An Introduction to MCMC

Jascha Schewtschenko

Institute of Cosmology and Gravitation, University of Portsmouth

July 24, 2019

Outline

Homework (Data Collection)

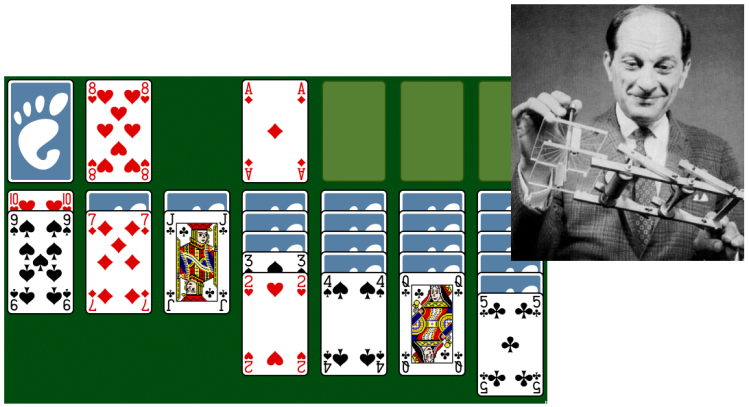
- What do you belief is the share of winnable deals?

Homework (Data Collection)

- What do you belief is the share of winnable deals?
- How many deals have you played? How many won?

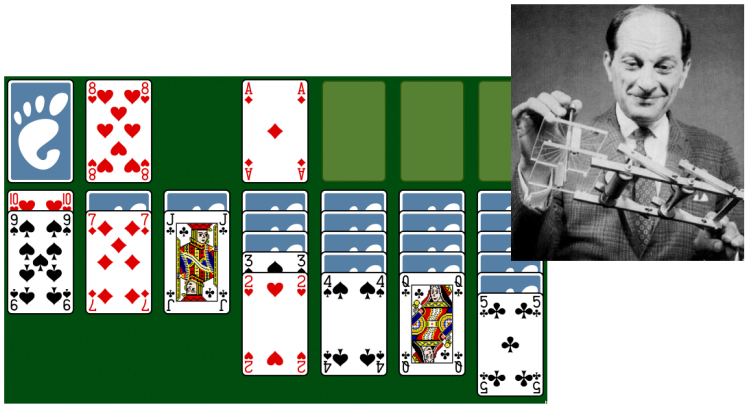
MARCOV CHAIN **MONTE CARLO** SAMPLING

Motivation



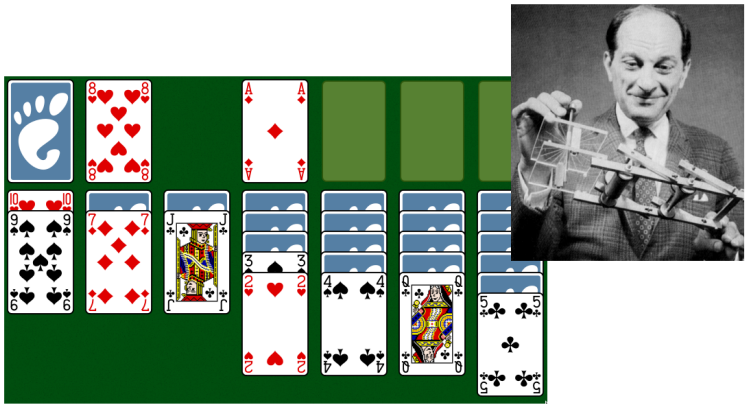
- In 1946, while recovering from a brain surgery, Stanislav Ulam, who worked on the Manhattan Project (cf. Teller-Ulam design) & playing a solitaire card game, started wondering what the chances of winning are.

Motivation



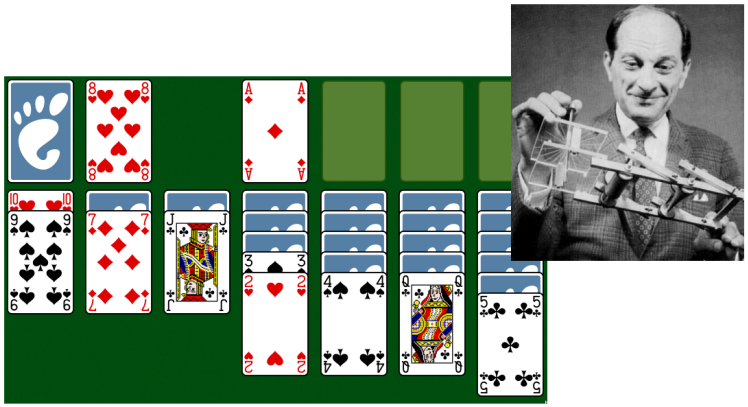
- In 1946, while recovering from a brain surgery, Stanislav Ulam, who worked on the Manhattan Project (cf. Teller-Ulam design) & playing a solitaire card game, started wondering what the chances of winning are.
- Given the huge amount of possible moves (cards dealt & played), the problem turned out to be very challenging to be solved purely combinatorically.

Motivation



- In 1946, while recovering from a brain surgery, Stanislav Ulam, who worked on the Manhattan Project (cf. Teller-Ulam design) & playing a solitaire card game, started wondering what the chances of winning are.
- Given the huge amount of possible moves (cards dealt & played), the problem turned out to be very challenging to be solved purely combinatorically.
- Ulam had the idea to simply count the number of won and lost games instead. He reasoned that with enough games, the ratio of won vs played games should become a good approximation of the true chance of winning.

Motivation



- In 1946, while recovering from a brain surgery, Stanislav Ulam, who worked on the Manhattan Project (cf. Teller-Ulam design) & playing a solitaire card game, started wondering what the chances of winning are.
- Given the huge amount of possible moves (cards dealt & played), the problem turned out to be very challenging to be solved purely combinatorically.
- Ulam had the idea to simply count the number of won and lost games instead. He reasoned that with enough games, the ratio of won vs played games should become a good approximation of the true chance of winning.

Homework (Analysis)

- Our estimated ratio of winnable games is $f_{\text{winnable}} = W/N = \ldots$

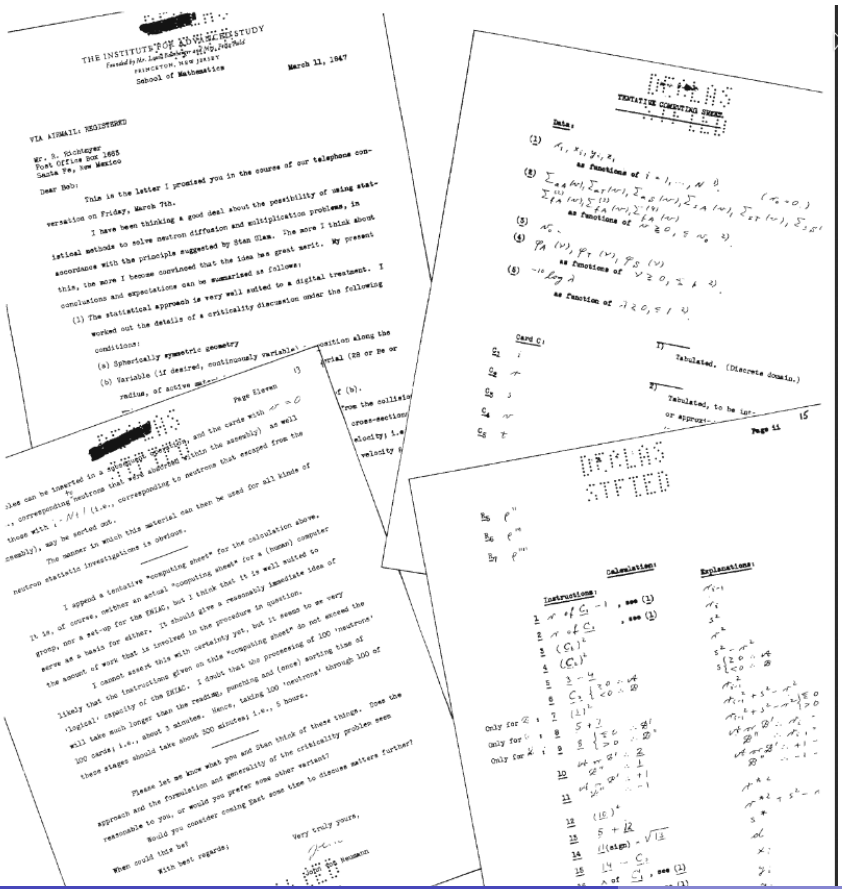
Homework (Analysis)

- Our estimated ratio of winnable games is $f_{\text{winnable}} = W/N = \ldots$
- Win-Lose scenario with fixed probability follows a **binomial distribution**. Hence, there is a 68% (95%) chance that the true probability lies within 1 (2) σ of our estimate with:

$$\sigma^2 = f_{\text{winnable}}(1 - f_{\text{winnable}})/N = \ldots$$

From cards to nukes

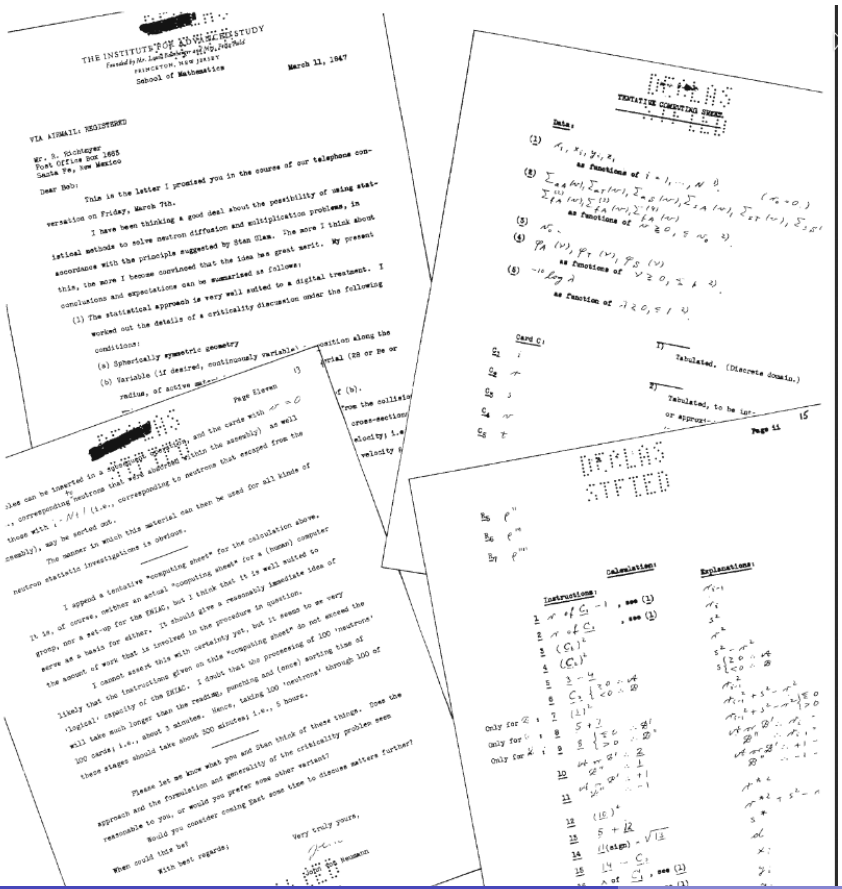
John von Neumann



- Ulam told his colleague John von Neumann, a pioneer in computer science (besides many other fields of expertise) about his idea.

From cards to nukes

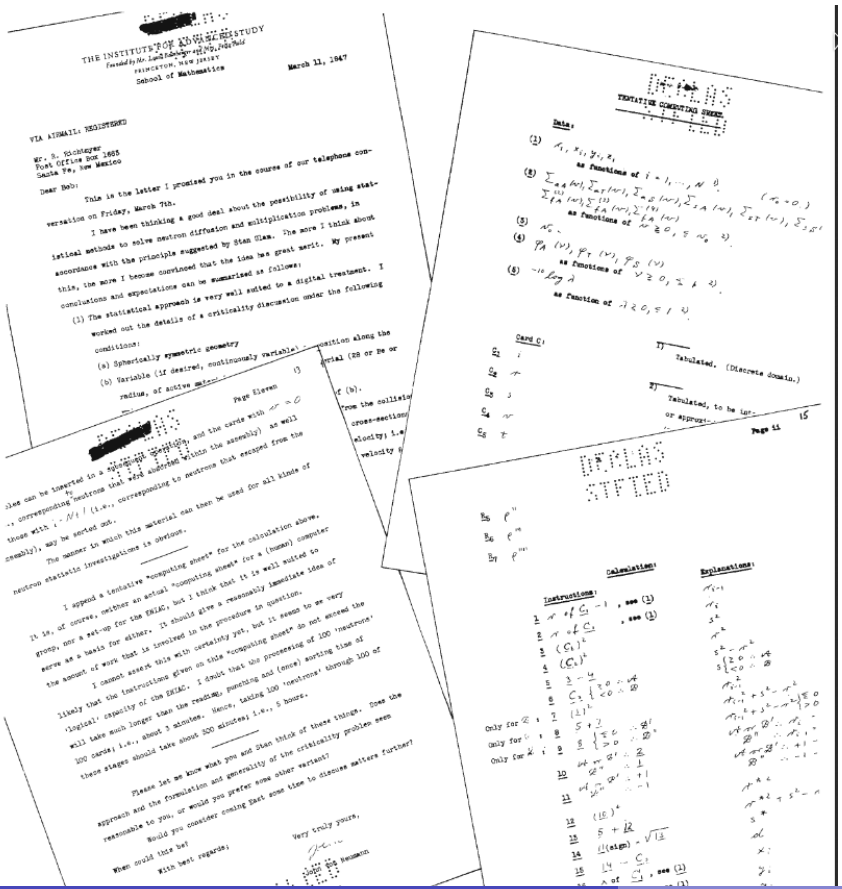
John von Neumann



- Ulam told his colleague John von Neumann, a pioneer in computer science (besides many other fields of expertise) about his idea.
- Von Neumann realized the potential of using it in combination with the new "electronic computing" and proposed a program to apply it to neutron scattering.

From cards to nukes

John von Neumann

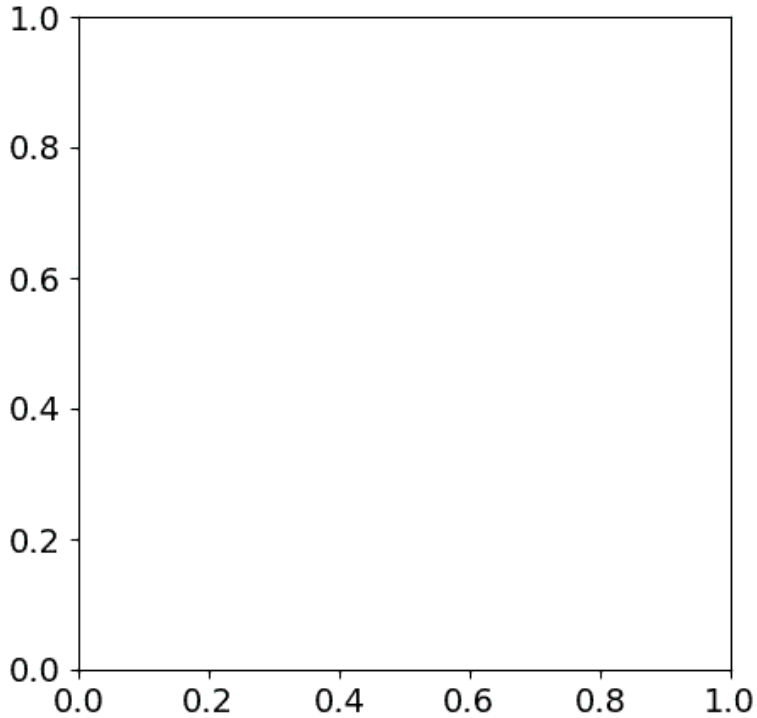


- Ulam told his colleague John von Neumann, a pioneer in computer science (besides many other fields of expertise) about his idea.
- Von Neumann realized the potential of using it in combination with the new "electronic computing" and proposed a program to apply it to neutron scattering.
- Instead of solving the problem statistically for the whole assemble of neutrons, he proposed to follow a subset of neutrons and decide randomly the outcome of events those neutron face (e.g. either fission, scattering or absorbtion)

Monte Carlo Sampling

Algorithm:

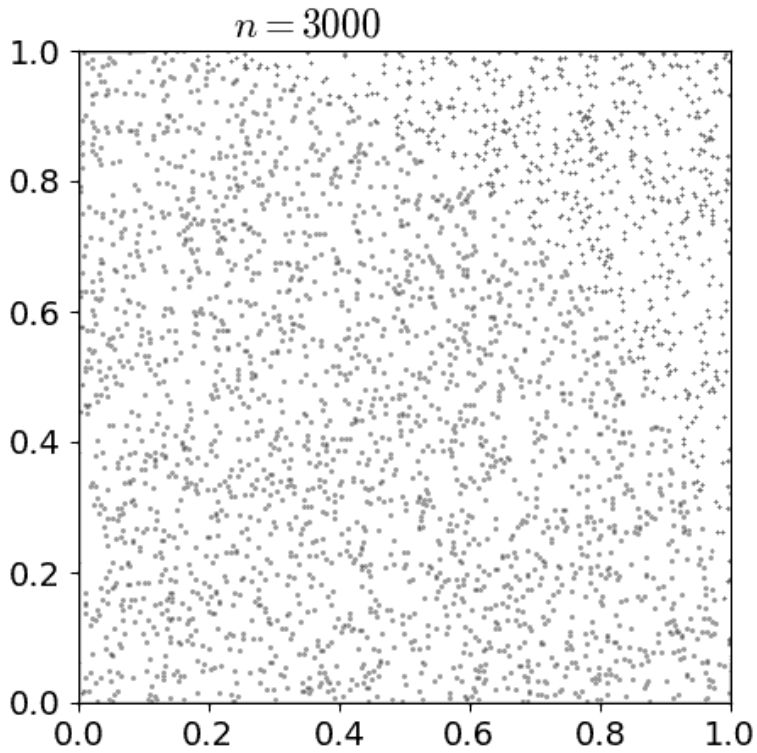
1. Define an domain for input



Monte Carlo Sampling

Algorithm:

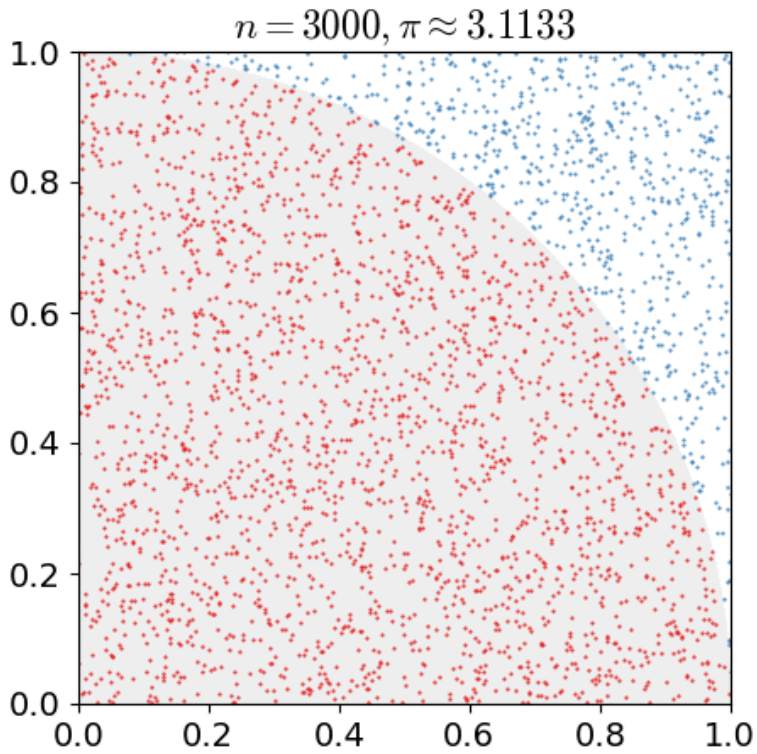
1. Define an domain for input
2. Generate random inputs from a probability distribution over the domain



Monte Carlo Sampling

Algorithm:

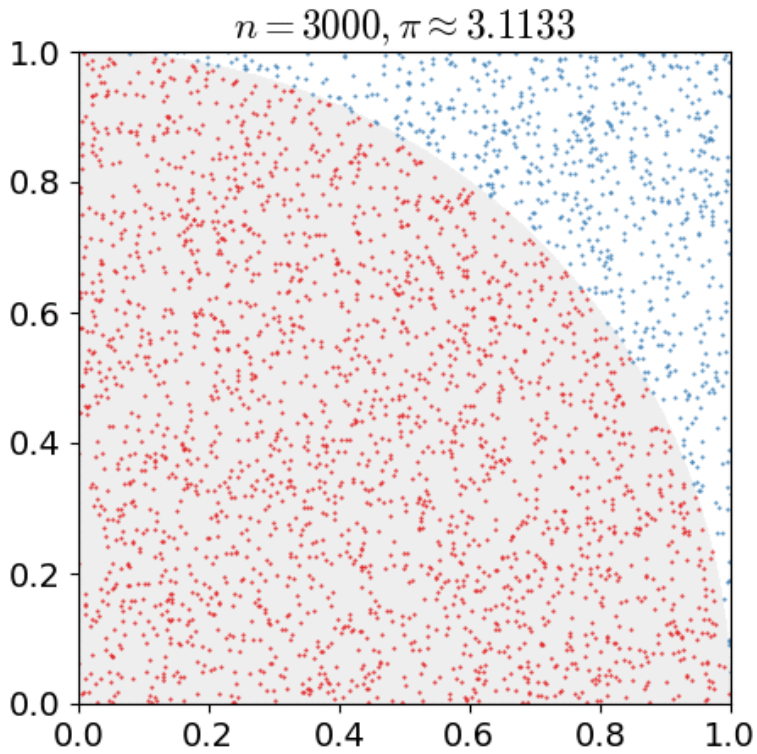
1. Define an domain for input
2. Generate random inputs from a probability distribution over the domain
3. Perform a deterministic computation on the inputs



Monte Carlo Sampling

Algorithm:

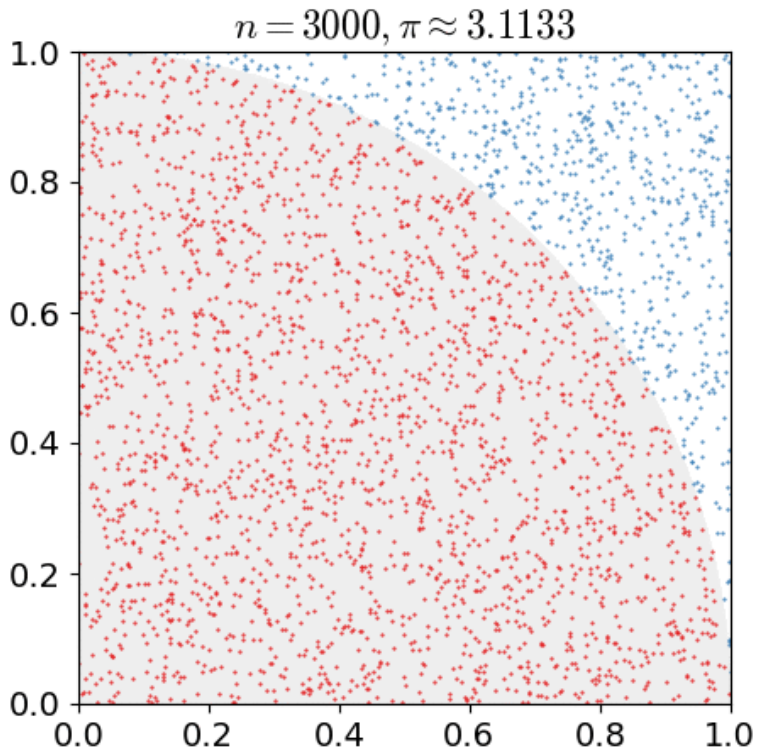
1. Define an domain for input
2. Generate random inputs from a probability distribution over the domain
3. Perform a deterministic computation on the inputs
4. Aggregate results



Monte Carlo Sampling

Algorithm:

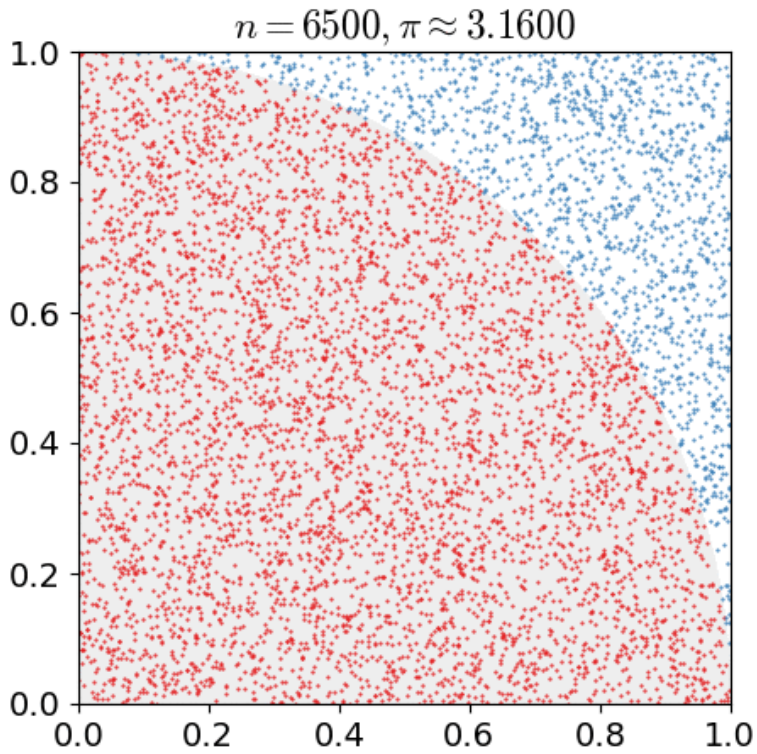
1. Define an domain for input
2. Generate random inputs from a probability distribution over the domain
3. Perform a deterministic computation on the inputs
4. Aggregate results



Monte Carlo Sampling

Algorithm:

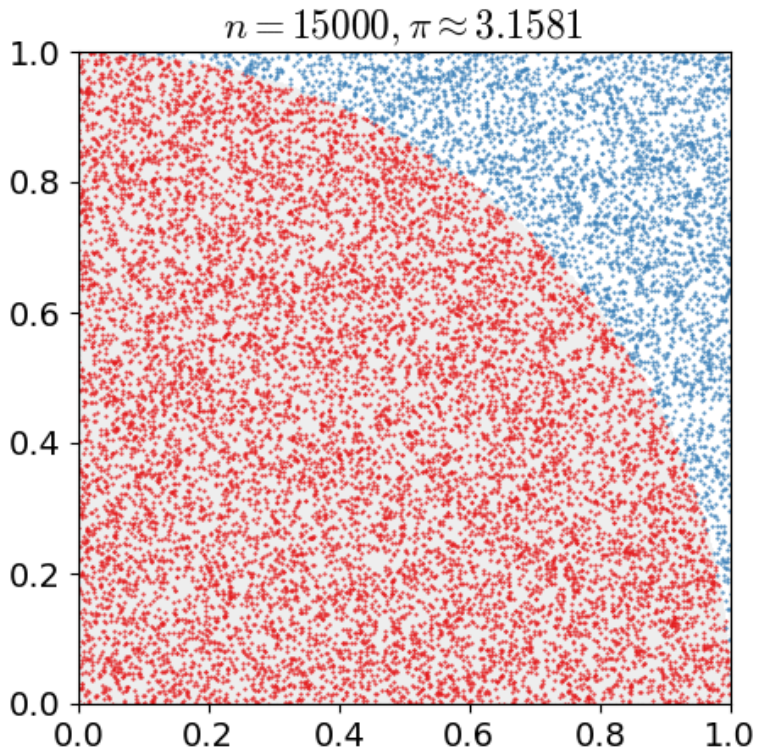
1. Define an domain for input
2. Generate random inputs from a probability distribution over the domain
3. Perform a deterministic computation on the inputs
4. Aggregate results



Monte Carlo Sampling

Algorithm:

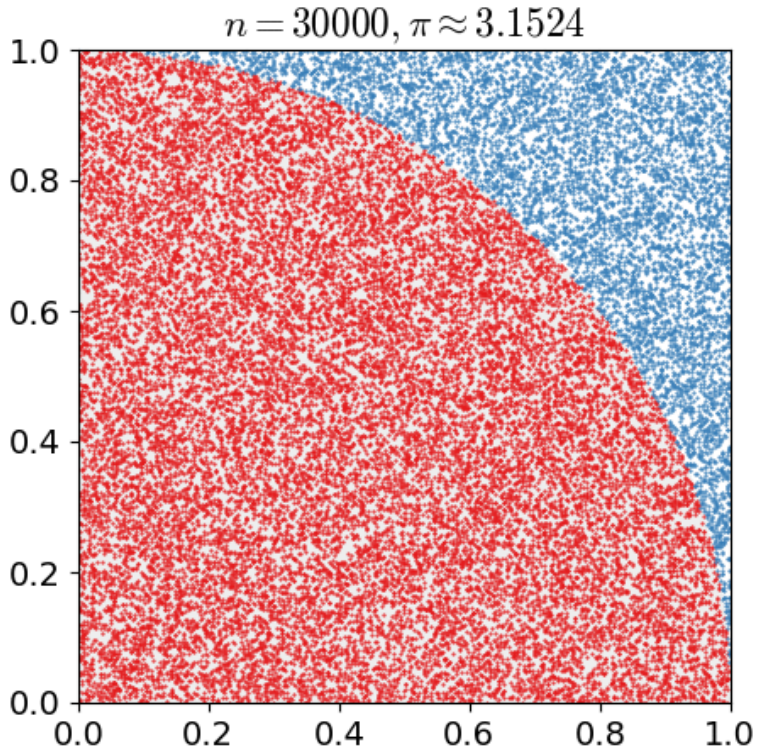
1. Define an domain for input
2. Generate random inputs from a probability distribution over the domain
3. Perform a deterministic computation on the inputs
4. Aggregate results



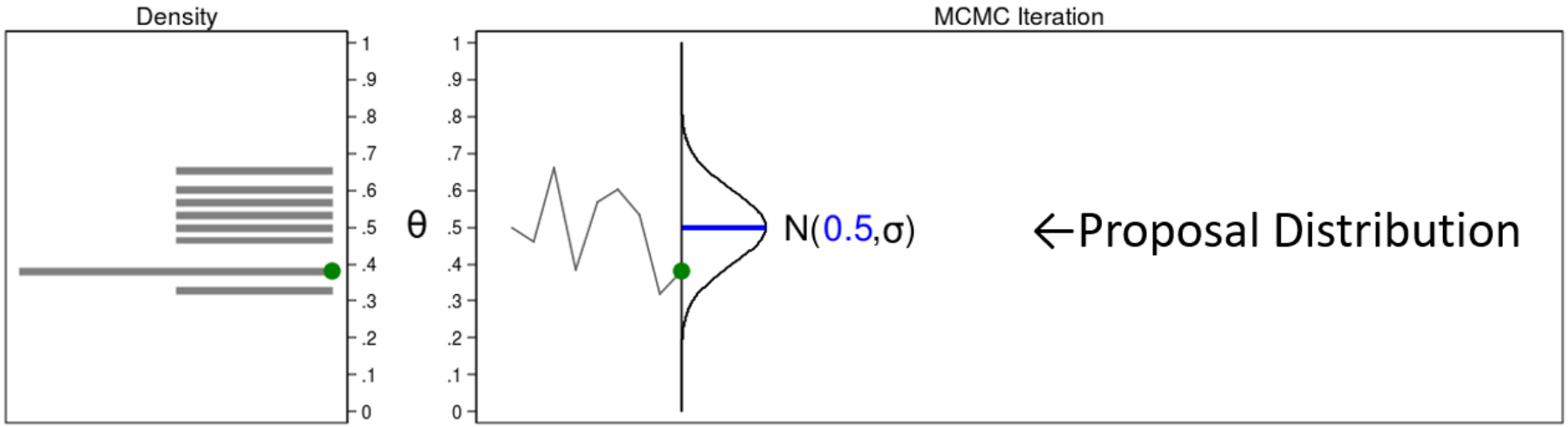
Monte Carlo Sampling

Algorithm:

1. Define an domain for input
2. Generate random inputs from a probability distribution over the domain
3. Perform a deterministic computation on the inputs
4. Aggregate results



Monte Carlo Sampling



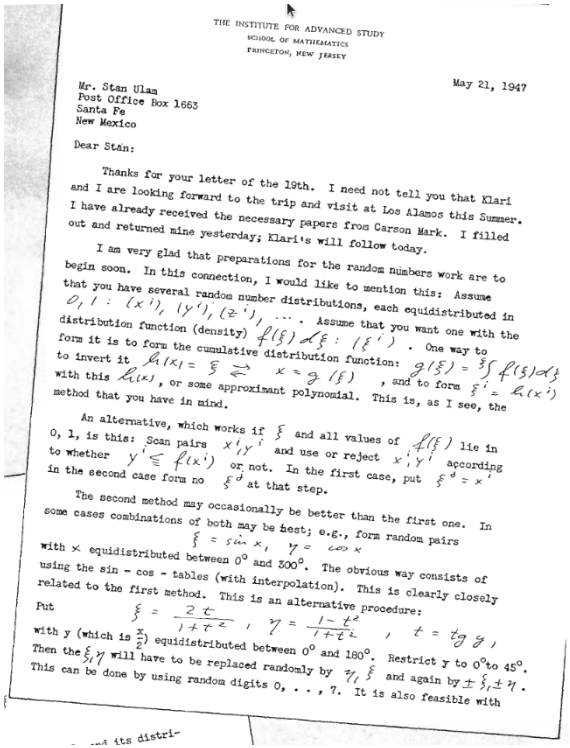
Draw $\theta_t \sim \text{Normal}(0.5,\sigma) = 0.381$

MARCOV CHAIN MONTE CARLO

SAMPLING

How to sample the proposal function: Acceptance-Rejection Method

Von Neumann further came up with an algorithm to sample an arbitrary (quasi-)probability density q using another, tractable one h.



THE INSTITUTE FOR ADVANCED STUDY
SCHOOL OF MATHEMATICS
PRINCETON, NEW JERSEY

May 21, 1947

Mr. Stan Ulam
Post Office Box 1663
Santa Fe
New Mexico

Dear Stan:

Thanks for your letter of the 19th. I need not tell you that Klari and I are looking forward to the trip and visit at Los Alamos this Summer. I have already received the necessary papers from Carson Mark. I filled out and returned mine yesterday; Klari's will follow today.

I am very glad that preparations for the random numbers work are to begin soon. In this connection, I would like to mention this: Assume that you have several random number distributions, each equidistributed in $0, 1$: $(x^1), (y^1), (z^1), \ldots$. Assume that you want one with the distribution function (density) $f(\xi)\,d\xi$: (ξ^1). One way to form it is to form the cumulative distribution function: $g(\xi) = \int^{\xi} f(\eta)\,d\eta$ to invert it $h_1(x) = \xi \rightleftarrows x = g(\xi)$, and to form $\xi^1 = h_1(x^1)$ with this $h_1(x)$, or some approximant polynomial. This is, as I see, the method that you have in mind.

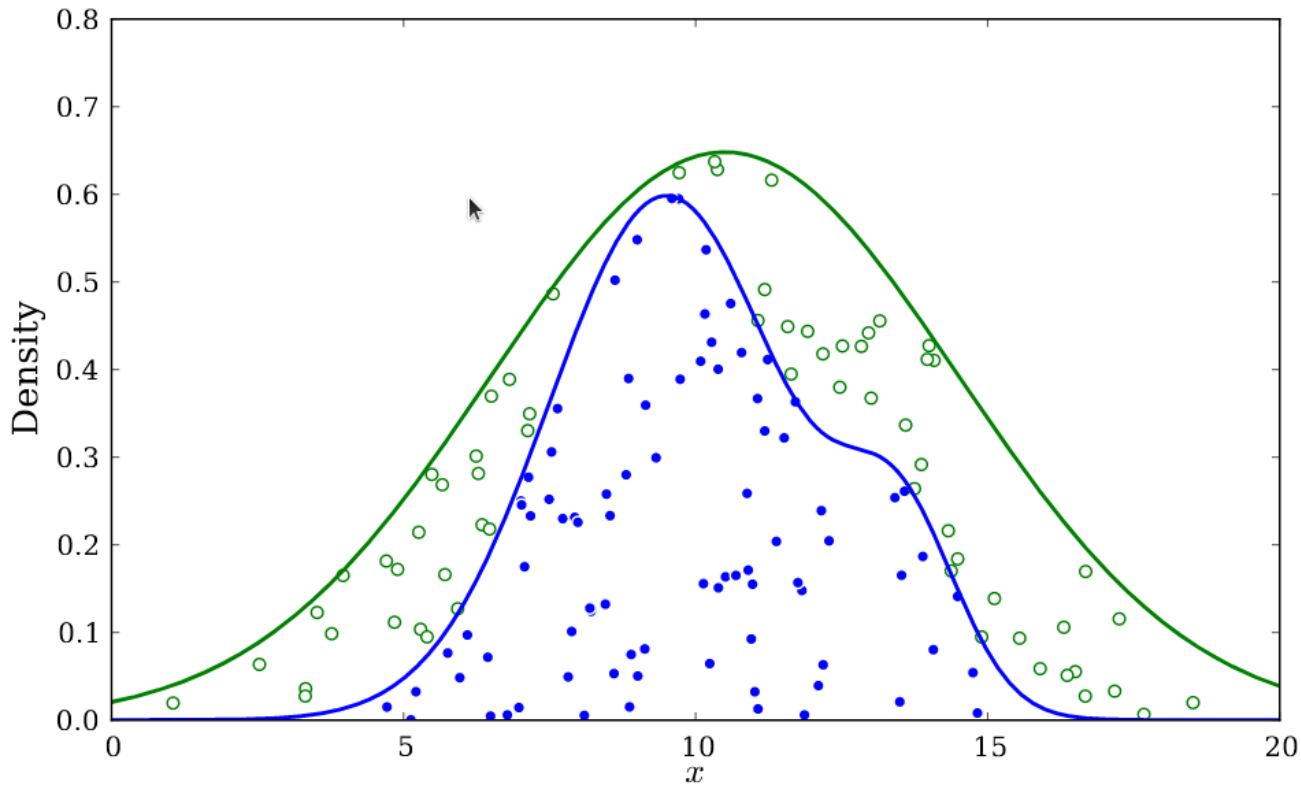
An alternative, which works if ξ and all values of $f(\xi)$ lie in $0, 1$, is this: Scan pairs x^1, y^1 and use or reject x^1, y^1 according to whether $y^1 \leqq f(x^1)$ or not. In the first case, put $\xi^1 = x^1$ in the second case form no ξ^1 at that step.

The second method may occasionally be better than the first one. In some cases combinations of both may be best; e.g., form random pairs $\xi = \sin x$, $\eta = \cos x$ with x equidistributed between 0° and 360°. The obvious way consists of using the sin - cos - tables (with interpolation). This is clearly closely related to the first method. This is an alternative procedure:

Put $\xi = \frac{2t}{1+t^2}$, $\eta = \frac{1-t^2}{1+t^2}$, $t = \tan \frac{y}{2}$,

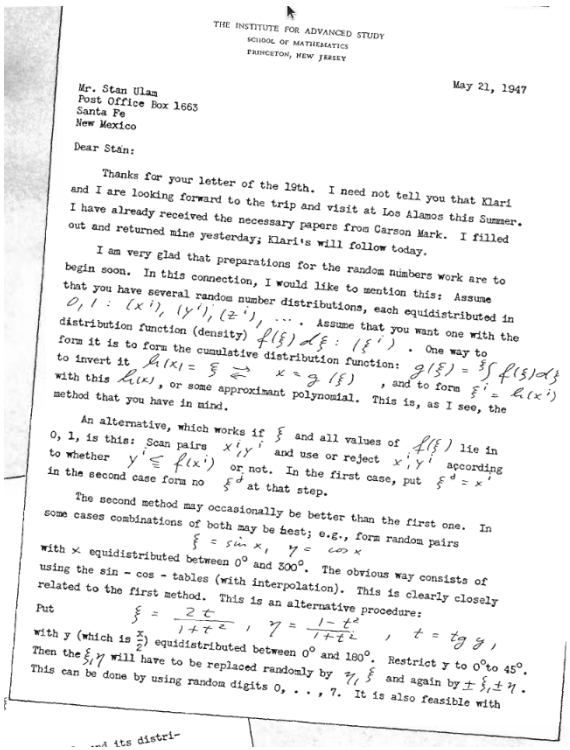
with y (which is $\frac{x}{2}$) equidistributed between 0° and 180°. Restrict y to 0° to 45°. Then the ξ, η will have to be replaced randomly by η, ξ and again by $\pm\xi, \pm\eta$. This can be done by using random digits $0, \ldots, 7$. It is also feasible with

... and its distri-



How to sample the proposal function: Acceptance-Rejection Method

Von Neumann further came up with an algorithm to sample an arbitrary (quasi-)probability density q using another, tractable one h.



THE INSTITUTE FOR ADVANCED STUDY
SCHOOL OF MATHEMATICS
PRINCETON, NEW JERSEY

May 21, 1947

Mr. Stan Ulam
Post Office Box 1663
Santa Fe
New Mexico

Dear Stan:

Thanks for your letter of the 19th. I need not tell you that Klari and I are looking forward to the trip and visit at Los Alamos this Summer. I have already received the necessary papers from Carson Mark. I filled out and returned mine yesterday; Klari's will follow today.

I am very glad that preparations for the random numbers work are to begin soon. In this connection, I would like to mention this: Assume that you have several random number distributions, each equidistributed in $0, 1$: $(x^{(i)})$, $(y^{(i)})$, $(z^{(i)})$, ... Assume that you want one with the distribution function (density) $f(\xi)\,d\xi$: $(\xi^{(i)})$. One way to form it is to form the cumulative distribution function: $g(\xi) = \int^{\xi} f(\xi)\,d\xi$ to invert it $h(x) = \xi \rightsquigarrow x = g(\xi)$, and to form $\xi^{(i)} = h(x^{(i)})$ with this $h(x)$, or some approximant polynomial. This is, as I see, the method that you have in mind.

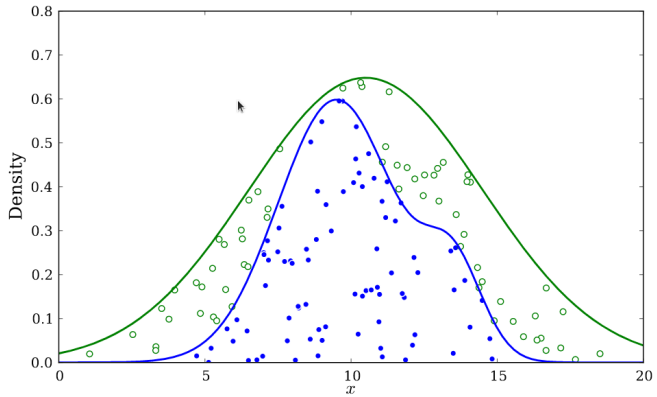
An alternative, which works if ξ and all values of $f(\xi)$ lie in $0, 1$, is this: Scan pairs $x^{(i)}, y^{(i)}$ and use or reject $x^{(i)}, y^{(i)}$ according to whether $y^{(i)} \leqq f(x^{(i)})$ or not. In the first case, put $\xi^{(d)} = x^{(i)}$ in the second case form no $\xi^{(d)}$ at that step.

The second method may occasionally be better than the first one. In some cases combinations of both may be best; e.g., form random pairs $\xi = \sin x$, $\eta = \cos x$ with x equidistributed between 0° and 300°. The obvious way consists of using the sin - cos - tables (with interpolation). This is clearly closely related to the first method. This is an alternative procedure:

Put $\xi = \frac{2t}{1+t^2}$, $\eta = \frac{1-t^2}{1+t^2}$, $t = \operatorname{tg} \frac{y}{2}$,

with y (which is $\frac{x}{2}$) equidistributed between 0° and 180°. Restrict y to 0° to 45°. Then the ξ, η will have to be replaced randomly by η, ξ and again by $\pm\xi, \pm\eta$. This can be done by using random digits $0, \ldots, 7$. It is also feasible with

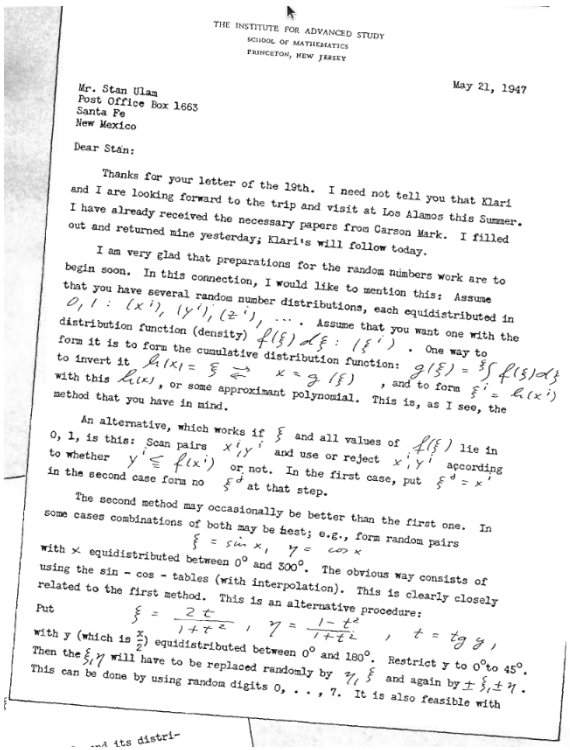
... and its distri-



1. Choose a tractable density $h(\Theta)$ and a constant C so $C \cdot h \geq q \; \forall \Theta$

How to sample the proposal function: Acceptance-Rejection Method

Von Neumann further came up with an algorithm to sample an arbitrary (quasi-)probability density q using another, tractable one h.



THE INSTITUTE FOR ADVANCED STUDY
SCHOOL OF MATHEMATICS
PRINCETON, NEW JERSEY

May 21, 1947

Mr. Stan Ulam
Post Office Box 1663
Santa Fe
New Mexico

Dear Stan:

Thanks for your letter of the 19th. I need not tell you that Klari and I are looking forward to the trip and visit at Los Alamos this Summer. I have already received the necessary papers from Carson Mark. I filled out and returned mine yesterday; Klari's will follow today.

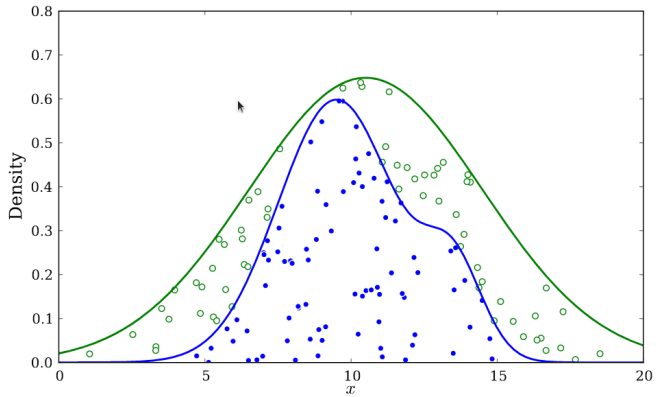
I am very glad that preparations for the random numbers work are to begin soon. In this connection, I would like to mention this: Assume that you have several random number distributions, each equidistributed in $0, 1$: $(x^{(i)})$, $(y^{(i)})$, $(z^{(i)})$, ... Assume that you want one with the distribution function (density) $f(\xi)\,d\xi$: $(\xi^{(i)})$. One way to form it is to form the cumulative distribution function: $g(\xi) = \int_{-\infty}^{\xi} f(\xi)\,d\xi$ to invert it $h(x) = \xi \rightleftharpoons x = g(\xi)$, and to form $\xi^{(i)} = h(x^{(i)})$ with this $h(x)$, or some approximant polynomial. This is, as I see, the method that you have in mind.

An alternative, which works if ξ and all values of $f(\xi)$ lie in $0, 1$, is this: Scan pairs $x^{(i)}, y^{(i)}$ and use or reject $x^{(i)}, y^{(i)}$ according to whether $y^{(i)} \leqq f(x^{(i)})$ or not. In the first case, put $\xi^{(d)} = x^{(i)}$ in the second case form no $\xi^{(d)}$ at that step.

The second method may occasionally be better than the first one. In some cases combinations of both may be best; e.g., form random pairs $\xi = \sin x$, $\eta = \cos x$ with x equidistributed between 0° and 360°. The obvious way consists of using the sin - cos - tables (with interpolation). This is clearly closely related to the first method. This is an alternative procedure:

Put $\xi = \frac{2t}{1+t^2}$, $\eta = \frac{1-t^2}{1+t^2}$, $t = \operatorname{tg} \frac{y}{2}$, with y (which is $\frac{x}{2}$) equidistributed between 0° and 180°. Restrict y to 0° to 45°. Then the ξ, η will have to be replaced randomly by η, ξ and again by $\pm\xi, \pm\eta$. This can be done by using random digits $0, \ldots, 7$. It is also feasible with

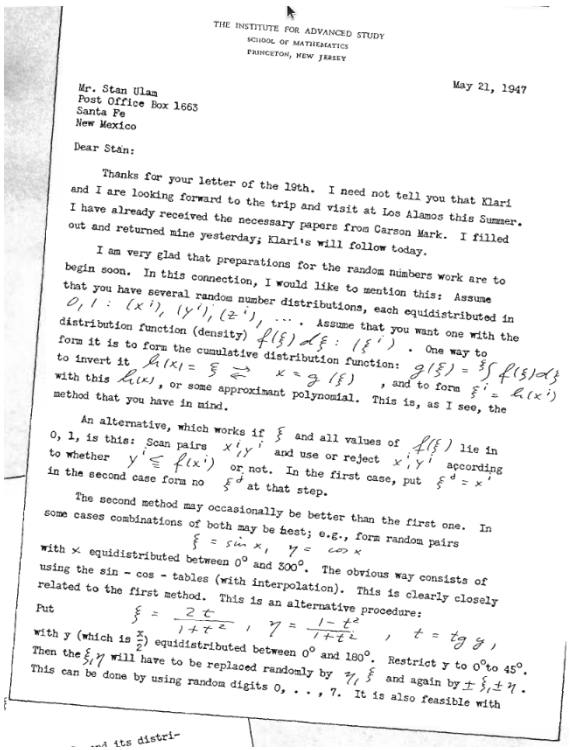
... and its distri-



1. Choose a tractable density $h(\Theta)$ and a constant C so $C \cdot h \geq q \ \forall \Theta$
2. Draw a candidate parameter value Θ' from h

How to sample the proposal function: Acceptance-Rejection Method

Von Neumann further came up with an algorithm to sample an arbitrary (quasi-)probability density q using another, tractable one h.



THE INSTITUTE FOR ADVANCED STUDY
SCHOOL OF MATHEMATICS
PRINCETON, NEW JERSEY

May 21, 1947

Mr. Stan Ulam
Post Office Box 1663
Santa Fe
New Mexico

Dear Stan:

Thanks for your letter of the 19th. I need not tell you that Klari and I are looking forward to the trip and visit at Los Alamos this Summer. I have already received the necessary papers from Carson Mark. I filled out and returned mine yesterday; Klari's will follow today.

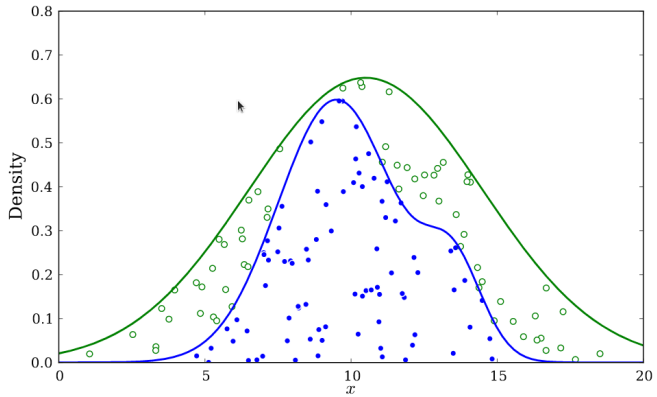
I am very glad that preparations for the random numbers work are to begin soon. In this connection, I would like to mention this: Assume that you have several random number distributions, each equidistributed in $0, 1$: $(x^{(i)})$, $(y^{(i)})$, $(z^{(i)})$, ... Assume that you want one with the distribution function (density) $f(\xi)\,d\xi$: $(\xi^{(i)})$. One way to form it is to form the cumulative distribution function: $g(\xi) = \int_{-\infty}^{\xi} f(\eta)\,d\eta$ to invert it $h(x) = \xi \rightleftharpoons x = g(\xi)$, and to form $\xi^{(i)} = h(x^{(i)})$ with this $h(x)$, or some approximant polynomial. This is, as I see, the method that you have in mind.

An alternative, which works if ξ and all values of $f(\xi)$ lie in $0, 1$, is this: Scan pairs $x^{(i)}, y^{(i)}$ and use or reject $x^{(i)}, y^{(i)}$ according to whether $y^{(i)} \leqq f(x^{(i)})$ or not. In the first case, put $\xi^{(d)} = x^{(i)}$ in the second case form no $\xi^{(d)}$ at that step.

The second method may occasionally be better than the first one. In some cases combinations of both may be best; e.g., form random pairs $\xi = \sin x$, $\eta = \cos x$ with x equidistributed between 0° and 360°. The obvious way consists of using the sin - cos - tables (with interpolation). This is clearly closely related to the first method. This is an alternative procedure:

Put $\xi = \frac{2t}{1+t^2}$, $\eta = \frac{1-t^2}{1+t^2}$, $t = \operatorname{tg} \frac{x}{2}$, with y (which is $\frac{x}{2}$) equidistributed between 0° and 180°. Restrict y to 0° to 45°. Then the ξ, η will have to be replaced randomly by η, ξ and again by $\pm\xi, \pm\eta$. This can be done by using random digits $0, \ldots, 7$. It is also feasible with

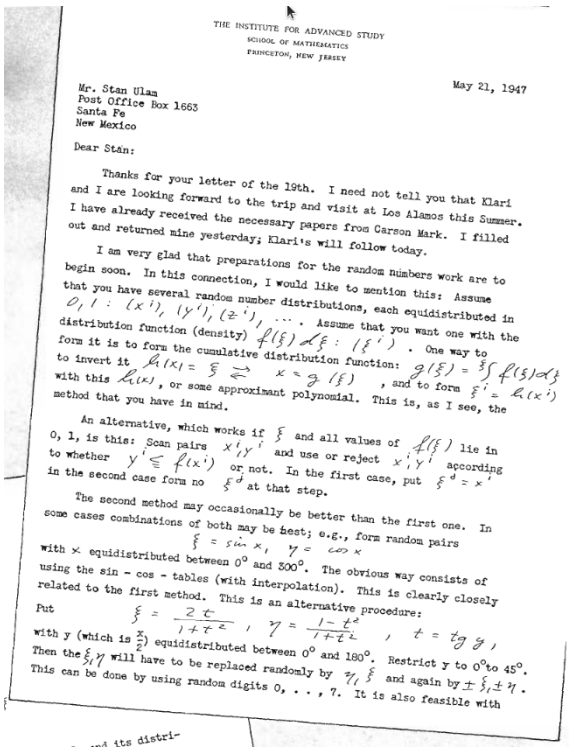
... and its distri-



1. Choose a tractable density $h(\Theta)$ and a constant C so $C \cdot h \geq q \; \forall \Theta$
2. Draw a candidate parameter value Θ' from h
3. Draw a random uniform number $u \in [0, 1]$

How to sample the proposal function: Acceptance-Rejection Method

Von Neumann further came up with an algorithm to sample an arbitrary (quasi-)probability density q using another, tractable one h.



THE INSTITUTE FOR ADVANCED STUDY
SCHOOL OF MATHEMATICS
PRINCETON, NEW JERSEY

May 21, 1947

Mr. Stan Ulam
Post Office Box 1663
Santa Fe
New Mexico

Dear Stan:

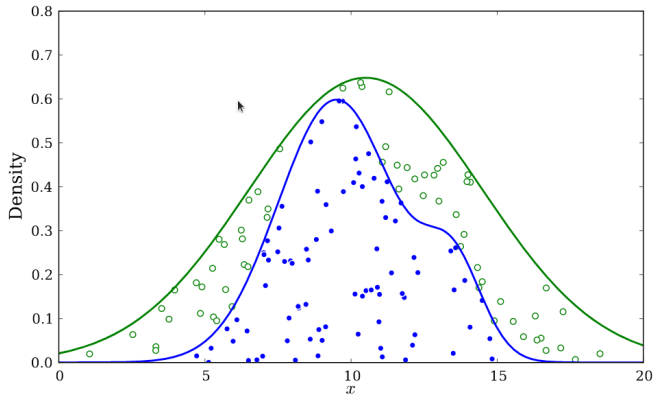
Thanks for your letter of the 19th. I need not tell you that Klari and I are looking forward to the trip and visit at Los Alamos this Summer. I have already received the necessary papers from Carson Mark. I filled out and returned mine yesterday; Klari's will follow today.

I am very glad that preparations for the random numbers work are to begin soon. In this connection, I would like to mention this: Assume that you have several random number distributions, each equidistributed in $0, 1$: $(x^{(1)}), (y^{(1)}), (z^{(1)}), \ldots$. Assume that you want one with the distribution function (density) $f(\xi)\,d\xi$: $(\xi^{(1)})$. One way to form it is to form the cumulative distribution function: $g(\xi) = \int^{\xi} f(\xi)\,d\xi$ to invert it $h(x) = \xi \rightarrow x = g^{-1}(\xi)$, and to form $\xi^{(1)} = h(x^{(1)})$ with this $h(x)$, or some approximant polynomial. This is, as I see, the method that you have in mind.

An alternative, which works if ξ and all values of $f(\xi)$ lie in $0, 1$, is this: Scan pairs $x^{(1)}, y^{(1)}$ and use or reject $x^{(1)}, y^{(1)}$ according to whether $y^{(1)} \leqq f(x^{(1)})$ or not. In the first case, put $\xi^{(1)} = x^{(1)}$ in the second case form no $\xi^{(1)}$ at that step.

The second method may occasionally be better than the first one. In some cases combinations of both may be best; e.g., form random pairs $\xi = \sin x$, $\eta = \cos x$ with x equidistributed between 0° and 360°. The obvious way consists of using the sin - cos - tables (with interpolation). This is clearly closely related to the first method. This is an alternative procedure:

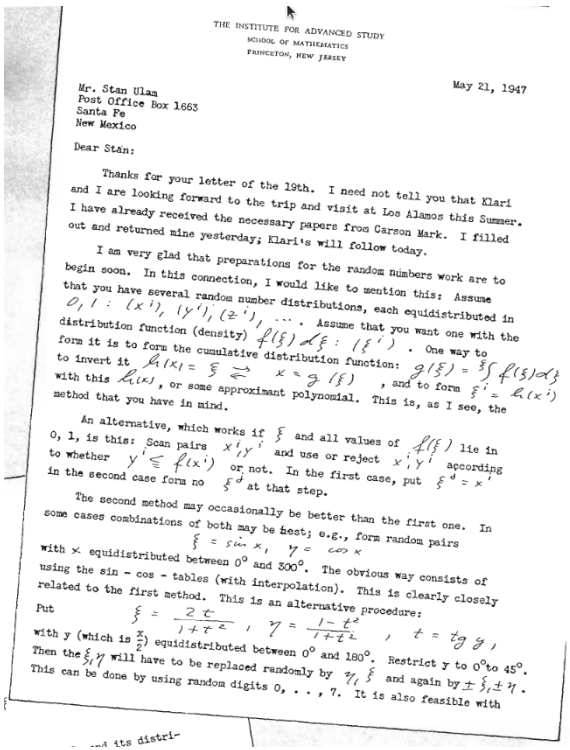
Put $\xi = \frac{2t}{1+t^2}$, $\eta = \frac{1-t^2}{1+t^2}$, $t = \operatorname{tg} \frac{x}{2}$, with y (which is $\frac{x}{2}$) equidistributed between 0° and 180°. Restrict y to 0° to 45°. Then the ξ, η will have to be replaced randomly by η, ξ and again by $\pm\xi, \pm\eta$. This can be done by using random digits $0, \ldots, 7$. It is also feasible with



1. Choose a tractable density $h(\Theta)$ and a constant C so $C \cdot h \geq q \ \forall \Theta$
2. Draw a candidate parameter value Θ' from h
3. Draw a random uniform number $u \in [0, 1]$
4. If $u < \frac{C \cdot h(\Theta')}{q(\Theta')}$ accept Θ' as a sample

How to sample the proposal function: Acceptance-Rejection Method

Von Neumann further came up with an algorithm to sample an arbitrary (quasi-)probability density q using another, tractable one h.



THE INSTITUTE FOR ADVANCED STUDY
SCHOOL OF MATHEMATICS
PRINCETON, NEW JERSEY

May 21, 1947

Mr. Stan Ulam
Post Office Box 1663
Santa Fe
New Mexico

Dear Stan:

Thanks for your letter of the 19th. I need not tell you that Klari and I are looking forward to the trip and visit at Los Alamos this Summer. I have already received the necessary papers from Carson Mark. I filled out and returned mine yesterday; Klari's will follow today.

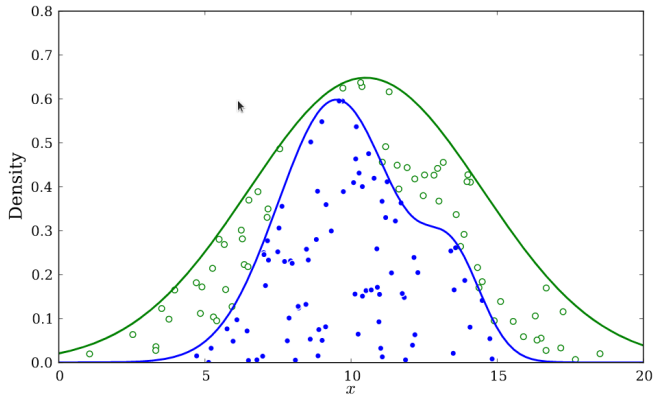
I am very glad that preparations for the random numbers work are to begin soon. In this connection, I would like to mention this: Assume that you have several random number distributions, each equidistributed in $0, 1$: $(x^{(1)}), (y^{(1)}), (z^{(1)}), \ldots$. Assume that you want one with the distribution function (density) $f(\xi)$ $d\xi$: $(\xi^{(1)})$. One way to form it is to form the cumulative distribution function: $g(\xi) = \int_{-\infty}^{\xi} f(\xi)\,d\xi$ to invert it $h(x) = \xi \rightleftharpoons x = g(\xi)$, and to form $\xi^{(1)} = h(x^{(1)})$ with this $h(x)$, or some approximant polynomial. This is, as I see, the method that you have in mind.

An alternative, which works if ξ and all values of $f(\xi)$ lie in $0, 1$, is this: Scan pairs $x^{(1)}, y^{(1)}$ and use or reject $x^{(1)}, y^{(1)}$ according to whether $y^{(1)} \leqq f(x^{(1)})$ or not. In the first case, put $\xi^{(1)} = x^{(1)}$ in the second case form no $\xi^{(1)}$ at that step.

The second method may occasionally be better than the first one. In some cases combinations of both may be best; e.g., form random pairs $\xi = \sin x$, $\eta = \cos x$ with x equidistributed between 0° and 360°. The obvious way consists of using the sin - cos - tables (with interpolation). This is clearly closely related to the first method. This is an alternative procedure:

Put $\xi = \frac{2t}{1+t^2}$, $\eta = \frac{1-t^2}{1+t^2}$, $t = \operatorname{tg} \frac{y}{2}$, with y (which is $\frac{x}{2}$) equidistributed between 0° and 180°. Restrict y to 0° to 45°. Then the ξ, η will have to be replaced randomly by η, ξ and again by $\pm\xi, \pm\eta$. This can be done by using random digits $0, \ldots, 7$. It is also feasible with

... and its distri-



1. Choose a tractable density $h(\Theta)$ and a constant C so $C \cdot h \geq q \ \forall \Theta$
2. Draw a candidate parameter value Θ' from h
3. Draw a random uniform number $u \in [0, 1]$
4. If $u < \frac{C \cdot h(\Theta')}{q(\Theta')}$ accept Θ' as a sample
5. Goto (2)

Motivation: Bayesian inference

- In Bayesian inference, we take both new data (aka evidence) as well as a prior belief into account when computing the probability for a certain model and its parameters $\{\Theta_i\}$

Motivation: Bayesian inference

- In Bayesian inference, we take both new data (aka evidence) as well as a prior belief into account when computing the probability for a certain model and its parameters $\{\Theta_i\}$
- In case of Solitaire, we consider a deal as a binary random event with a given probability p to obtain one result (win) and $(1 - p)$ the other (lose).

Motivation: Bayesian inference

- In Bayesian inference, we take both new data (aka evidence) as well as a prior belief into account when computing the probability for a certain model and its parameters $\{\Theta_i\}$
- In case of Solitaire, we consider a deal as a binary random event with a given probability p to obtain one result (win) and $(1-p)$ the other (lose).
- According to Bayes' theorem, this posterior belief is given by

$$\underbrace{P(\{\Theta_i\}|\mathrm{Data})}_{\text{Posterior Prob.}} = \underbrace{P(\{\Theta_i\})}_{\text{Prior Prob.}} \cdot \underbrace{P(\mathrm{Data}|\{\Theta_i\})}_{\text{Likelihood}} / \underbrace{P(\mathrm{Data})}_{\text{Model evidence}}$$

Bayesian inference - Prior Belief/Probability

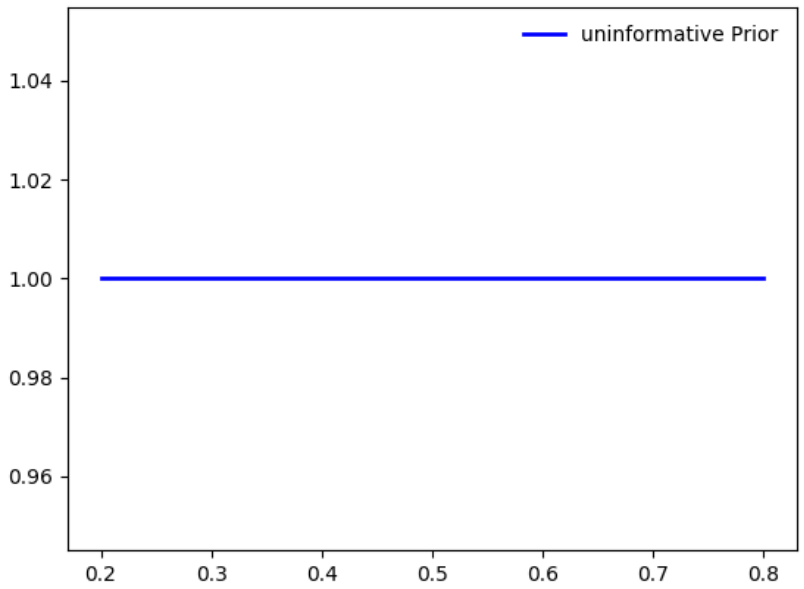
$$P(\{\Theta_i\}|\text{Data}) = \underbrace{P(\{\Theta_i\})}_{\text{Prior Prob.}} \cdot P(\text{Data}|\{\Theta_i\})/P(\text{Data})$$

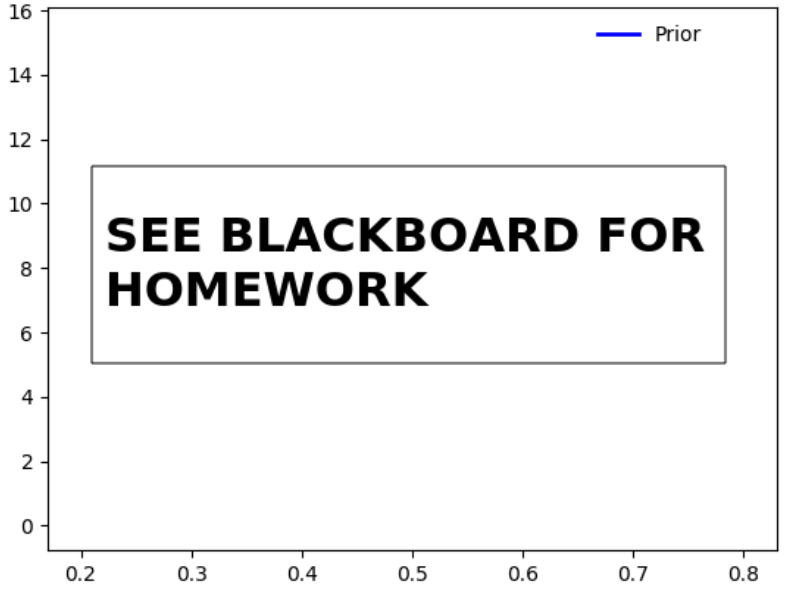
- mathematical expression (i.e. pdf) quantifying our belief about the model parameters

Bayesian inference - Prior Belief/Probability

$$P(\{\Theta_i\}|\text{Data}) = \underbrace{P(\{\Theta_i\})}_{\text{Prior Prob.}} \cdot P(\text{Data}|\{\Theta_i\})/P(\text{Data})$$

- mathematical expression (i.e. pdf) quantifying our belief about the model parameters
- two approaches: least-/uninformative vs informative prior





Bayesian inference - Data Likelihood

$$P(\{\Theta_i\}|\mathrm{Data}) = P(\{\Theta_i\}) \cdot \underbrace{P(\mathrm{Data}|\{\Theta_i\})}_{\text{Likelihood}} / P(\mathrm{Data})$$

- probability of the data for a given parameter set for a specific model

Bayesian inference - Data Likelihood

$$P(\{\Theta_i\}|\mathrm{Data}) = P(\{\Theta_i\}) \cdot \underbrace{P(\mathrm{Data}|\{\Theta_i\})}_{\text{Likelihood}} / P(\mathrm{Data})$$

- probability of the data for a given parameter set for a specific model
- for Solitaire, the likelihood function is the binomial distribution, i.e. the likelihood that exactly k out of n random deals are winnable assuming an underlying propability to win p is given by

$$L(\{k, n\}|\{p\}) = Binom(k, n, p) = \binom{n}{k} p^k (1-p)^{n-k}$$

Bayesian inference - Data Likelihood

$$P(\{\Theta_i\}|\mathrm{Data}) = P(\{\Theta_i\}) \cdot \underbrace{P(\mathrm{Data}|\{\Theta_i\})}_{\text{Likelihood}} / P(\mathrm{Data})$$

- probability of the data for a given parameter set for a specific model
- for Solitaire, the likelihood function is the binomial distribution, i.e. the likelihood that exactly k out of n random deals are winnable assuming an underlying propability to win p is given by

$$L(\{k,n\}|\{p\}) = Binom(k,n,p) = \binom{n}{k} p^k (1-p)^{n-k}$$

- in general, more complicated (i.e. no analytical expression, can only be computed numerically)

Bayesian inference - Probability of Evidence

$$P(\{\Theta_i\}|\mathrm{Data}) = P(\{\Theta_i\}) \cdot P(\mathrm{Data}|\{\Theta_i\}) / \underbrace{P(\mathrm{Data})}_{\textbf{Modelevidence}}$$

- Obtained by marginalizing the Likelihood over all possible parameters of a model type, i.e. the probabilty to obain specific data independently of the choice of model parameters

$$P(\mathrm{Data}) = \int P(\mathrm{Data}|\{\Theta_i\}) P(\{\Theta_i\}) d\{\Theta_i\}$$

Bayesian inference - Probability of Evidence

$$P(\{\Theta_i\}|\mathrm{Data}) = P(\{\Theta_i\}) \cdot P(\mathrm{Data}|\{\Theta_i\}) / \underbrace{P(\mathrm{Data})}_{\textbf{Modelevidence}}$$

- Obtained by marginalizing the Likelihood over all possible parameters of a model type, i.e. the probabilty to obain specific data independently of the choice of model parameters

$$P(\mathrm{Data}) = \int P(\mathrm{Data}|\{\Theta_i\}) P(\{\Theta_i\}) d\{\Theta_i\}$$

- Usually very difficult/expensive to compute

Bayesian inference - Probability of Evidence

$$P(\{\Theta_i\}|\mathrm{Data}) = P(\{\Theta_i\}) \cdot P(\mathrm{Data}|\{\Theta_i\}) / \underbrace{P(\mathrm{Data})}_{\textbf{Modelevidence}}$$

- Obtained by marginalizing the Likelihood over all possible parameters of a model type, i.e. the probabilty to obain specific data independently of the choice of model parameters

$$P(\mathrm{Data}) = \int P(\mathrm{Data}|\{\Theta_i\}) P(\{\Theta_i\}) d\{\Theta_i\}$$

- Usually very difficult/expensive to compute
- Fortunately as a constant proportionality factor Z, we do not need to know its value if we are only interested in ratios of posterior propabilities:

$$P(\{\Theta_i\}|\mathrm{Data}) \simeq P(\{\Theta_i\}) \cdot P(\mathrm{Data}|\{\Theta_i\})$$

Bayesian inference - Posterior Probability

$$\underbrace{P(\{\Theta_i\}|\mathrm{Data})}_{\text{Posterior Prob.}} \simeq P(\{\Theta_i\}) \cdot P(\mathrm{Data}|\{\Theta_i\})$$

- In some cases, posterior likelihood can be analytically calculated e.g. for our Solitaire model and an uninformative, flat prior or certain analytical prior probability (here e.g. beta-distribution)

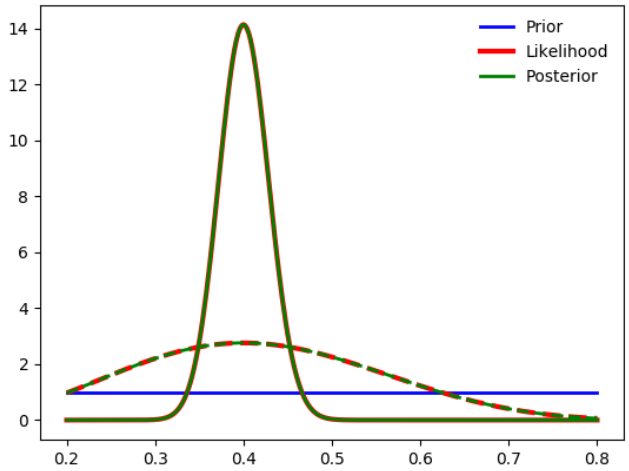
Bayesian inference - Posterior Probability (uninformative prior)

In general, the posterior for a flat, uninformative prior is identical to the likelihood function (in the part of the parameter space of interest and vanishes elsewhere)

$$P(\{\Theta_i\}|\mathrm{Data}) = P(\mathrm{Data}|\{\Theta_i\})\ |_{\mathrm{support}}$$

Thus, for our Solitaire example we get

$$P(\{p\}|\{n,k\}) = Binom(n,k,p)$$



Bayesian inference - Posterior Prob. (inform. prior)

- If we have an informative analytical prior, e.g. here a beta-distribution, we may still be able to calculate the posterior analytically. For our Solitaire example, we obtain

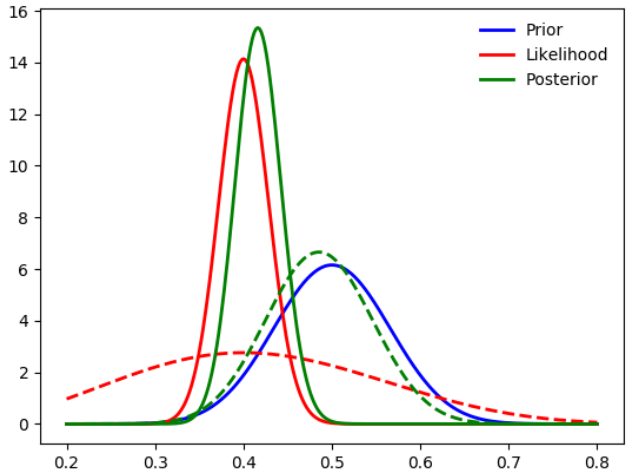
$$\begin{aligned} P(\{p\}|\{n,k\}) &= Beta(p,x,y) \cdot Binom(n,k,p) \\ &= Beta(p,x+k,y+(n-k)) \end{aligned}$$

Bayesian inference - Posterior Prob. (inform. prior)

- If we have an informative analytical prior, e.g. here a beta-distribution, we may still be able to calculate the posterior analytically. For our Solitaire example, we obtain

$$\begin{aligned} P(\{p\}|\{n,k\}) &= Beta(p,x,y) \cdot Binom(n,k,p) \\ &= Beta(p,x+k,y+(n-k)) \end{aligned}$$

- Here, posterior is again a beta-distr. i.e prior and posterior belong to the same distribution family. Hence, this is called a **conjugate prior** to this binomial likelihood.



Bayesian inference - Posterior Probability (cont.)

- In general, this is not possible as the prior or the likelihood may not be available as an analytical pdf (see e.g. your Solitaire prior beliefs)

Bayesian inference - Posterior Probability (cont.)

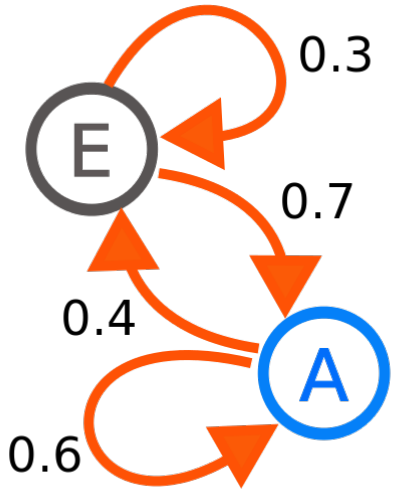
- In general, this is not possible as the prior or the likelihood may not be available as an analytical pdf (see e.g. your Solitaire prior beliefs)
- cannot simply use MC + Acceptance-Rejection method to sample/compute posterior as we don't know our maximum density. Even if we know it, for a high dimensional parameter space, rejection rate will be very high due to difficult choice of bounding probability density.

Bayesian inference - Posterior Probability (cont.)

- In general, this is not possible as the prior or the likelihood may not be available as an analytical pdf (see e.g. your Solitaire prior beliefs)
- cannot simply use MC + Acceptance-Rejection method to sample/compute posterior as we don't know our maximum density. Even if we know it, for a high dimensional parameter space, rejection rate will be very high due to difficult choice of bounding probability density.
- need to find an alternative more efficient way to sample such a posterior

MARCOV CHAIN MONTE CARLO SAMPLING

Markov Chains: Definition



- A Markov Chain is a sequence of events $\{S_t\}$ in which the probability of each event depends only on the state attained in the previous event (Markov property):

$$P(S_t|\{S_{t-1}, S_{t-2}, \dots\}) = P(S_t|S_{t-1})$$

- (Time-homog.) Transition kernel

$$P(S_t = y|S_{t-1} = x) = T(y|x)$$

Markov Chains: Definition (cont.)

- Prob. to be in state y at time t:

$$P(S_t = y) = P(\text{stay at y}) + P(\text{move to y}) - P(\text{move from y})$$

in discrete case:

$$\begin{aligned} P(S_t = y) &= P(S_{t-1} = y) + \sum_{x \neq y} P(S_{t-1} = x) T(y|x) \\ &- \sum_{x \neq y} P(S_{t-1} = y) T(x|y)) \end{aligned}$$

can also be written as:

$$P(S_t = y) = P(S_{t-1} = y) T(y|y) + \sum_{x \neq y} P(S_{t-1} = x) T(y|x)$$

$$\Rightarrow \quad \vec{P}(S_t) = \vec{P}(S_{t-1}) T = \vec{P}(S_{t-2}) T^2 = \ldots$$

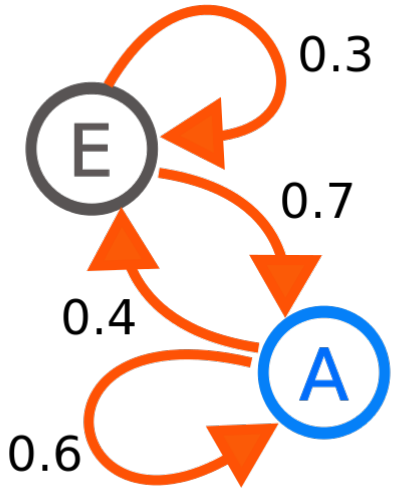
Markov Chains: Stationary Equilibrium

- Let's consider

$$\vec{P}(S_0) = \begin{pmatrix} P(S_0 = A) & P(S_0 = E) \end{pmatrix} = \begin{pmatrix} 1 & 0 \end{pmatrix}$$

$$T = \begin{pmatrix} T(A|A) & T(A|E) \\ T(E|A) & T(E|E) \end{pmatrix} = \begin{pmatrix} 0.6 & 0.7 \\ 0.4 & 0.3 \end{pmatrix}$$

and calculate the first few transitions using $\vec{P}(S_t) = \vec{P}(S_{t-1})T$:

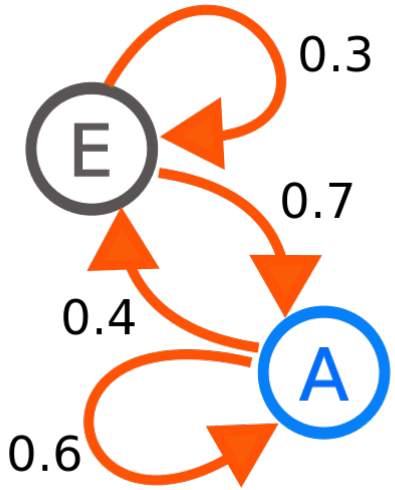


Markov Chains: Stationary Equilibrium

- Let's consider

$$\vec{P}(S_0) = \begin{pmatrix} P(S_0 = A) & P(S_0 = E) \end{pmatrix} = \begin{pmatrix} 1 & 0 \end{pmatrix}$$

$$T = \begin{pmatrix} T(A|A) & T(A|E) \\ T(E|A) & T(E|E) \end{pmatrix} = \begin{pmatrix} 0.6 & 0.7 \\ 0.4 & 0.3 \end{pmatrix}$$



and calculate the first few transitions using $\vec{P}(S_t) = \vec{P}(S_{t-1})T$:

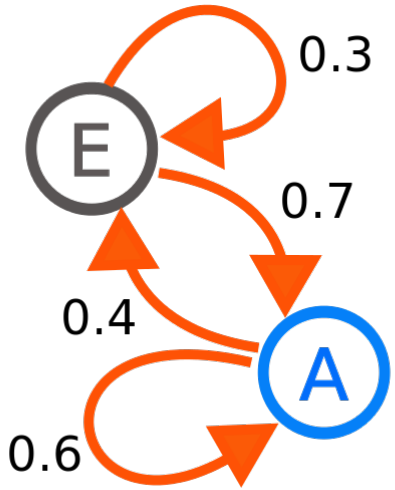
$$\begin{pmatrix} 0 \\ 1 \end{pmatrix} \rightarrow \begin{pmatrix} 0.7 \\ 0.3 \end{pmatrix} \rightarrow \begin{pmatrix} 0.63 \\ 0.37 \end{pmatrix} \rightarrow \begin{pmatrix} 0.637 \\ 0.363 \end{pmatrix} \rightarrow \begin{pmatrix} 0.6363 \\ 0.3637 \end{pmatrix} \rightarrow \cdots$$

Markov Chains: Stationary Equilibrium

- Let's consider

$$\vec{P}(S_0) = \begin{pmatrix} P(S_0 = A) & P(S_0 = E) \end{pmatrix} = \begin{pmatrix} 1 & 0 \end{pmatrix}$$

$$T = \begin{pmatrix} T(A|A) & T(A|E) \\ T(E|A) & T(E|E) \end{pmatrix} = \begin{pmatrix} 0.6 & 0.7 \\ 0.4 & 0.3 \end{pmatrix}$$



and calculate the first few transitions using $\vec{P}(S_t) = \vec{P}(S_{t-1})T$:

$$\begin{pmatrix} 0 \\ 1 \end{pmatrix} \rightarrow \begin{pmatrix} 0.7 \\ 0.3 \end{pmatrix} \rightarrow \begin{pmatrix} 0.63 \\ 0.37 \end{pmatrix} \rightarrow \begin{pmatrix} 0.637 \\ 0.363 \end{pmatrix} \rightarrow \begin{pmatrix} 0.6363 \\ 0.3637 \end{pmatrix} \rightarrow \dots$$

- In general, for the Markov chain to have a stationary equilibrium distribution, we need

$$P_{\text{eq}}(S_t = y) = P_{\text{eq}}(\text{stay at y}) \; \forall y$$

Hence $P_{\text{eq}}(\text{move to y}) - P_{\text{eq}}(\text{move from y})$ has to vanish.

Markov Chains: Stationary Equilibrium (cont.)

- in discrete case, we get:

$$\sum_{x \neq y} [P_{\text{eq}}(S_{t-1} = x)\, T(y|x) - P_{\text{eq}}(S_{t-1} = y)\, T(x|y))] = 0$$

Markov Chains: Stationary Equilibrium (cont.)

- in discrete case, we get:

$$\sum_{x \neq y} [P_{\text{eq}}(S_{t-1} = x) T(y|x) - P_{\text{eq}}(S_{t-1} = y) T(x|y))] = 0$$

- One sufficient (but not necessary) condition to satisfy this is:

$$P_{\text{eq}}(S_{t-1} = x) T(y|x) = P_{\text{eq}}(S_{t-1} = y) T(x|y))$$

which is called **detailed balance** condition.

Markov Chains: Stationary Equilibrium (cont.)

- in discrete case, we get:

$$\sum_{x \neq y} [P_{\text{eq}}(S_{t-1} = x)T(y|x) - P_{\text{eq}}(S_{t-1} = y)T(x|y))] = 0$$

- One sufficient (but not necessary) condition to satisfy this is:

$$P_{\text{eq}}(S_{t-1} = x)T(y|x) = P_{\text{eq}}(S_{t-1} = y)T(x|y))$$

  which is called **detailed balance** condition.
- if we want to sample a (non-tracable) probability function q using a Markov chain, we simply have to pick a transistion kernel with q as its equilibrium distribution !

MARCOV CHAIN MONTE CARLO SAMPLING

MCMC to the rescue!

- Markov chains can be used to randomly sample a posterior q/Z with the correct transition kernel T. But how to construct T?

MCMC to the rescue!

- Markov chains can be used to randomly sample a posterior q/Z with the correct transition kernel T. But how to construct T?
- Start with a proposal (or candidate) distribution $k(y|x)$

MCMC to the rescue!

- Markov chains can be used to randomly sample a posterior q/Z with the correct transition kernel T. But how to construct T?
- Start with a proposal (or candidate) distribution $k(y|x)$
- Using any k as T will not guarantee detailed balance, i.e. $q(x)k(y|x) \neq q(y)k(x|y)$ for some $x \neq y$

MCMC to the rescue!

- Markov chains can be used to randomly sample a posterior q/Z with the correct transition kernel T. But how to construct T?
- Start with a proposal (or candidate) distribution $k(y|x)$
- Using any k as T will not guarantee detailed balance, i.e. $q(x)k(y|x) \neq q(y)k(x|y)$ for some $x \neq y$
- Let's assume $q(x)k(y|x) > q(y)k(x|y)$ for some x, y

MCMC to the rescue!

- Markov chains can be used to randomly sample a posterior q/Z with the correct transition kernel T. But how to construct T?
- Start with a proposal (or candidate) distribution $k(y|x)$
- Using any k as T will not guarantee detailed balance, i.e. $q(x)k(y|x) \neq q(y)k(x|y)$ for some $x \neq y$
- Let's assume $q(x)k(y|x) > q(y)k(x|y)$ for some x, y
- Borrow from acceptance/rejection idea: Introduce acceptance criterion with acceptance prob. $\alpha(y|x)$ to lower probability on LHS and maximise it on RHS, i.e. construct transition kernel as follows:

$$T(y|x) = k(y|x)\alpha(y|x) + [1 - \alpha(y|x)]\delta_{x,y}$$

MCMC to the rescue!

- Markov chains can be used to randomly sample a posterior q/Z with the correct transition kernel T. But how to construct T?
- Start with a proposal (or candidate) distribution $k(y|x)$
- Using any k as T will not guarantee detailed balance, i.e. $q(x)k(y|x) \neq q(y)k(x|y)$ for some $x \neq y$
- Let's assume $q(x)k(y|x) > q(y)k(x|y)$ for some x, y
- Borrow from acceptance/rejection idea: Introduce acceptance criterion with acceptance prob. $\alpha(y|x)$ to lower probability on LHS and maximise it on RHS, i.e. construct transition kernel as follows:

$$T(y|x) = k(y|x)\alpha(y|x) + [1 - \alpha(y|x)]\delta_{x,y}$$

- In case above, we maximise RHS by setting $\alpha(x|y) = 1$, thus obtaining for $x \neq y$ for the detailed balance condition:

$$q(x)T = q(y)T \Leftrightarrow q(x)k(y|x)\alpha(x|y) = q(y)k(x|y)$$

$$\Leftrightarrow \alpha(y|x) = \frac{q(y)k(x|y)}{q(x)k(y|x)}$$

MCMC to the rescue! (cont.)

- If $q(x)k(y|x) < q(y)k(x|y)$, simply switch roles of x and y

MCMC to the rescue! (cont.)

- If $q(x)k(y|x) < q(y)k(x|y)$, simply switch roles of x and y
- Hence we obtain criterion

$$\alpha(y|x) = \begin{cases} \frac{q(y)k(x|y)}{q(x)k(y|x)} & \text{if } q(x)k(y|x) > q(y)k(x|y) \\ 1 & \text{otherwise} \end{cases}$$

MCMC - Metropolis-Hastings algorithm

- Given a target quasi-distribution $q(\Theta)$:
  1. Specify a proposal distribution $k(y|x)$

MCMC - Metropolis-Hastings algorithm

- Given a target quasi-distribution $q(\Theta)$:
  1. Specify a proposal distribution $k(y|x)$
  2. Choose a starting point Θ i.e. $S_{t=0} = \Theta$; set $t = 0$

MCMC - Metropolis-Hastings algorithm

- Given a target quasi-distribution $q(\Theta)$:
  1. Specify a proposal distribution $k(y|x)$
  2. Choose a starting point Θ i.e. $S_{t=0} = \Theta$; set $t = 0$
  3. Increment t

MCMC - Metropolis-Hastings algorithm

- Given a target quasi-distribution $q(\Theta)$:
  1. Specify a proposal distribution $k(y|x)$
  2. Choose a starting point Θ i.e. $S_{t=0} = \Theta$; set $t = 0$
  3. Increment t
  4. Draw a new state proposal $\Theta' \sim k(\Theta'|\Theta)$

MCMC - Metropolis-Hastings algorithm

- Given a target quasi-distribution $q(\Theta)$:
  1. Specify a proposal distribution $k(y|x)$
  2. Choose a starting point Θ i.e. $S_{t=0} = \Theta$; set $t = 0$
  3. Increment t
  4. Draw a new state proposal $\Theta' \sim k(\Theta'|\Theta)$
  5. Draw a uniform random number $u \in [0, 1]$

MCMC - Metropolis-Hastings algorithm

- Given a target quasi-distribution $q(\Theta)$:
  1. Specify a proposal distribution $k(y|x)$
  2. Choose a starting point Θ i.e. $S_{t=0} = \Theta$; set $t = 0$
  3. Increment t
  4. Draw a new state proposal $\Theta' \sim k(\Theta'|\Theta)$
  5. Draw a uniform random number $u \in [0, 1]$
  6. If $u < \frac{q(\Theta')k(\Theta|\Theta')}{q(\Theta)k(\Theta'|\Theta)}$, set $S_t = \Theta'$; else $setS_t = \Theta$

MCMC - Metropolis-Hastings algorithm

- Given a target quasi-distribution $q(\Theta)$:
  1. Specify a proposal distribution $k(y|x)$
  2. Choose a starting point Θ i.e. $S_{t=0} = \Theta$; set $t = 0$
  3. Increment t
  4. Draw a new state proposal $\Theta' \sim k(\Theta'|\Theta)$
  5. Draw a uniform random number $u \in [0, 1]$
  6. If $u < \frac{q(\Theta')k(\Theta|\Theta')}{q(\Theta)k(\Theta'|\Theta)}$, set $S_t = \Theta'$; else $setS_t = \Theta$
  7. Goto (3)

MCMC - Metropolis-Hastings algorithm

- Given a target quasi-distribution $q(\Theta)$:
  1. Specify a proposal distribution $k(y|x)$
  2. Choose a starting point Θ i.e. $S_{t=0} = \Theta$; set $t = 0$
  3. Increment t
  4. Draw a new state proposal $\Theta' \sim k(\Theta'|\Theta)$
  5. Draw a uniform random number $u \in [0, 1]$
  6. If $u < \frac{q(\Theta')k(\Theta|\Theta')}{q(\Theta)k(\Theta'|\Theta)}$, set $S_t = \Theta'$; else $setS_t = \Theta$
  7. Goto (3)
- The art of MCMC is in specifying the proposal distribution $k(y|x)$

MCMC - Random Walk Metropolis (example)

- Assume that k does not depend on state, but solely on difference between states i.e.

$$k(y|x) = K(y - x)$$

This is the jumping function for a random walk if always accepted.

MCMC - Random Walk Metropolis (example)

- Assume that k does not depend on state, but solely on difference between states i.e.

$$k(y|x) = K(y - x)$$

  This is the jumping function for a random walk if always accepted.

- Assume furthermore, that k is symmetric i.e.

$$k(y|x) = k(x|y)$$

  This simplifies our acceptance function to

$$\alpha(y|x) = \min\left(\frac{q(y)}{q(x)}, 1\right)$$

MCMC - Random Walk Metropolis (example)

- Assume that k does not depend on state, but solely on difference between states i.e.

$$k(y|x) = K(y - x)$$

  This is the jumping function for a random walk if always accepted.

- Assume furthermore, that k is symmetric i.e.

$$k(y|x) = k(x|y)$$

  This simplifies our acceptance function to

$$\alpha(y|x) = \min\left(\frac{q(y)}{q(x)}, 1\right)$$

- Popular choice: Multivariate Gaussian distribution

$$k(y|x) \simeq \exp(-\frac{1}{2}(y - x)^T \Sigma^{-1}(y - x))$$

MCMC - Random Walk Metropolis - Log densities

Switching to logarithms:

- When implementing MCMC, it is advisable to work with logarithmic densities instead as multiplications and ratios become sums and substractions

MCMC - Random Walk Metropolis - Log densities

Switching to logarithms:

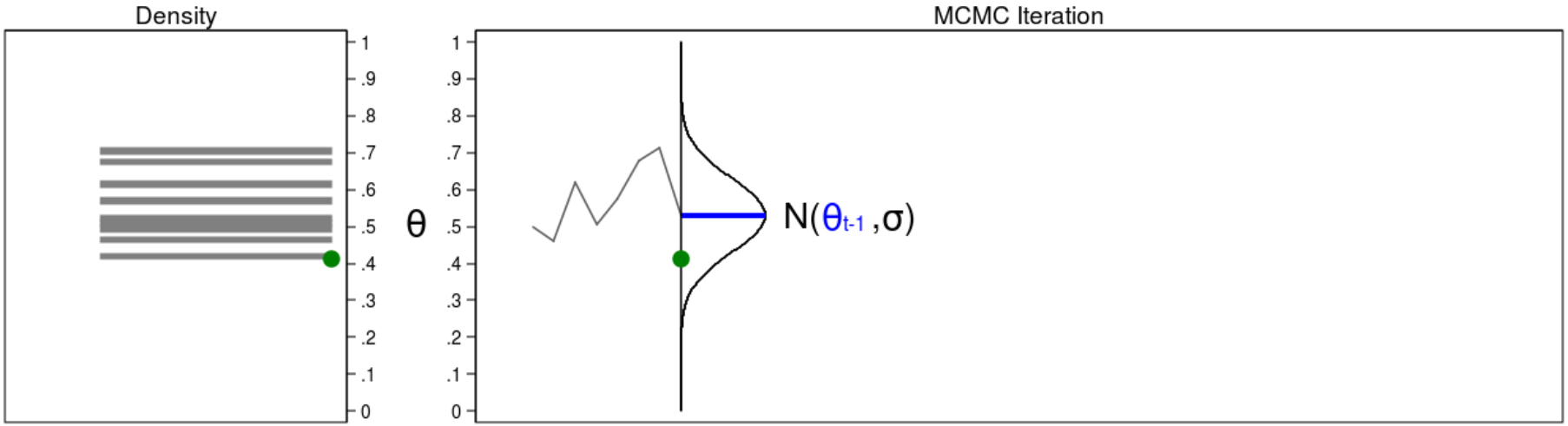
- When implementing MCMC, it is advisable to work with logarithmic densities instead as multiplications and ratios become sums and substractions
- This is important when dealing with a wide dynamical range to avoid over-/underflows (but may cause some occurrences of $-\infty$ we have to deal with)

MCMC - Random Walk Metropolis - Log densities

Switching to logarithms:

- When implementing MCMC, it is advisable to work with logarithmic densities instead as multiplications and ratios become sums and substractions
- This is important when dealing with a wide dynamical range to avoid over-/underflows (but may cause some occurrences of $-\infty$ we have to deal with)
- The acceptance-rejection step then reads:
  6. If $\ln u < \ln q(\Theta') - \ln q(\Theta)$, set $S_t = \Theta'$; else $set S_t = \Theta$

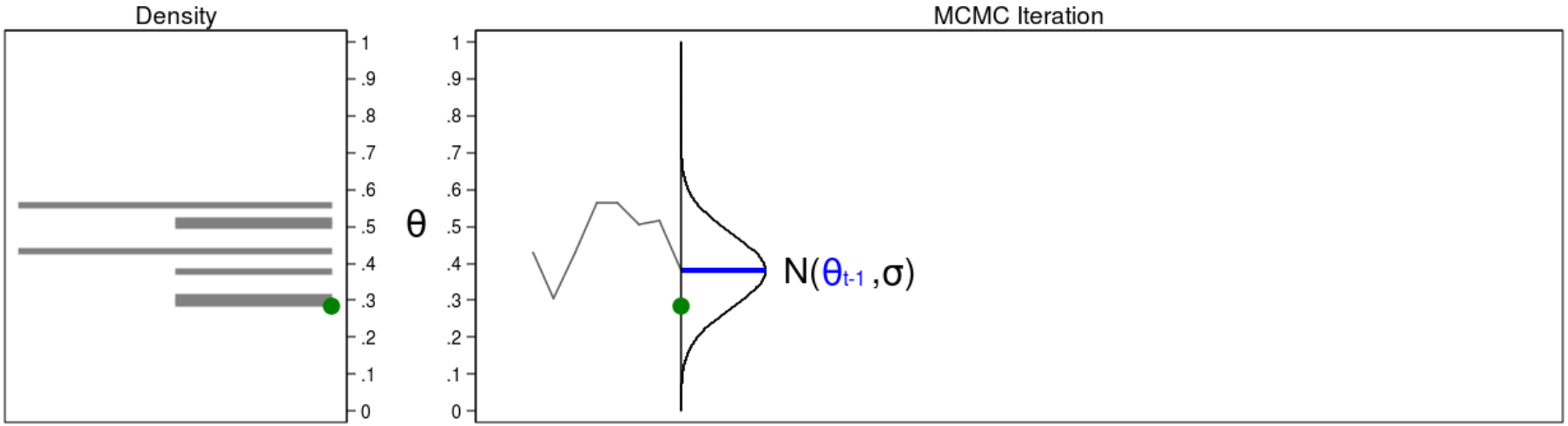
MCMC - Random Walk Metropolis (example)



Draw $\theta_t \sim \text{Normal}(\theta_{t-1}, \sigma)$

$\text{Normal}(0.530, \sigma) = 0.411$

MCMC - Random Walk Metropolis (example)



Step 1: $r(\theta_{new}, \theta_{t-1}) = \frac{\text{Posterior}(\theta_{new})}{\text{Posterior}(\theta_{t-1})} = \frac{\text{Beta}(1,1,0.286) \times \text{Binomial}(10,4,0.286)}{\text{Beta}(1,1,0.380) \times \text{Binomial}(10,4,0.380)} = 0.747$

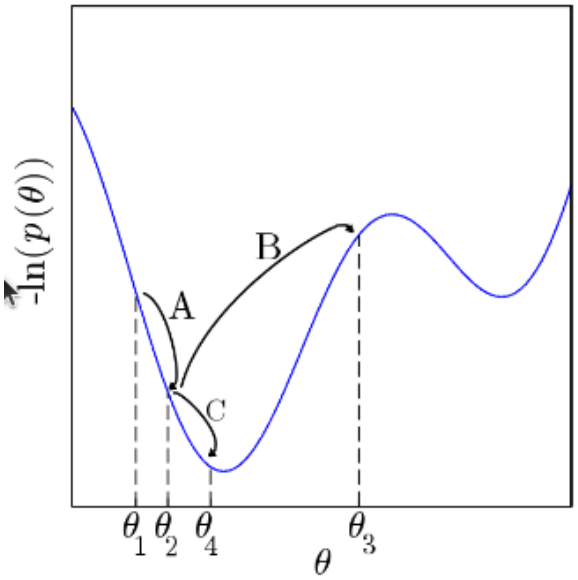
Step 2: Acceptance probability $\alpha(\theta_{new}, \theta_{t-1}) = \min\{r(\theta_{new}, \theta_{t-1}), 1\} = \min\{0.747, 1\} = 0.747$

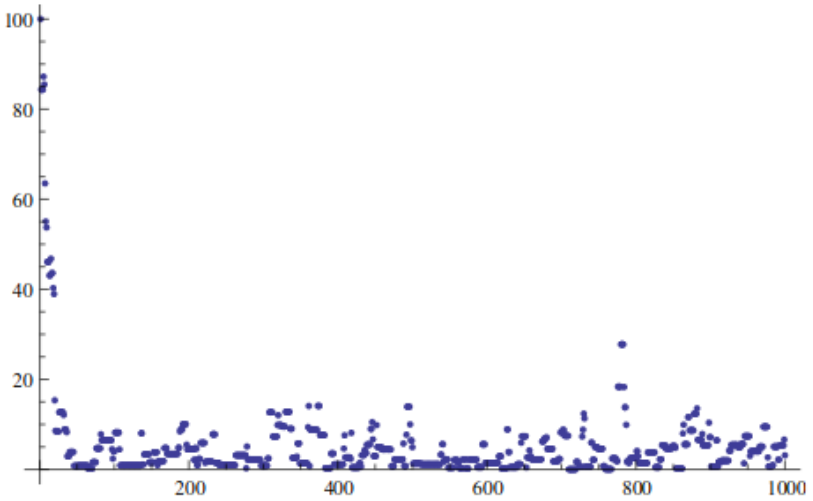
Step 3: Draw $u \sim \text{Uniform}(0,1) = 0.094$

Step 4: If $u < \alpha(\theta_{new}, \theta_{t-1})$ → If $0.094 < 0.747$ Then $\theta_t = \theta_{new} = 0.286$

Otherwise $\theta_t = \theta_{t-1} = 0.380$

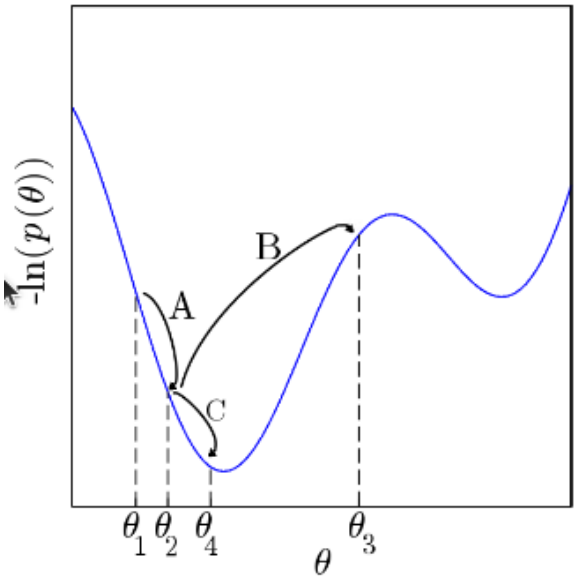
MCMC - Convergence

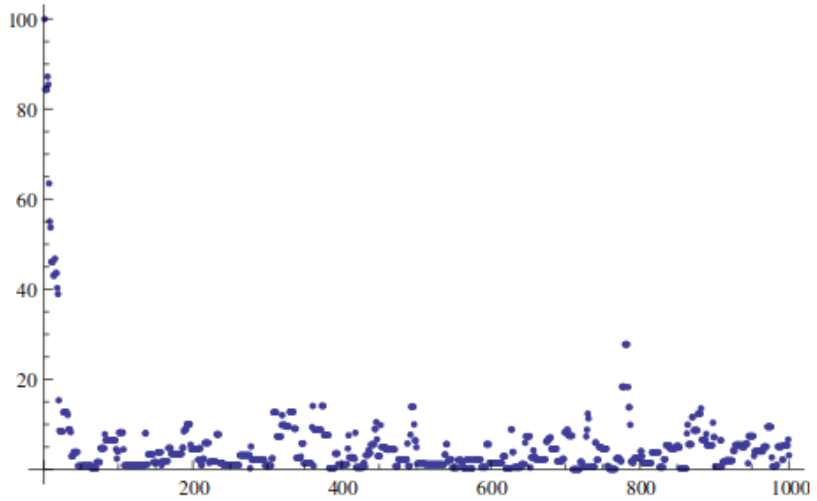




- How long do we have to run MCMC until sampling good approximation of underlying posterior?

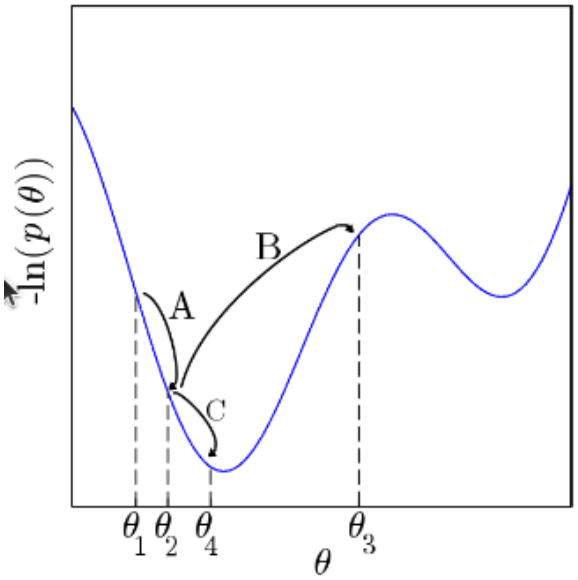
MCMC - Convergence

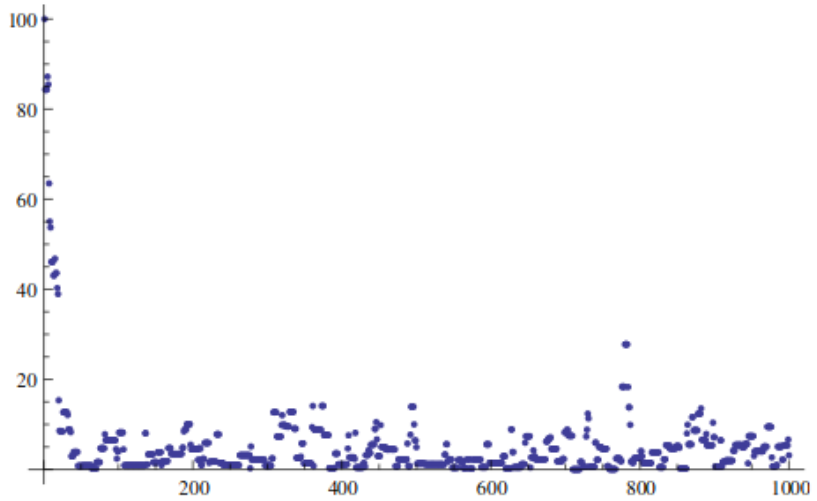




- How long do we have to run MCMC until sampling good approximation of underlying posterior?
- no simply/reliable answer to this key question; absolute convergence of pdf difficult to quantify/prove

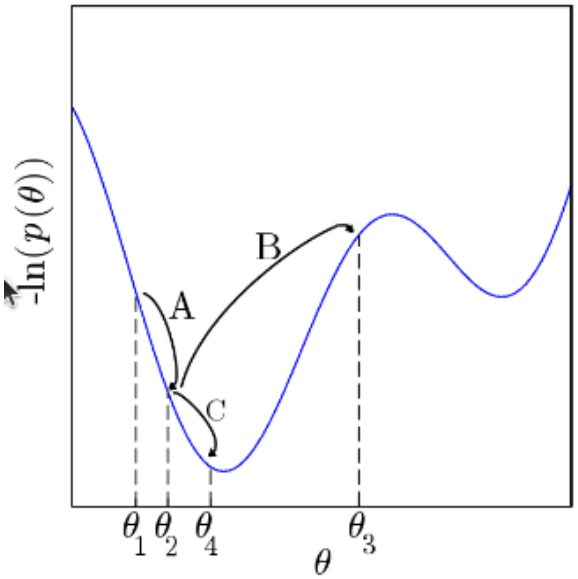
MCMC - Convergence

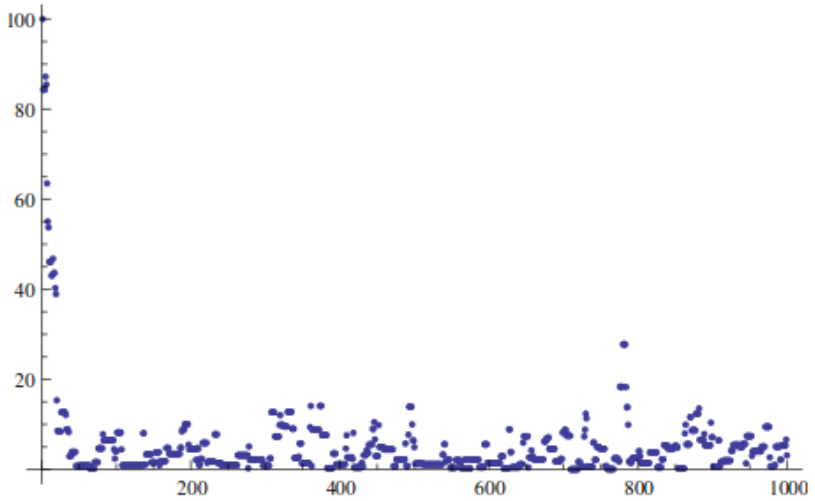




- How long do we have to run MCMC until sampling good approximation of underlying posterior?
- no simply/reliable answer to this key question; absolute convergence of pdf difficult to quantify/prove
- relying on heuristics e.g. to check whether your walker has traversed the high density reasons at least a couple of times. This implies e.g. any substantial subset of the chain shows the same post. morph.

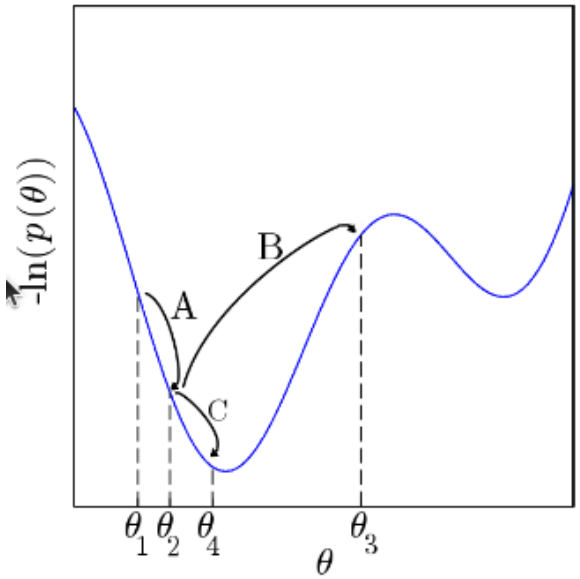
MCMC - Convergence (cont.)

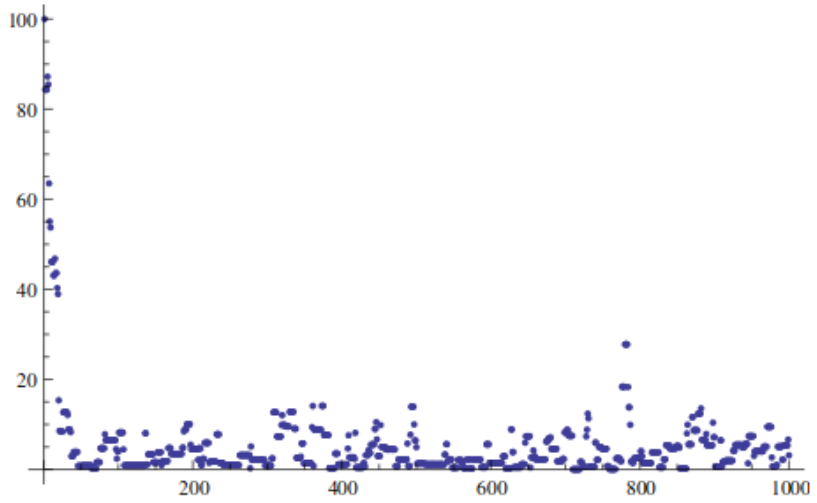




- This similarity can be quantified by e.g. determining deviations in means and variances between subsets

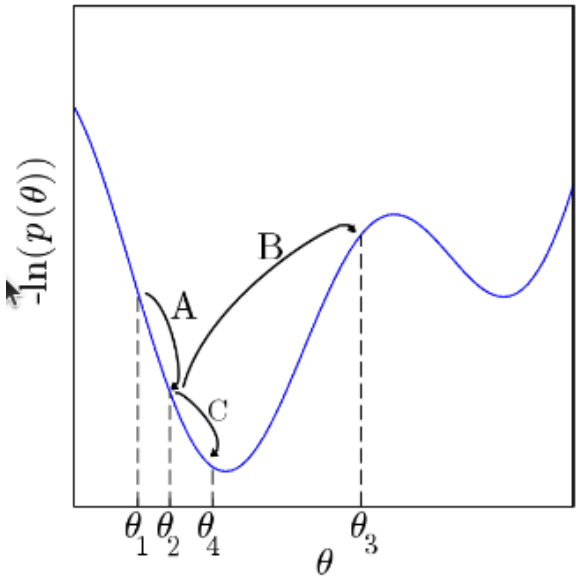
MCMC - Convergence (cont.)

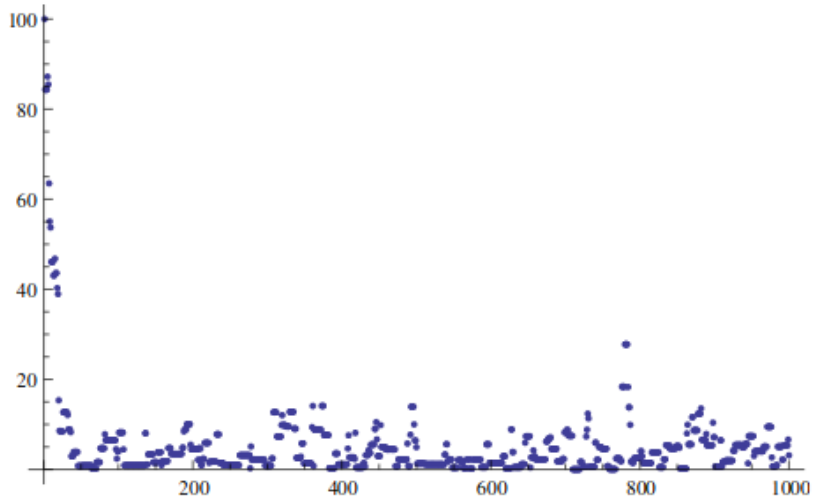




- This similarity can be quantified by e.g. determining deviations in means and variances between subsets
- Most important tool is the so-called **autocorrelation time** τ_X. It tells us how many steps it takes for a chain to ensure that samples on both end of the interval are (virtually) independent; is 2-point statistic, thus requires significant amount of data to be estimated precisely.

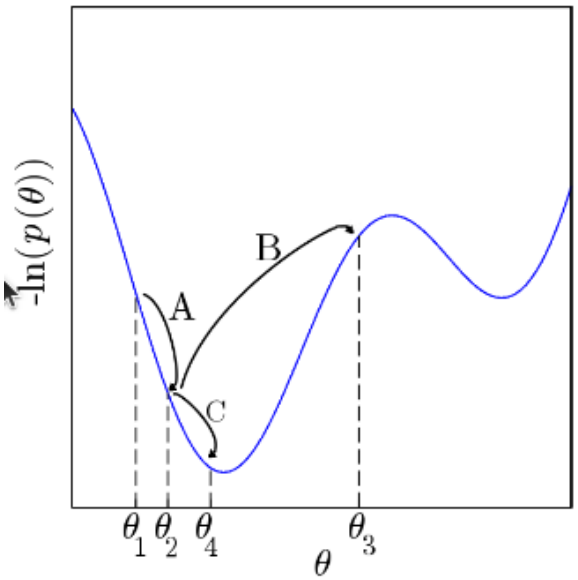
MCMC - Convergence (cont.)

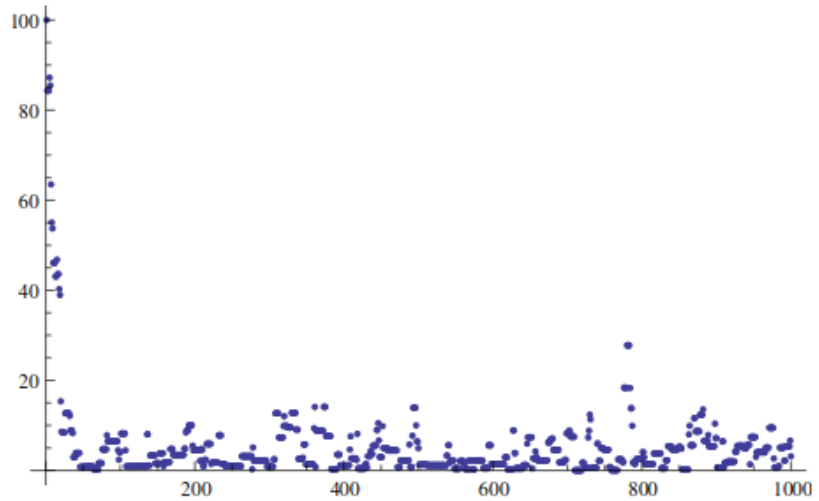




- This similarity can be quantified by e.g. determining deviations in means and variances between subsets
- Most important tool is the so-called **autocorrelation time** τ_X. It tells us how many steps it takes for a chain to ensure that samples on both end of the interval are (virtually) independent; is 2-point statistic, thus requires significant amount of data to be estimated precisely.
- For multiple chains, there is the **Gelman-Rubin diagnostic** which compares the variance with a chain with that across chains

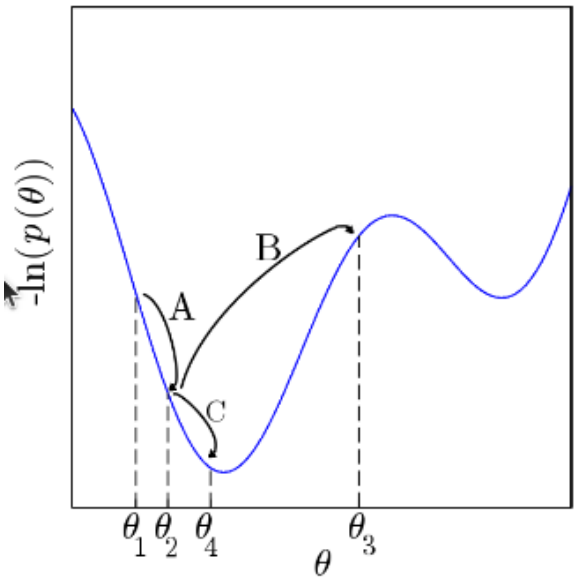
MCMC - Initialisation/Burn-In

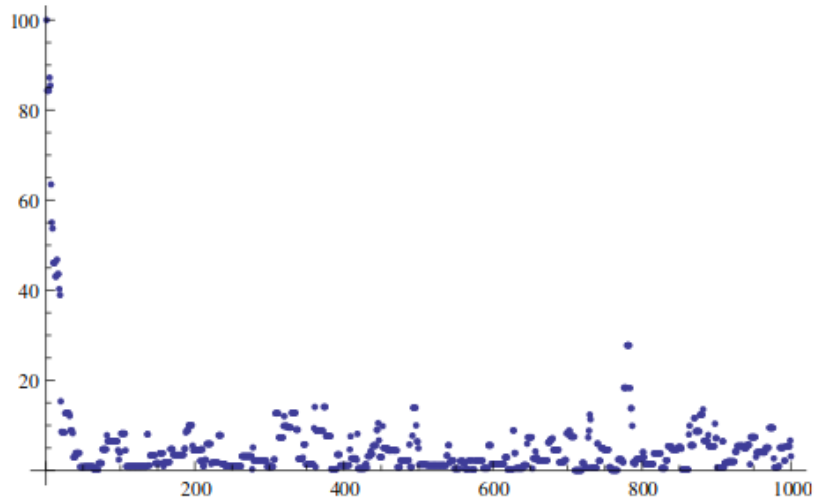




- Initial state is often randomly chosen, but not from the underlying posterior, thus may end up in a "non-typical" place and then over-samples the region around initial $\rightarrow$ not good sample of posterior

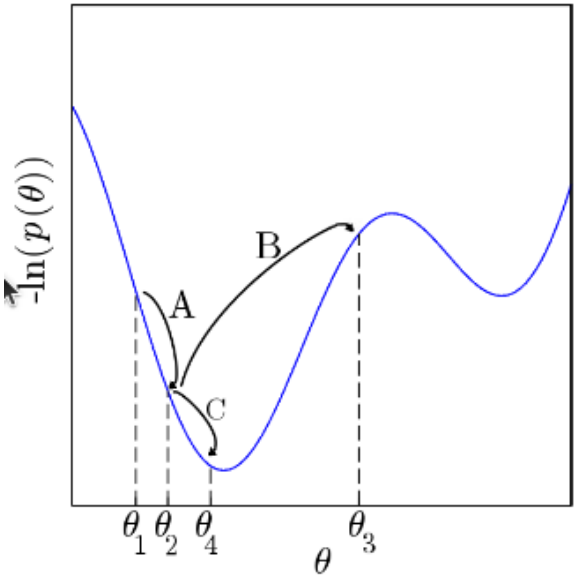
MCMC - Initialisation/Burn-In

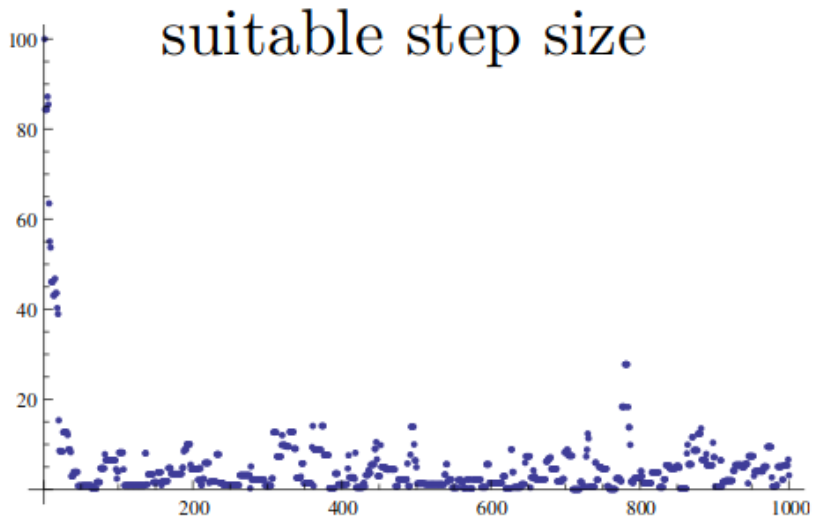




- Initial state is often randomly chosen, but not from the underlying posterior, thus may end up in a "non-typical" place and then over-samples the region around initial $\rightarrow$ not good sample of posterior
- Ignore/discard initial steps aka **burn-in** before you do any inference on the chain(s)

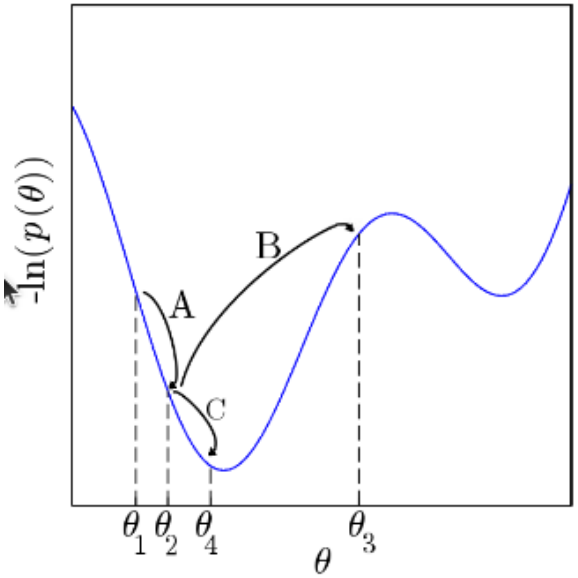
MCMC - Tuning

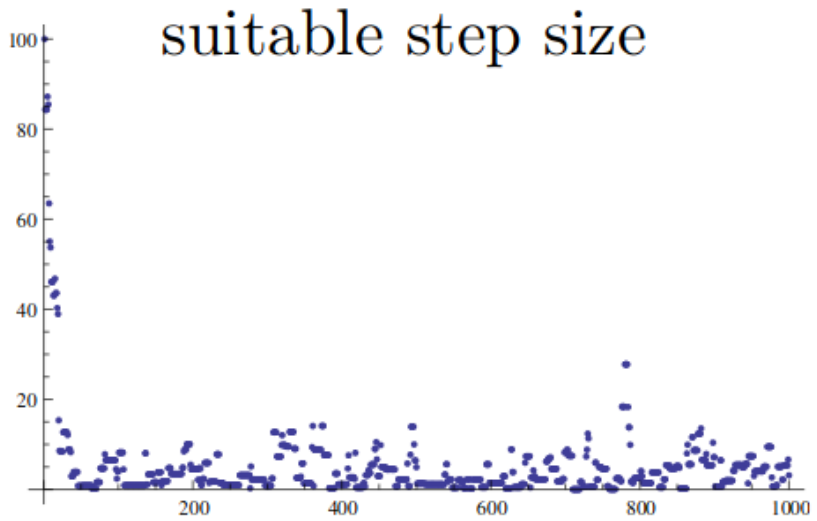




- The proposal function k can have various parameters to determine how the MC random-walks through the parameter space

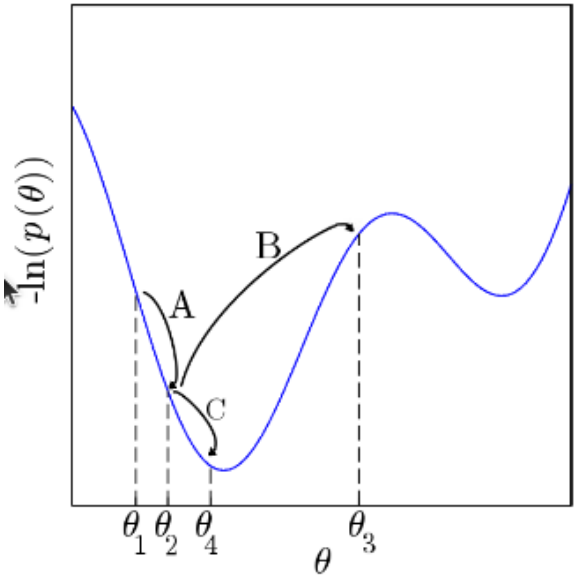
MCMC - Tuning

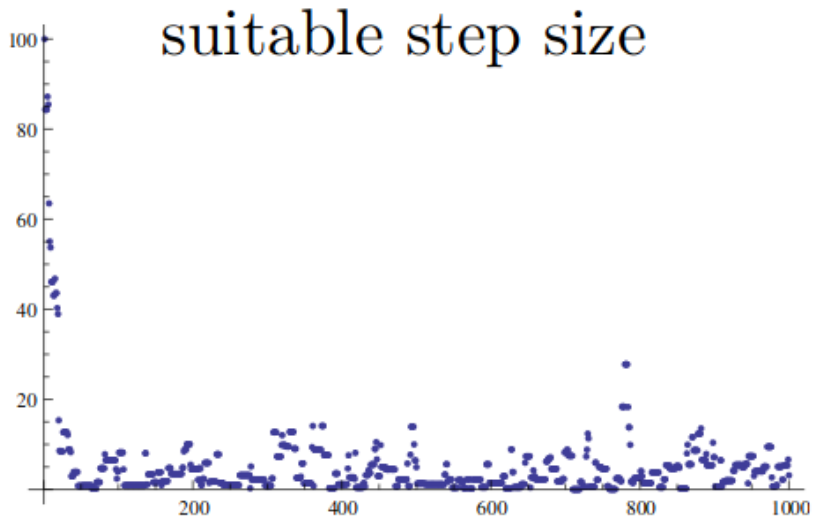




- The proposal function k can have various parameters to determine how the MC random-walks through the parameter space
- e.g. for the multivariante Gaussian distribution

MCMC - Tuning





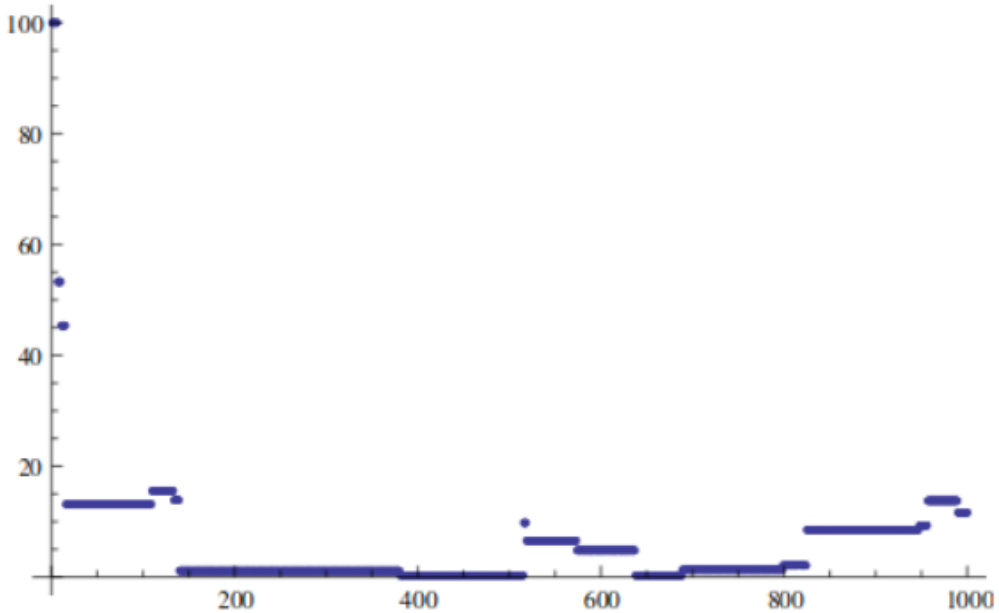
- The proposal function k can have various parameters to determine how the MC random-walks through the parameter space
- e.g. for the multivariante Gaussian distribution

$$k(y|x) \simeq \exp(-\frac{1}{2}(y-x)^T\Sigma^{-1}(y-x))$$

there are $D(D+1)/2$ parameters in the $D \times D$ sym.,pos.def. covariance matrix to be set

- Question: How to find the optimal parameters ?

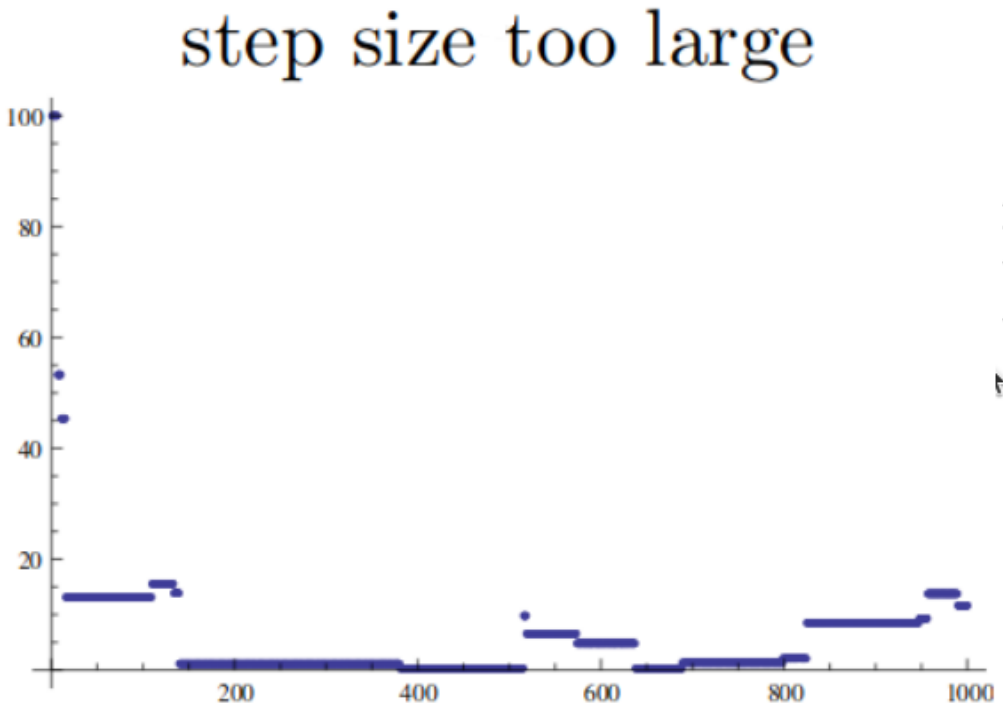
MCMC - Tuning (Problem 1)

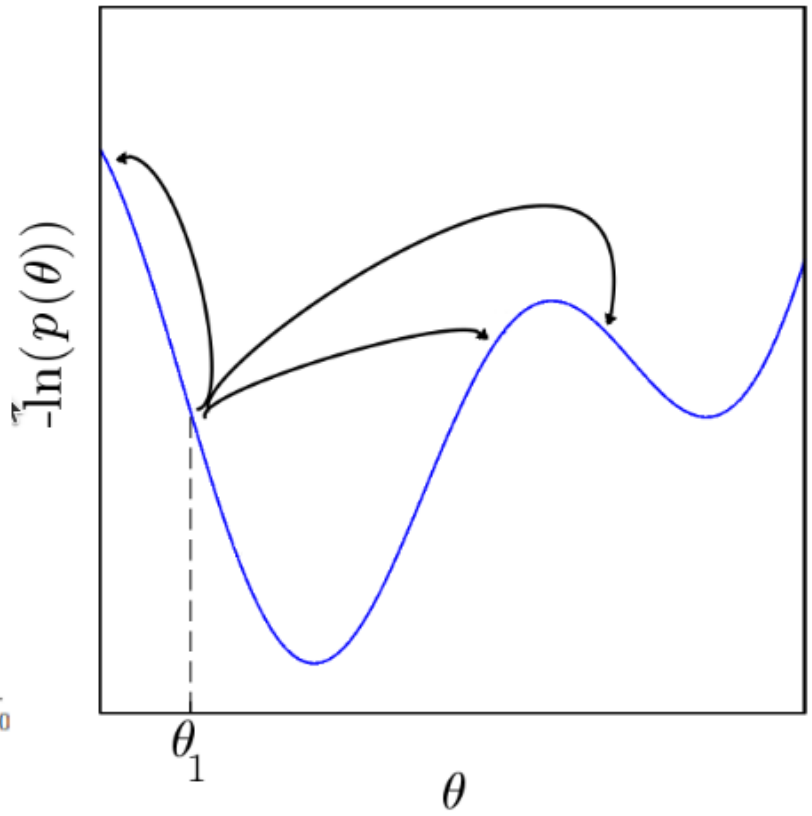


WHAT IS THE PROBLEM HERE?

- if distribution is **too wide**, while steps cover parameter space easily, almost all proposals are rejected

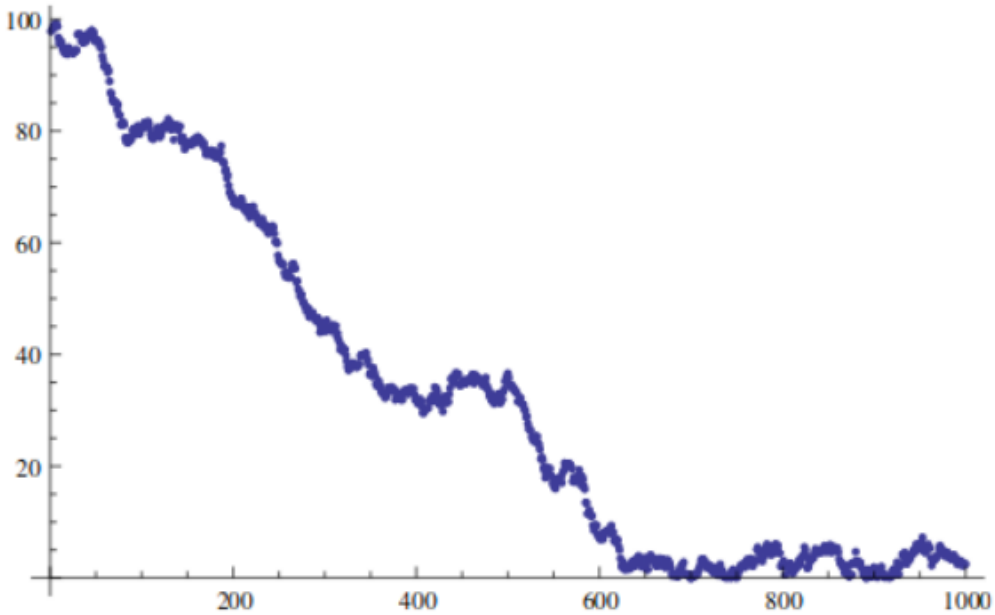
MCMC - Tuning (Problem 1)





- if distribution is **too wide**, while steps cover parameter space easily, almost all proposals are rejected

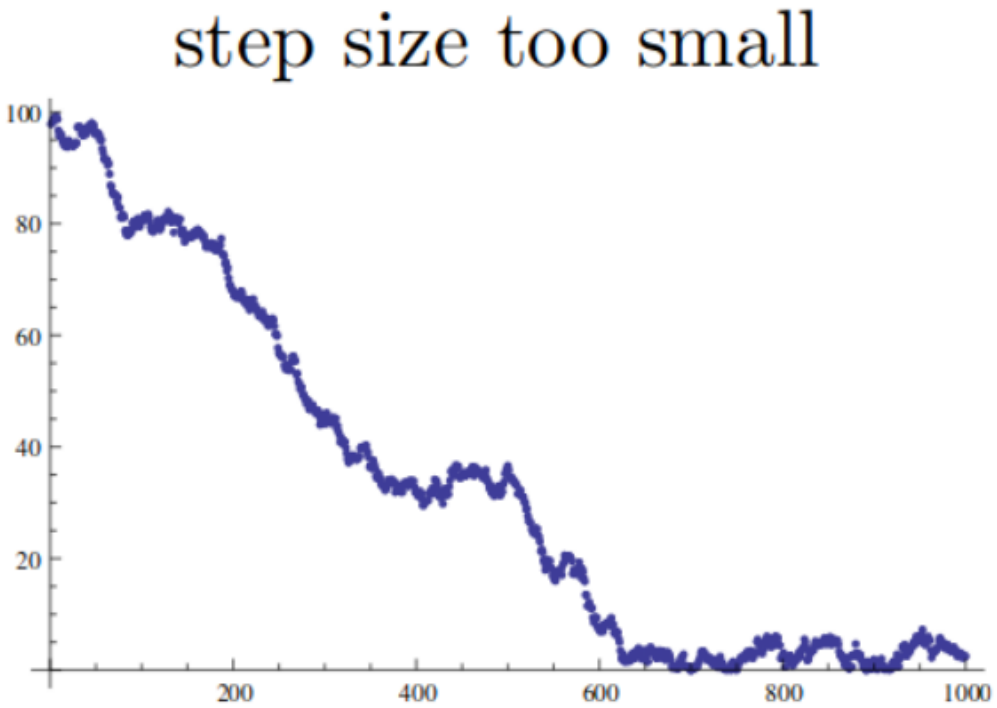
MCMC - Tuning (Problem 2)

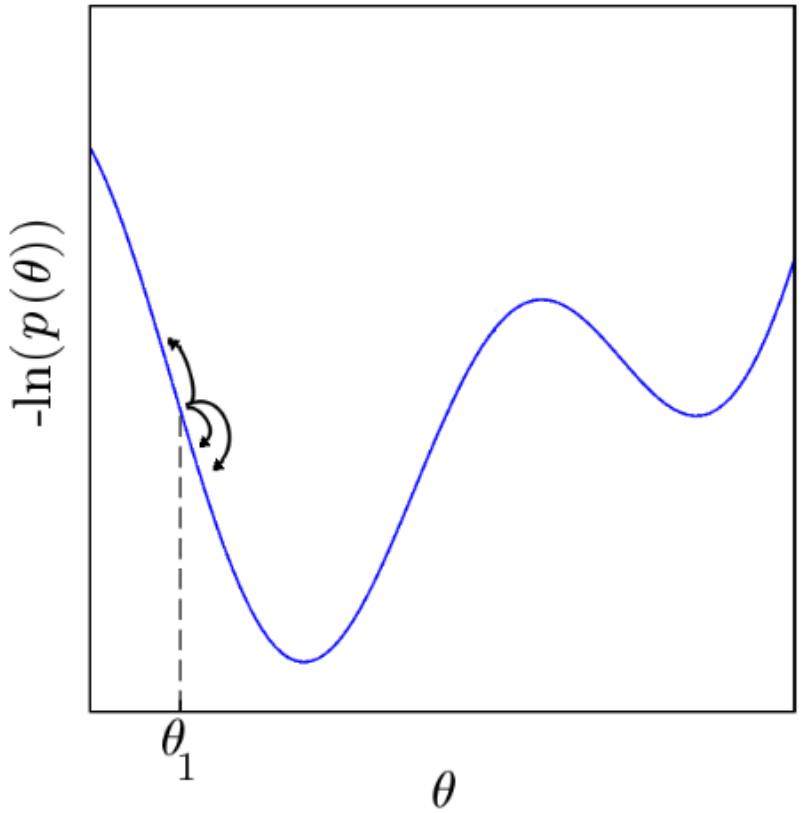


WHAT IS THE PROBLEM HERE?

- if distribution is **too narrow**, almost all proposals are accepted, but full exploration of parameter space takes very long

MCMC - Tuning (Problem 2)





- if distribution is **too narrow**, almost all proposals are accepted, but full exploration of parameter space takes very long

MCMC - Tuning (cont.)

- Best parameters for proposal function are those that minimize the autocorrelation time τ_x

MCMC - Tuning (cont.)

- Best parameters for proposal function are those that minimize the autocorrelation time τ_x
- As previously mentioned, τ_x can be difficult to measure without significantly huge amounts of data. Instead we can use proxy statistics:

  Acceptance fraction With shrinking step size the acceptance ratio grows and vice versa. The Goldilock value is roughly between 0.5 and 0.25. Can be used during the Burn-in phase to update the parameters if observed acceptance differs significantly.

MCMC - Tuning (cont.)

- Best parameters for proposal function are those that minimize the autocorrelation time τ_x
- As previously mentioned, τ_x can be difficult to measure without significantly huge amounts of data. Instead we can use proxy statistics:

  Acceptance fraction With shrinking step size the acceptance ratio grows and vice versa. The Goldilock value is roughly between 0.5 and 0.25. Can be used during the Burn-in phase to update the parameters if observed acceptance differs significantly.

  You can decompose the D-dim. tuning problem into D 1-dim. problems by updating each model parameter separately.

  Accepted Squared Jump Distance This is the mean squared distance the walker moves and maximises when the acceptance rate is reasonable and the step size is large - easy to measure and well correlated with the autocorrelation time

MCMC - Final Notes

We saw that MCMC is by construction a fair sampler for our posterior (or any other) probability, which allows us to integrate over the probability function (e.g. to calculate the mean or median), but we have to also point out, what it is not ...

MCMC - Final Notes

We saw that MCMC is by construction a fair sampler for our posterior (or any other) probability, which allows us to integrate over the probability function (e.g. to calculate the mean or median), but we have to also point out, what it is not ...

- It is **NOT** a **good search algorithm**. Chains are not guaranteed to find/sample every local maximum, let alone a global one in the whole parameter space.

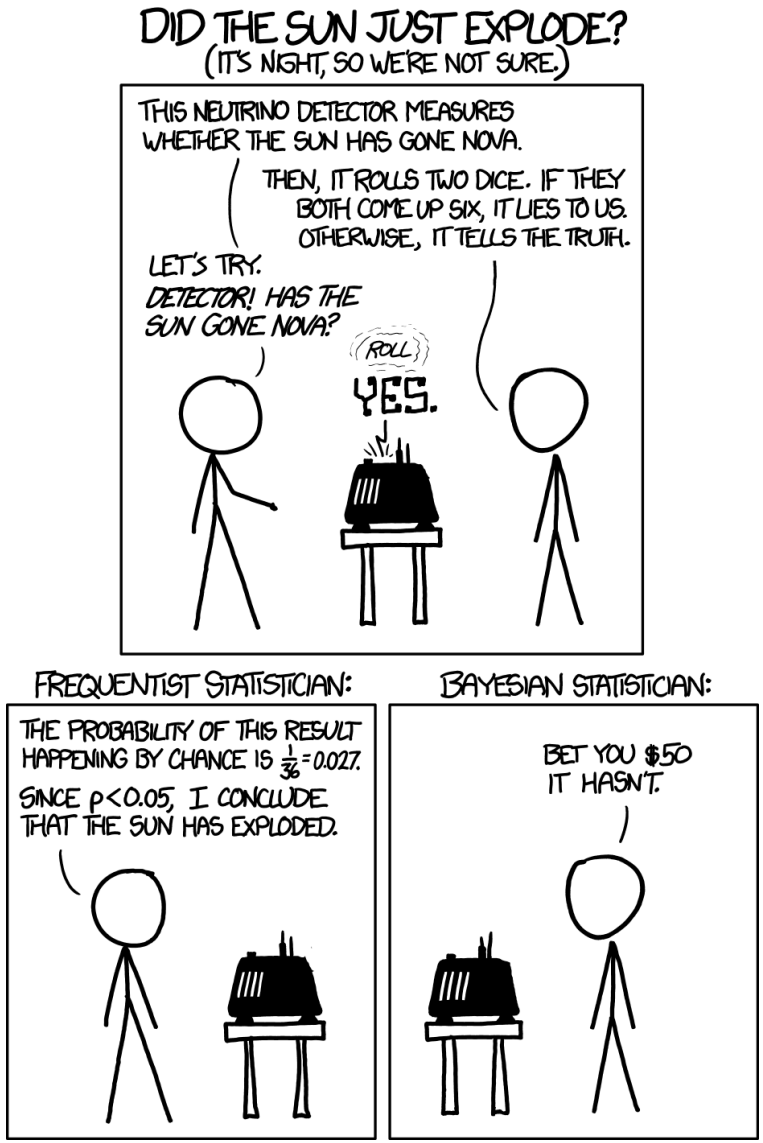
MCMC - Final Notes

We saw that MCMC is by construction a fair sampler for our posterior (or any other) probability, which allows us to integrate over the probability function (e.g. to calculate the mean or median), but we have to also point out, what it is not ...

- It is **NOT** a **good search algorithm**. Chains are not guaranteed to find/sample every local maximum, let alone a global one in the whole parameter space.
- It is definitely also **NOT** a **good optimizer** i.e. generically samples will not lie close the maximum of the sampled pdf (gets worse with higher dimensionality) $\Rightarrow$ "Best-Fit" value is meaningless.

Web-/Bibliography

- **Hogg et al.,Data Analysis Recipes using MCMC** arXiv:1710.06068
- **Leclercq et al., Cosmology: from theory to data, from data to theory** arXiv:1403.1260v3
- **Loredo,T., Lectures on Bayesian inference & MCMC** `https://astrostatistics.psu.edu/su14/lectures/`


DID THE SUN JUST EXPLODE?
(IT'S NIGHT, SO WE'RE NOT SURE.)
THIS NEUTRINO DETECTOR MEASURES WHETHER THE SUN HAS GONE NOVA.
THEN, IT ROLLS TWO DICE. IF THEY BOTH COME UP SIX, IT LIES TO US. OTHERWISE, IT TELLS THE TRUTH.
LET'S TRY.
DETECTOR! HAS THE SUN GONE NOVA?
ROLL
YES.
FREQUENTIST STATISTICIAN:
THE PROBABILITY OF THIS RESULT HAPPENING BY CHANCE IS 1/36 = 0.027.
SINCE p<0.05, I CONCLUDE THAT THE SUN HAS EXPLODED.
BAYESIAN STATISTICIAN:
BET YOU $50 IT HASN'T.