *Choose* l *with probability* β_{kl}.
2. *Generate* $u_{kl} \sim f_{kl}(\cdot)$.
3. *Set* $(z, u_{lk}) = T_{kl}(x, u_{kl})$ } so have to specify this diffeomorphism
4. *Accept* $(k_t, x_t) = (l, z)$ *with probability*

$$\min\left(1, \frac{\pi_l(z)\beta_{lk}f_{lk}(u_{lk})}{\pi_k(x)\beta_{kl}f_{kl}(u_{kl})}\left|\frac{\partial T_{kl}(x, u_{kl})}{\partial(x, u_{kl})}\right|\right),$$

and $(k_t, x_t) = (k_{t-1}, x_{t-1})$ *otherwise.*

- The transformation T_{kl} can be difficult to construct
- Tuning of a reversible jump algorithm can be difficult } in terms of how long does it have to run in order to sample whole space

* Have still to write down the application of it.

Accuracy of MCMC

14 December 2020 / 10:03

This section explores accuracy and convergence properties of MCMC.

This is a difficult task due to the following:

- Our goal is to estimate

$$\theta = \int h(x)\pi(dx)$$

- We do this by generating a Markov chain $X_1, X_2, X_3, \ldots$ and use the following estimator

$$\hat{\theta}_N = \frac{1}{N}\sum_{t=1}^{N} h(X_t)$$

Difficulty for accuracy results:

1. The random variables $X_1, X_2, X_3, \ldots$ are dependent
2. The $X_t's$ are not identically distributed since X_t has the distribution π only asymptotically for $t \to \infty$

From this it follows that we cannot use the standard CLT

cannot use the standard CLT results for independent and identically distributed samples and the analysis of accuracy metrics for such methods is rough.

i.e. in general

Bias

- We make a systematic error:

$$\mathbb{E}(\hat{\theta}_N) = \boxed{\frac{1}{N}} \sum_{t=1}^{N} \mathbb{E}(h(X_t)) \neq \theta$$

- This bias is of the order $O(1/N)$, provided

$$\sum_{t=1}^{\infty} |\mathbb{E}(h(X_t)) - \theta| < \infty$$

Moreover, observe the following:

- The variance is given by

$$\mathrm{Var}(\hat{\theta}_N) = \frac{1}{N^2} \sum_{s=1}^{N} \sum_{t=1}^{N} \mathrm{Cov}(h(X_s), h(X_t))$$

- The mean square error (MSE) is

- The mean square error (MSE) is

$$\mathbb{E}\left((\hat{\theta}_N - \theta)^2\right) = \underbrace{\left(\mathbb{E}(\hat{\theta}_N) - \theta\right)^2}_{\text{Bias}^2} + \text{Var}\left(\hat{\theta}_N\right)$$

- For an error bound, we need to estimate both the bias and the variance. Both tasks are rather difficult, in general

We can then see that the following holds:

- We will see that typically the variance of $\hat{\theta}_N$ is still of order $O(1/N)$
- Since the bias is of order $O(1/N)$, its contribution to the mean square error is asymptotically negligible

↳ as entering in a squared way in the MSE.

Notice now that it is possible to try to filter out the bias in the following way

bias in the following way

- For the bias, we are often satisfied with graphical tools, e.g., so called **trace plots** of $h(X_t)$ versus t
  - Try to find a time t_0 called the **burn-in** time after which systematic deviations no longer occur
  - Use only the values after iteration t_0
  - Run several chains with different initial values to verify that the chains reach the stationary distribution

So that the idea is to tackle the two issues for computing the accuracy of MCMC methods separately.

→ for the identical distribution

↳ try to tackle it taking into account the burn-in period so that you would ultimately sample from an identical distribution

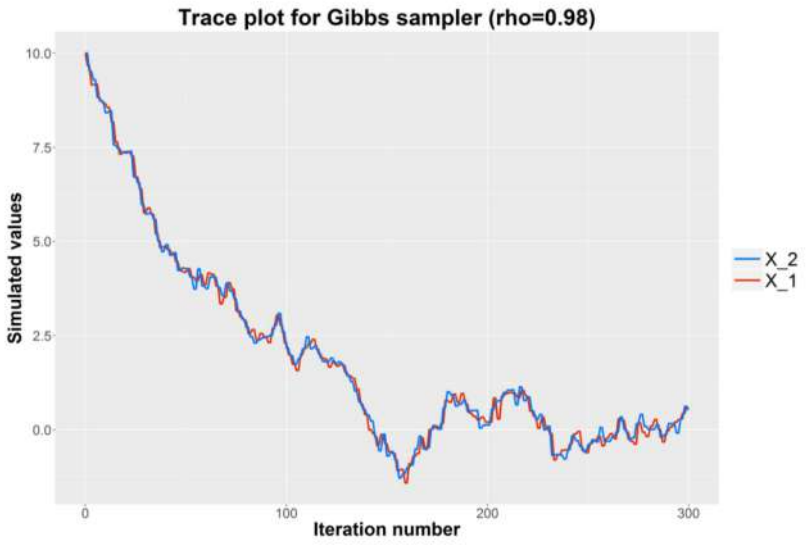
→ Once you sample from an identical distribution, there are CLT versions for dependent samples.

↳ Use these to calculate accuracy of your estimator.

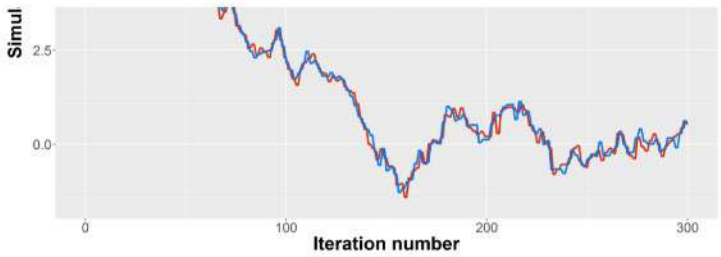
Example of burn-in period:

Example of trace plot

Gibbs sampler for bivariate normal distribution



⇒ Guess burn-in time of approx. $t_0 = 200$



$\Rightarrow$ Guess burn-in time of approx. $t_0 = 200$

A more rigorous mathematical approach for the bias would be to inspect the convergence speed of the Markov Chain X_t to the stationary distribution, i.e the converge of our sample estimate $\mathbb{E}(h(X_t))$ to the true estimate $\int h(x)\,\pi(dx)$.

- Let (X_t) be a Markov chain with initial distribution ν, transition P and invariant distribution π
- **Goal:** Estimate how quickly

$$\mathbb{E}(h(X_t)) - \int h(x)\pi(dx) = \int P^t h(x)\nu_0(dx) - \int P^t h(x)\pi(dx)$$

converges to zero

notice can also leave this out here as π is the stationary distribution.

- In general, very difficult without particular conditions on h and the proposal kernel Q

- In general, very difficult without particular conditions on π and the proposal kernel Q
- We restrict ourselves to the case where π is a probability on the discrete space $\{1, 2, ...n\}$

Then

- If $\mathbb{X}$ is discrete, we have

$$\left|\mathbb{E}(h(X_t)) - \int h(x)\pi(dx)\right| = \left|\sum_j (\nu_0 P^t(j) - \pi P^t(j))h(j)\right|$$
$$\leq \max_i |h(i)| \sum_j |\nu_0 P^t(j) - \pi P^t(j)|$$

- It is sufficient to bound the L_1-distance

$$||\nu_0 P^t - \pi P^t||_1 = \sum_j |\nu_0 P^t(j) - \pi P^t(j)|$$

This can be done in two ways that we will explore next:

1. An algebraic Way

2. A [illegible]

(2) A stochastic method based on the idea of coupling Markov Chains

(1) On the algebraic approach:

Algebraic approach

- The Frobenius theorem states that the eigenvalue with the largest absolute value of a stochastic irreducible and aperiodic matrix equals 1 and its multiplicity is 1

⇒ write down definition, see later in the chapter

- The convergence speed is determined by the eigenvalue of the transition matrix P with the second largest absolute value

(2) On the stochastic Approach

Coupling of Markov chains

- Construct a Markov process $(X_t^{(1)}, X_t^{(2)})$ on the state space $(1, 2, \ldots n)^2$ such that:
  - Marginally, $(X_t^{(1)})$ and $(X_t^{(2)})$ are both Markov chains with transition matrix P and initial distributions μ and ν

Product space, i.e. doubled

space, i.e. doubled dimension.

- Marginally, $(X_t^{(1)})$ and $(X_t^{(2)})$ are both Markov chains with transition matrix P and initial distributions μ and ν
- The two chains are dependent: they stay together after they have met for the first time:

$$\text{if } X_t^{(1)} = X_t^{(2)} \text{ for some } \Rightarrow X_s^{(1)} = X_s^{(2)} \ \forall s > t$$

- How we make the transition as long as $X_{t-1}^{(1)} \neq X_{t-1}^{(2)}$ is left open

Mathematically:

- Denote by ***Q* transition matrix for the coupled process**

$$Q(i,j;k,l) = \mathbb{P}(X_t^{(1)} = k, X_t^{(2)} = l | X_{t-1}^{(1)} = i, X_{t-1}^{(2)} = j)$$

- The following must hold true

$$\begin{aligned} \sum_l Q(i,j;k,l) &= P(i,k) \text{ for all } i \neq j, k \\ \sum_k Q(i,j;k,l) &= P(j,l) \text{ for all } i \neq j, l \end{aligned}$$

$$\begin{aligned} Q(i,i;k,k) &= P(i,k), \\ Q(i,i;k,l) &= 0 \text{ if } k \neq l \end{aligned}$$

- There are still many choices for Q, the easiest one being $Q(i,j;k,l) = P(i,k)P(j,l)$ if $i \neq j$ (independence as long as the chains have not met)

Given such construction the following Lemma holds:

Lemma

Lemma

For any coupling satisfying the above properties,

$$||\nu P^t - \mu P^t||_1 \leq 2\mathbb{P}(X_t^{(1)} \neq X_t^{(2)}).$$

Moreover, there is a coupling such that

$$\mathbb{P}(X_t^{(1)} \neq X_t^{(2)}) \leq \alpha^t \mathbb{P}(X_0^{(1)} \neq X_0^{(2)})$$

where

$$\alpha = \frac{1}{2} \max_{i,j} ||P(i,.) - P(j,.)||_1.$$

Such that putting it all together you can compute a bound for

$$\| \nu P^t - \mu P^t \|_1$$ norm,

and such that you have a bound for the convergence speed of your sample estimator, i.e. a bound for the bias.

Given such an analysis we turn to an investigation of the properties of the Estimator variance. We restrict ourself on an analysis of the variance in the case of sampling from a stationary distribution, i.e. after the burn-in samples have been removed.

- Assume that that $(X_1, X_2, ..., X_k)$ and $(X_{i+1}, X_{i+2}, ..., X_{i+k})$ have the same distribution for all i and for all k, i.e., (X_t) is **stationary**
- If (X_i) is a Markov chain with a transition kernel that does not depend on time, then we have **stationarity if and only if** $X_1 \sim \pi$ (π is the stationary distribution)

↓

i.e. the first must already come from the invariant, stationary distribution.

Given now the sampling from a

Given now the sampling from a stationary distribution, the following holds:

Lemma

Let (X_i) be stationary, $Y_i = h(X_i)$, and $R(k) = Cov(Y_i, Y_{i+k})$. Then

1. $$Var\left(\frac{1}{N}\sum_{i=1}^{N} Y_i\right) = \frac{1}{N}\sum_{k=-N+1}^{N-1}\left(1 - \frac{|k|}{N}\right) R(k).$$

2. *If $\sum_{k=1}^{\infty} |R(k)| < \infty$, then as $N \to \infty$*

$$N Var(\hat{\theta}_N) \to \sigma_\infty^2 = \sum_{k=-\infty}^{\infty} R(k) = Var(Y_1) + 2\sum_{k=1}^{\infty} R(k).$$

(due to symmetry; $Var(Y_1)$ if $k=0$)

3. *If $\sum_{k=1}^{\infty} |R(k)| < \infty$, then*

$$Corr\left(\frac{1}{N}\sum_{i=1}^{N} Y_i, \frac{1}{N}\sum_{i=N+1}^{2N} Y_i\right) \to 0.$$

(i.e. two blocks that follows are indep.)

Proof of the Lemma

(1.) $$Var\left(\frac{1}{N}\sum_{i=1}^{N} Y_i\right) = \frac{1}{N^2} Var\left(\sum_{i=1}^{N} Y_i\right)$$

$$= \frac{1}{N^2}\sum_{i,j=1}^{N} Cov(Y_i, Y_j)$$

$$= \frac{1}{N^2}\sum_{k=-N+1}^{N-1} (N - |k|) R(k)$$

$$= \frac{1}{N^2} \cdot \sum_{k=-N+1}^{N-1} R(k) \cdot \underbrace{(N-|k|)}_{= \#\text{ of comb for which diff } (j-i)=k}$$

(2.) From (1) we know that

$$Var(\hat{\theta}_N) = \frac{1}{N^2} \cdot \frac{N}{N} \sum_{k=-N+1}^{N-1} R(k) \, (N-|k|)$$

$$N \, Var(\hat{\theta}_N) = \sum_{k=-N+1}^{N-1} R(k) \left(1 - \frac{|k|}{N}\right)$$

$$\leq \sum_{k=-\infty}^{\infty} \max\left(0, 1 - \frac{|k|}{N}\right) R(k)$$

$\max\left(0, 1 - \frac{|k|}{N}\right) \leq 1$

With that it is easy to see that

1. $1 - \frac{|k|}{N} = 1$ [illegible] that

$$\lim_{N\to\infty} 1 - \frac{|k|}{N} = 1 \quad \text{so that}$$

$$\lim_{N\to\infty} N \operatorname{Var}\left(\hat{\theta}_N\right) = \lim_{N\to\infty} \sum_{-\infty}^{\infty} \max\left(0, 1 - \frac{|k|}{N}\right) \cdot R(k)$$

$$= \sum_{-\infty}^{\infty} \lim_{N\to\infty} \max\left(0, 1 - \frac{|k|}{N}\right) \cdot R(k)$$

"lebesgue dominated convergence apply so that you can bring the lim inside the "integral" sum

$$= \sum_{-\infty}^{\infty} R(k) := \sigma_\infty^2$$

(3.) We have to show that if $\sum_{k=1}^{\infty} R(k) < \infty$ then

$$\lim_{N\to\infty} \operatorname{Corr}\left(\frac{1}{N}\sum_{i=1}^{N} Y_i, \frac{1}{N}\sum_{j=N+1}^{2N} Y_j\right) = 0$$

Notice that this means:

Notice that this means:

$$\lim_{N\to\infty} \frac{\mathrm{Cov}\left(\sum_{i=1}^{N} Y_i, \sum_{j=N+1}^{2N} Y_j\right)}{\sqrt{\mathrm{Var}\left(\sum_{i=1}^{N} Y_i\right)}\sqrt{\mathrm{Var}\left(\sum_{j=N+1}^{2N} Y_j\right)}} = 0$$

Notice here $\frac{1}{N}$ omitted as same holds.

we know from (1.) that these two are asymptotically equal.

Notice now:

$$\mathrm{Var}\left(\frac{1}{N}\sum Y_i\right) = \frac{1}{N^2}\,\mathrm{Var}\left(\sum Y_i\right)$$

So that:

$$\mathrm{Var}\left(\sum Y_i\right) = N^2 \cdot \mathrm{Var}\left(\hat{\theta}_N\right)$$

$$= N \cdot \underbrace{N \cdot \mathrm{Var}\left(\hat{\theta}_N\right)}_{\sigma_\infty^2 + o(1)}$$

such that $\lim N\,\mathrm{Var}(\hat{\theta}_N) - \sigma_\infty^2 = 0$

$\sigma_\infty^2 + o(1)$ such that $\lim N \mathrm{Var}(\hat{\theta}_N) - \sigma_\infty^2 = 0$

So that generally

$$\sqrt{\mathrm{Var}\left(\sum Y_i\right)} \cdot \sqrt{\mathrm{Var}\left(\sum Y_j\right)} = N\sigma_\infty^2 + o(1)$$

Moreover for the Nominator:

$$\mathrm{Cov}\left(\sum_{i=1}^{N} Y_i, \sum_{j=1+N}^{2N} Y_j\right) = \left|\sum_{k=1}^{2N-1} \underbrace{\min(k, 2N-k)}_{\text{number of occurrences}} R(k)\right|$$

$f(x) =$

function always

$$\leq \sqrt{N} \sum_{k=1}^{\sqrt{N}} |R(k)| + N \cdot \sum_{k=\sqrt{N}+1}^{\infty} |R(k)|$$

$$\leq \sqrt{N} \sum_{k=1}^{\infty} |R(k)| +$$

function always smaller N

function smaller than $\sqrt{N}$

$k=1$

$+N \sum_{k=\sqrt{N}+1}^{\infty} |R(k)|$

It is the clear that , both are of order $o(N)$, for the first is clear as

$\lim_{N\to\infty} \frac{\sqrt{N}}{N} \sum_{k=1}^{\infty} |R(k)| = 0$, for the second it follows as it is a convergent series.

It follows therefore all in all

$$\lim_{N\to\infty} Cov\left(\sum_{i=1}^{N} Y_i , \sum_{j=N+1}^{2N} Y_j\right) = \frac{o(N)}{N\sigma_N^2 + o(1)} = 0$$

$$\lim_{N\to\infty} \mathrm{Corr}\left(\sum_{i=1}^{N} Y_i \,,\, \sum_{j=N+1}^{2N} Y_j\right) = \frac{o(N)}{N\sigma_N^2 + o(1)} = 0$$

as $o(N)$ grows less rapidly than denominator N.

Given the above and the assumption that $\sum |R(k)| < \infty$, we can then use the Chebyshev Inequality to compute the probability for an error bound ε, i.e.

$$P_r\left(|\hat{\theta}_N - \theta| > \varepsilon\right) \le \frac{Var(\hat{\theta}_N)}{\varepsilon^2} \approx \frac{\sigma_\infty^2}{N\varepsilon^2}$$

Important Note: Often such bound is not sharp, i.e. it

not sharp, i.e. it is not of practical relevance, for instance right hand side > 1.

For such reason we are rather interested on a CLT theorem.

Aside from this two other questions arise in relation to such Chebyshev bound:

1. When is $\sum |R(k)| < \infty$?
2. How can we estimate σ^2_∞?
3. Does a central limit theorem hold?

When is $\sum |R(k)| < \infty$?

was the case

- What matters is how quickly $P^t h(x) - \theta$ goes to zero, as in the analysis of the bias

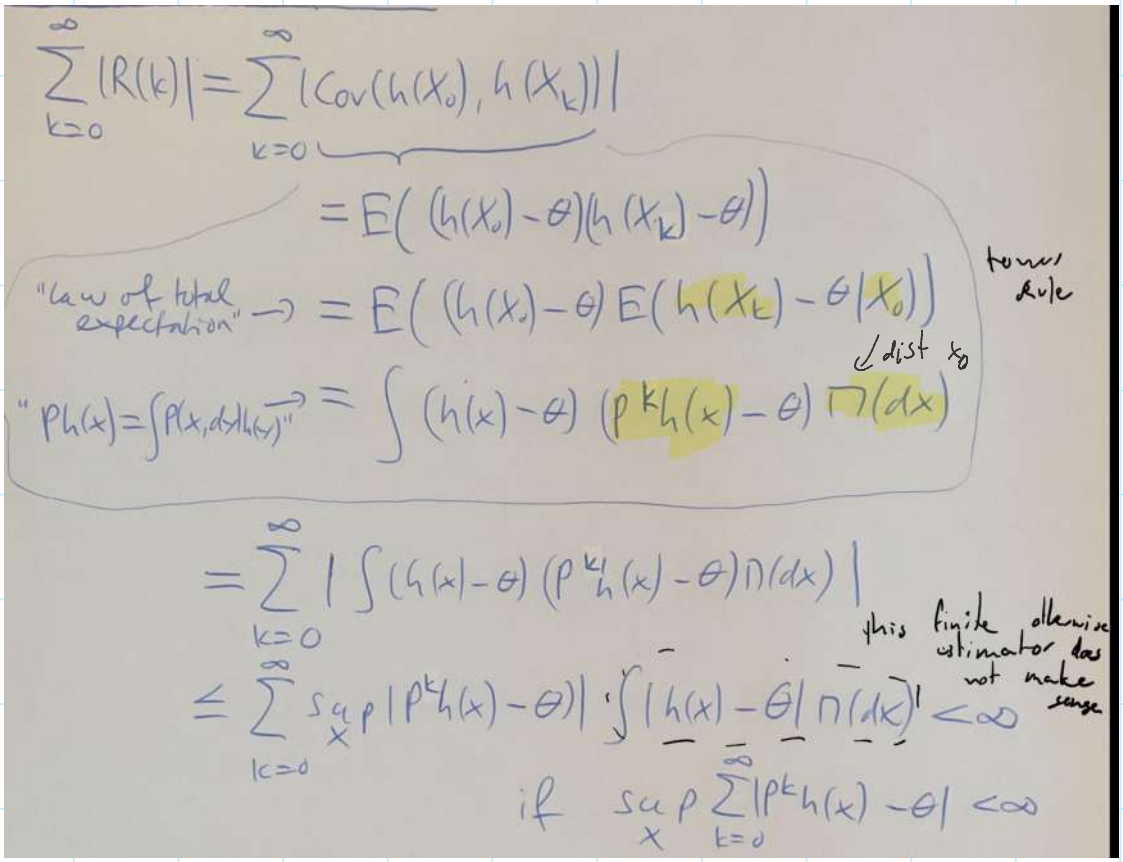
analysis of the bias

- In particular, the following condition is sufficient:

$$\sup_x \sum_t |P^t h(x) - \theta| < \infty$$

} quite strict condition.

↳ Proof



↳ i.e. sum of all individual bias must be bounded.

How can we estimate σ_∞^2?

How can we estimate σ_∞^2?

- Use

$$\hat{R}(k) = \frac{1}{N} \sum_{i=1}^{N-|k|} (Y_i - \hat{\theta}_N)(Y_{i+|k|} - \hat{\theta}_N)$$

- And

$$\hat{\sigma}_\infty^2 = \sum_{k=-m}^{m} w_k \hat{R}(k)$$

(weighted average of sample autocovariance)

where

- w_k are symmetric weights with $w_0 = 1 \geq w_1 \geq \ldots \geq w_{m+1} = 0$
- In theory, we should have $m \to \infty$ and $m = o(N)$
- In practice, $m \approx N^{1/3}$ is often a sensible choice

Notice you should not have $w_k = 1$ $\forall k$ as then it is possible to show

$$\sum_{k=-N+1}^{N-1} 1 \cdot \hat{R}(k) = \left(\sum_{i=1}^{N} (Y_i - \hat{\theta}_N)^2 \right) = 0$$

so not suitable to estimate σ_∞^2

Does a central limit theorem hold?

- There is a large literature concerning central limit theorems for

- There is a large literature concerning central limit theorems for stationary random variables
- One of the simplest and most important result for Markov chain Monte Carlo is the following:

  If (X_i) is an irreducible and aperiodic Markov chain with transition kernel P and π is reversible, then $\frac{1}{N}\sum h(X_i)$ is asymptotically normal if $\sum_k |R(k)| < \infty$

Aperiodicity

- An m-cycle for an irreducible chain with kernel P is a collection of disjoint sets $\{E_0, \ldots, E_{m-1}\}$ such that $P(x, E_j) = 1$ for $j = i + 1 \bmod m$ and all $x \in E_i$

  3 periodically jumping from one set to the other.
- The period d of a chain is the largest m for which an m-cycle exist
- A chain with transition kernel P is called aperiodic if $d = 1$

  ↳ are then can just jump between two sets

  the larger m the more sets jumps are possible.

↳ Easy to verify then that many Markov Chains

↳ Easy to verify then that many Markov Chains display such behaviour.

↳ An example might be for instance RW-Metropolis Hasting MCMC.

Such that:

Confidence interval for θ

- A straightforward choice for a confidence interval is

$$\hat{\theta}_N \pm \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\frac{1}{\sqrt{N}}\hat{\sigma}_\infty$$

} usual MC-CI

- Another possibility is the so-called **batch means** method:
  1. Compute the means of b consecutive Y_i's:

$$\hat{\theta}_{i,b} = \frac{1}{b}\sum_{j=(i-1)b+1}^{ib} Y_j$$

  2. Consider the means $\hat{\theta}_{i,b}$, $i = 1, 2, ..., k = N/b$ as independent and normally distributed → see above
  3. Use the following confidence interval

$$\hat{\theta}_N \pm \frac{1}{\sqrt{k}} t_{k-1,1-\frac{\alpha}{2}} \sqrt{\frac{1}{k-1}\sum_{i=1}^{k}(\hat{\theta}_{i,b}-\hat{\theta}_N)^2}$$

→ t-dist correct for estim. of variance

→ after burn-in as is the case here.

- **Advantage**: σ_∞ does not need to be estimated.
  Disadvantage: The choice of b can be difficult