Jump'n'Run

A speedrun through the world of Bayesian Statistics and MCMC sampling

Kathrin Grunthal

Fundi Tutorials 2025

DISCLAIMER and OVERVIEW

- Disclaimer:
  - This talk is a very subjective overview on the matter of Bayesian statistics and MCMC sampling
  - It's a loose collection of knowledge
    - mostly "learning by doing"
    - I tried to be more professional using lecture notes, "A Student's guide to Bayesian Statistics" and a bunch of blog entries

- Overview
  - Why do we need statistics after all, and what are we actually trying to do all day long?
  - What's the concept of Bayesian statistics, and is there more?
  - Why do we need MCMC sampling?
  - How can we understand samplers and use them efficiently?

Why care for statistics?

Astronomy: make **general** statements about the Universe, based on **specific** examples of its behaviour

- Astronomy/Astrophysics is unusual, because it is generally not an experimental science
  - Can’t add carbon to a star and see what happens

- Improving our state of knowledge …
  - … by incorporating new information into our physical models
  - … do so via “plausible reasoning”

Types of reasoning

(A) Alcyone is within 200pc of Earth

(B) Alcyone is a member of the Pleiades

(C) All stars within the Pleiades are within 200pc of Earth

Assume (C) is true
→ "hypothesis space" in which we are reasoning

DEDUCTIVE REASONING

- if (B) is true, then (A) is true
  if (A) is false, then (B) is false
- (A) is a logical consequence of (B) and (C)

BUT:
What can we say about (B) if (A) [and (C)] are true?

PLAUSIBLE REASONING

- if (A) is true, then (B) is more plausible
  if (B) is false, then (A) is less plausible

➢ Basis of physical model building

➢ NEED TO BRING THAT INTO MATHEMATICAL FORM

What kind of statistician are you?

BAYESIAN

- Logical reasoning that uses Bayes' Theorem
- Interpretation of 'probability':
  - 'degree of belief'
  - Number between 0 and 1 measuring the plausibility of a proposition when incomplete knowledge means we cannon know its truth or falsehood
- Both the youngest and oldest interpretation
  - Original idea introduced by Laplace, Bernoulli and Bayes
  - Eclipsed by 'frequentist interpretation' until it was put on firmer footing by Jeffreys (1939) and Jaynes (1950's)

FREQUENTIST

- One attempt to remove the subjectivity from Bayesian statistics
- A frequentist equates probability to a limiting relative frequency

  Rel. frequency (A) = events (A) / total number of tries

- Assumptions:
  - All experiments are done under the same conditions
  - Limit converges
  - Past frequencies predict future frequencies

Bayesian or Frequentist?

	Bayesian	Frequentist
Model parameter	Is a random variable	Is not a random variable
Jargon	Credibility interval; prior; posterior	Confidence interval; p-value, significance
Goal	Decide on an opinion to have, based on a prior belief	Decide on an action to take, compared to a default action
Pros	Intuitive definitions of concepts	Makes sense to talk about method quality and getting the answer right
Give up	Lose ability to talk about right answers; no such thing as statistically significant, or rejecting the null, only "more likely" and "less likely"	Core concepts are more difficult to understand and apply

Bayesian or Frequentist?

- Neither is better, they are two competing interpretations
- Neither is more objective, both are based on assumptions

- "When you have a small sample, you should use Bayesian Statistics!"
  - Frequentist approach is only usable with a sample size that is large enough
  - it is possible to proceed with as little as one data point in the Bayesian approach
    
    - only works because you use a lot of initial assumptions
    - Statistics is not Alchemy! It's not possible to gain more information by using a different interpretation.

Bayesian inference – Bayes' rule

Likelihood
- ❖ model expected to describe the data
- ❖ Probability we would have seen what we saw, assuming the validity of the hypothesis

Prior probability
- ❖ State of knowledge prior to acquiring the data

$$p(\text{hypothesis} \mid \text{data}, I) = \frac{p(\text{data} \mid \text{hypothesis}) \times p(\text{hypothesis} \mid I)}{p(\text{data} \mid I)}$$

Posterior probability
- ❖ Probability of the hypothesis/model parameters given the observed data

Evidence
- ❖ normalisation of the posterior
- ❖ Probability for a future data set given our choice of model

The likelihood

- What is the likelihood?
  - probability model to approximate a real-world process
  - represents the set of assumptions we make in our analysis

- Why is $p(\text{data}|\theta)$ a "likelihood" and not "probability"?
  - If we hold the parameters fixed, the resulting distribution of possible data samples is a valid PDF.
  - Bayesian inference: keep the data fixed, let the model parameters vary → resulting distribution must not be a valid PDF.
  - Emphasize this via: $\mathcal{L}(\theta|data) = p(data|\theta)$

- How to choose the likelihood?
  1. Evaluate the real-life behaviour the model should be capable of explaining & note down the necessary assumptions
  2. Choose a suitable distribution function (e.g. Chapter 8, A Student's guide to Bayesian Statistics)
  3. AFTER FITTING: test the model's ability to explain the data, and if necessary, choose a new model

Example: PTA likelihood

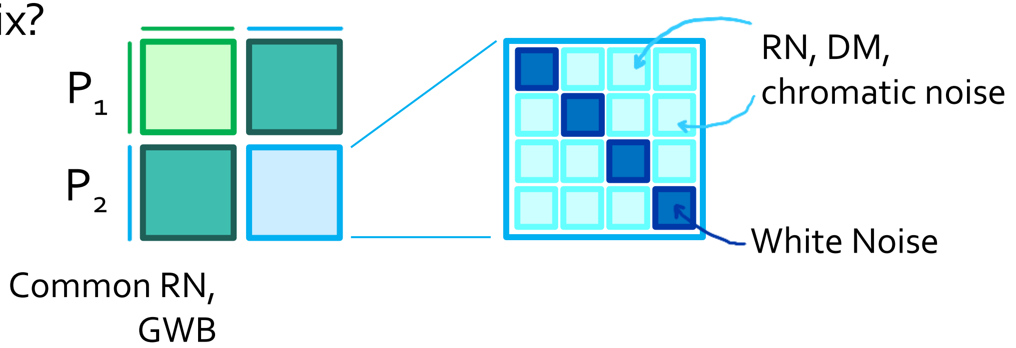
- Which distribution do we expect?
  - Model fit → describe the distribution of the **residuals**
  - if our timing model is correct, the residuals should be distributed like a **Gaussian around zero**

$$\mathcal{L} \sim \exp\left(\frac{1}{2}\,\overrightarrow{\delta t}^{T}\,\boldsymbol{C}^{-1}\,\overrightarrow{\delta t}\right)$$

- How are the residuals calculated?

$$\overrightarrow{\delta t} = \vec{t}_{obs} - \vec{t}_{theo} \simeq \vec{t}_{obs} - \boldsymbol{M}\vec{\epsilon} - \vec{d}(\theta)$$

- What is in the covariance matrix?



The prior

- What is the prior?
  - Represents our pre-data uncertainty for a parameter's true value
  - Needs to be a valid probability distribution!
  - Most controversial aspect of Bayesian statistics, due to their inherent subjectivity

- Why do we even need a prior?
  - Bayes' rule is only a way to update our initial belief in light of data → we must specify this initial belief

- Why can't we use a unity prior (in general)?
  - On the first sight, this sounds like a good idea, because it would apparently remove the criticized subjectivity
  - But: unbound, continuous parameter: $\int_{-\infty}^{+\infty} p(\theta)\mathrm{d}\theta = \infty$, and the prior must be a valid pdf!

- Construction of priors: uninformative & informative priors

The posterior

- What is the posterior
  - Golden goal of Bayesian inference
  - PDF that allows us to calculate expectation values, credible intervals etc. for our model parameters given the data that we have observed
  - Allows us to predict future data

- We have to ensure that it is a valid PDF!

The evidence

- What is the evidence?
  - Denominator in Bayes' rule
  - Normalising factor: Likelihood is not a valid PDF, and thus the object likelihood x prior is equally none
    ensure that the integral over the posterior is 1 → normalise with $\int_{\text{all }\theta} p(\text{data}|\theta)p(\theta)\mathrm{d}\theta$
  - Probability distribution: PDF for a future data set given our model, since $p(data) = \int_{\text{all }\theta} p(\text{data},\theta)\mathrm{d}\theta$

- The problem with the evidence
  - For relatively complex models, the computation of the integral becomes increasingly difficult
  - Example: model the exam scores, where score_{ij} for person i in school j is normally distributed as $\text{score}_{ij} \sim \mathcal{N}(\mu_j, \sigma_j)$

$$p(\text{data}) = \int_{\mu_1} \int_{\sigma_1} \dots \int_{\mu_{100}} \int_{\sigma_{100}} \mathrm{d}\mu_1 \mathrm{d}\sigma_1 \dots \mathrm{d}\mu_{100} \mathrm{d}\sigma_{100} \quad p(\text{data}|\mu_1, \sigma_1, \dots, \mu_{100}, \sigma_{100}) \times p(\mu_1, \sigma_1, \dots, \mu_{100}, \sigma_{100})$$

eValUAtE tHe EVidEnCE ^_^

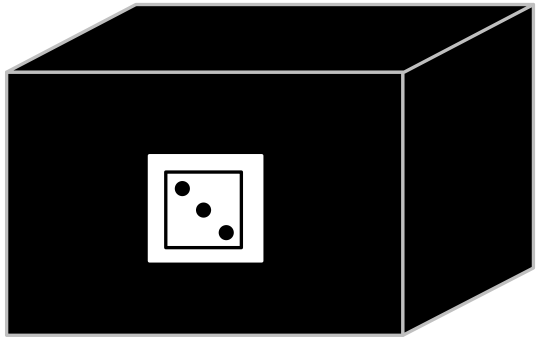
- Create a grid over the parameter space, calculate the posterior at each grid point
  - properly realisable for discrete random variables, more difficult for continuous parameters
  - Inefficient: area of parameter space that is relevant is likely small compared to the total grid
  - not feasible for large dimensions, as the numerical expense grows exponentially with the number of dimensions

- conjugate priors
  - Choose the prior such that, given your likelihood function, the posterior is in the same family of distributions
  - https://en.wikipedia.org/wiki/Conjugate_prior
  - not really useful in practice…

Or is there something more clever?

Integration via independent sampling

- Example:



Throw the die multiple times → Take sample mean → Estimate the true mean

- Mathematically speaking: $E(X) = \int_{-\infty}^{\infty} x\, p(x)\, \mathrm{d}x \approx \frac{1}{n}\sum_{i=1}^{n} X_i$
- Generalise to ANY function g(X) $E(g(X)) = \int_{-\infty}^{\infty} g(x)\, p(x)\, \mathrm{d}x \approx \frac{1}{n}\sum_{i=1}^{n} g(X_i)$
- Generalise to ANY dimensionality $E\left(g(\vec{X})\right) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} g(\vec{x})p(\vec{x})\, \mathrm{d}x_1 \mathrm{d}x_2 \dots \mathrm{d}x_k \approx \frac{1}{n}\sum_{i=1}^{n} g(\overrightarrow{X_i})$

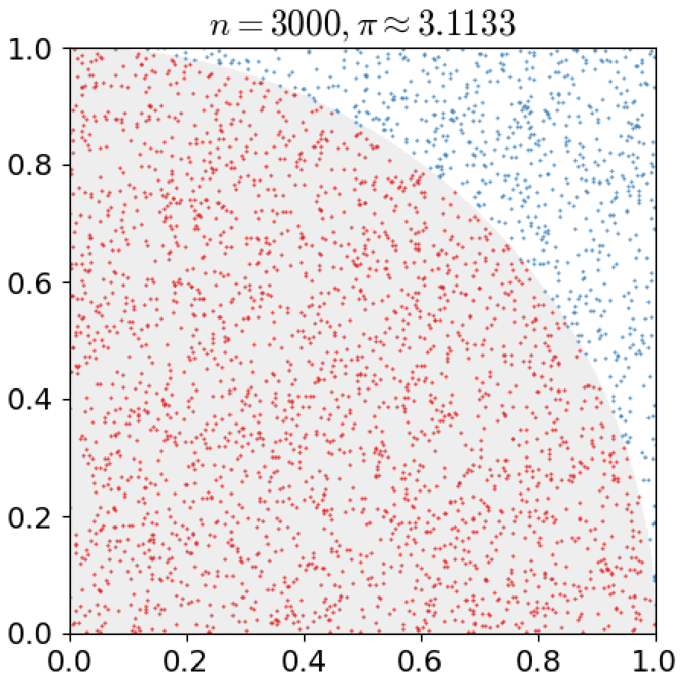
We can approximate multidimensional integrals like in Eq. (1),
as long as we can generate INDEPENDENT SAMPLES from the PDF

Monte Carlo simulation

- use randomness to solve problems that might be deterministic by relying on repeated random sampling

- General procedure
  1. define domain of possible inputs
  2. generate random inputs from PDF over the domain
  3. deterministic classification/computation of the outputs
  4. aggregate results

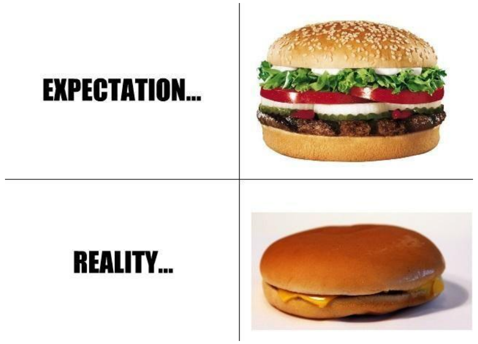
- Main applications:
  - optimisation
  - numerical integration
  - generating draws from a PDF



https://de.wikipedia.org/wiki/Monte-Carlo-Simulation#/media/Datei:Pi_monte_carlo_all.svg

Independent sampling

- Computers are deterministic machines → Random numbers are ALWAYS pseudo-random numbers
  - Rejection sampling: draw sample point (x,y) from the range of interest, accept if $y < p(x)$
  - Inverse transform sampling: sample $x \sim \text{Uniform}(0,1)$, calculate $y = \text{CDF}^{-1}(x)$

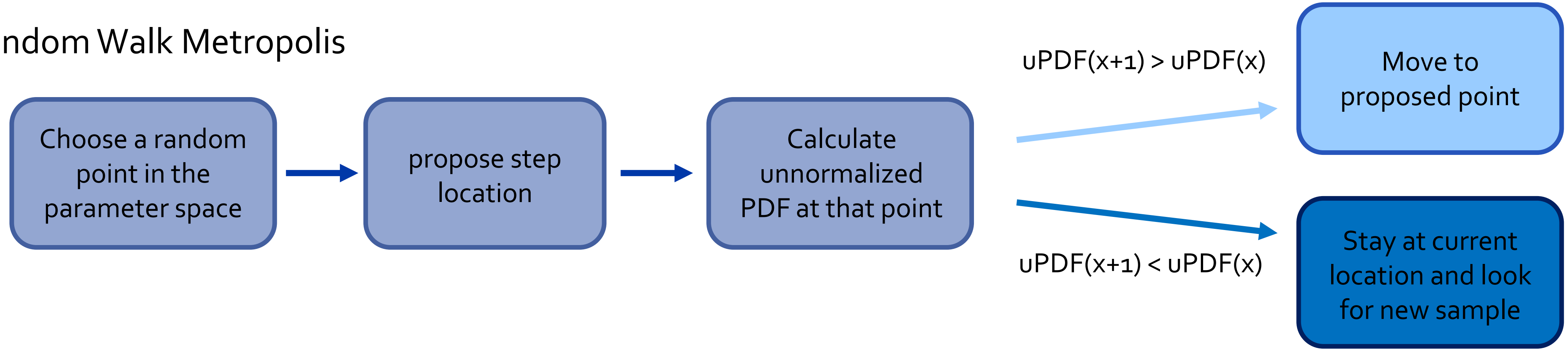


- What do we want our samples to look like?
  - $\frac{p(\theta_A|data)}{p(\theta_B|data)} = \frac{3}{1}$ → our sampler should generate 3 times more often from θ_A than from θ_B
  - Need only the relative height of the posterior, not the absolute
  - Bayes Theorem revisited: $\frac{p(\theta_A|data)}{p(\theta_B|data)} = \frac{\frac{\mathcal{L}(data|\,\theta_A)\times p(\theta_A)}{p(data)}}{\frac{\mathcal{L}(data|\,\theta_B)\times p(\theta_B)}{p(data)}} = \frac{\mathcal{L}(data|\,\theta_B)\times p(\theta_B)}{\mathcal{L}(data|\,\theta_B)\times p(\theta_B)}$

➢ knowledge of the UNNORMALISED POSTERIOR is enough to determine the relative sampling frequency

Dependent sampling

- Random Walk Metropolis



- Resulting series of parameter space points: “Markov chain”
- Effective sample size
  - Dependence of the sampler affects its ability of approximate the posterior
  - Effective sample size = independent sample size that gives the same error rate as the dependent sample size

Tl;dr – MCMC sampling

incorporate information from an experiment into our theory:
BAYES THEOREM!

calculate evidence in

multiple dimensions

integrate via independent sampling:
MONTE CARLO

How do we get independent samples?

sufficient, that points are sampled with the appropriate relative frequency

But we can’t evaluate

all relative frequencies

dependent sampling:
MARKOV CHAIN

MCMC in pseudo-code

1. Declare initial position θ_i
2. Calculate unnomalised posterior at θ_i: $\mathrm{post}(\theta_i)$
3. For N iterations do:
   1. draw new position θ_{i+1} from proposal distribution
   2. calculate unnomalised posterior at θ_{i+1}: $\mathrm{post}(\theta_{i+1})$
   3. draw random number u between 0 and 1
   4. if $\mathrm{post}(\theta_{i+1})/\mathrm{post}(\theta_i) > u$: move to θ_{i+1}, else stay at θ_i

Sample algorithms

- Random Walk Metropolis: can only be used to sample from unconstrained parameter space

- Metropolis-Hastings: ensures that the MC never strays outside of the bounds of the parameter space
  - Gibbs sampling: simplification for multidimensional PDFs, if marginal distribution of one (or multiple) parameters is known
  - Hamilton MC: uses Hamiltonian dynamics evolution to propose a new point
    reduces correlation between successive points
    No-U-Turn sampler (NUTS)
  - …

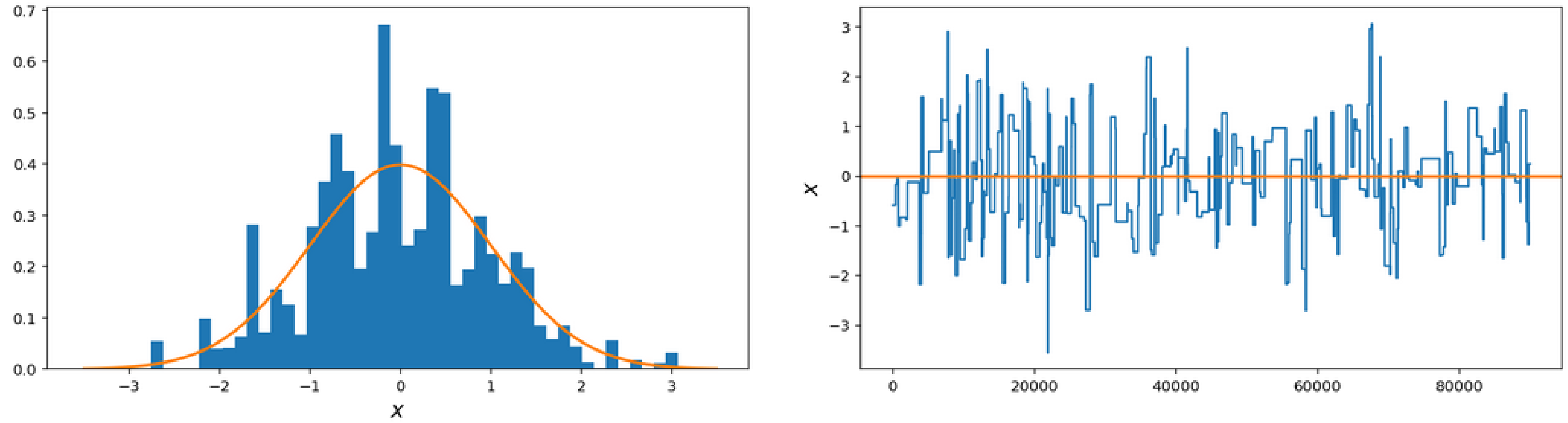
Jump proposals

- Proposal distribution = pdf that decides where to go next
  - Gaussian: mean = current position, variance = “jump size”, yours to choose

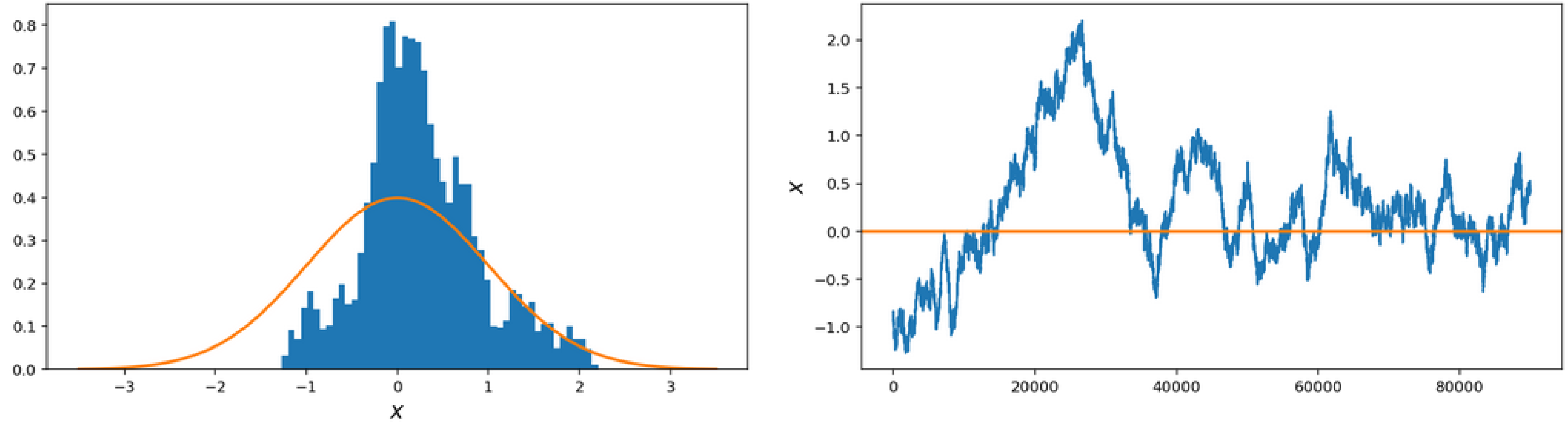
Jump proposals (next level)

- Adaptive Metropolis (AM)
  Haario et al. 2001
  update Gaussian proposal distribution based on previous samples
  can be slow in large parameter spaces
- Single-Component Adaptive Metropolis (SCAM)
  Haario et al. 2005
  only one correlated variable is updated in a proposal
  greatly improves mixing when running with many parameters
- Differential Evolution (DE)
  Braak 2006
  move by difference of two previous, random points in the chain
  used if strong multimodal structures expected in the posterior
- Uncorrelated Jumps
  typically draws from the prior distribution

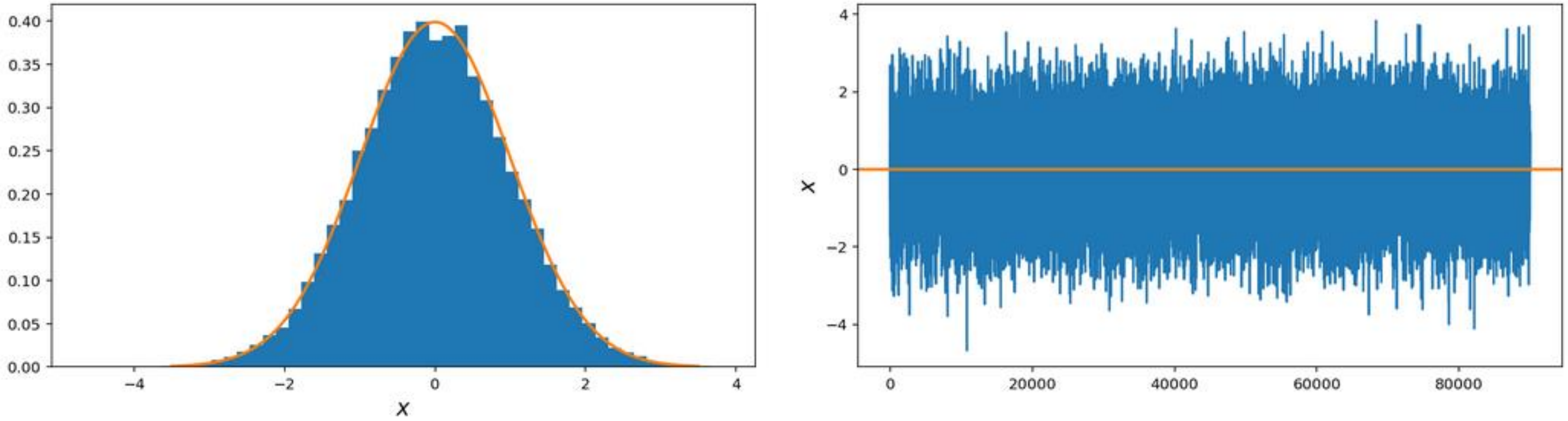
too large jump size



too small jump size

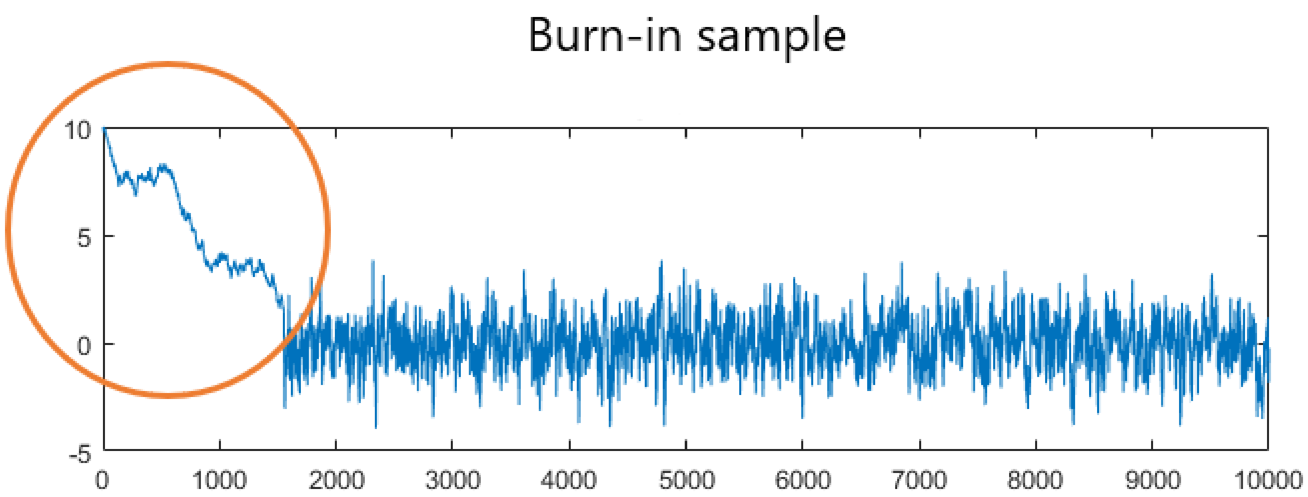


good jump size



Assess your chain

- Convergence
  - visual assessment of the chain
  - Gelman-Rubin-R statistic
- Burn-in

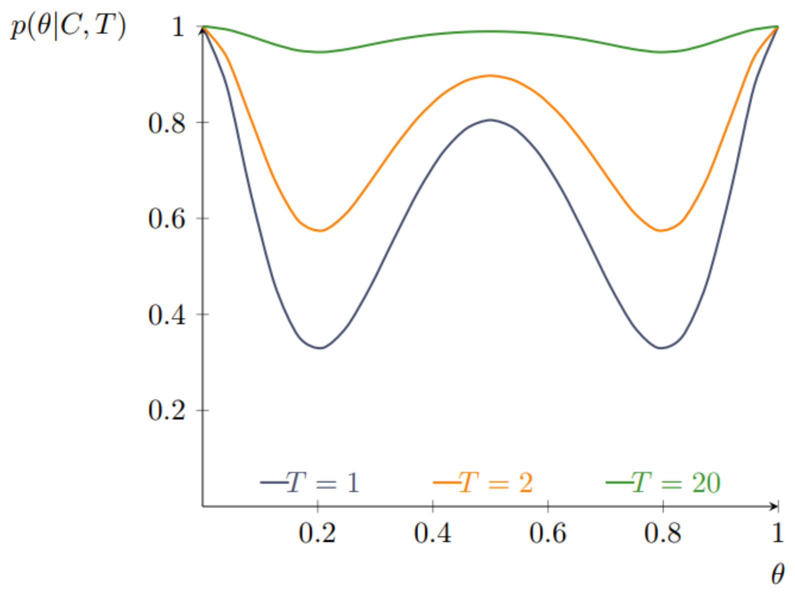


- Autocorrelation length
  - Sample $\mathrm{n}+i$ is uncorrelated from the sample n
  - Determine e.g. using the python package "acor"
  - Thin the MCMC chain by the autocorrelation length to increase the effective sample size

Beyond the simple MCMC

- Parallel tempering
  - Problem: posterior distribution has deep local minima, but at large distances
  - Solution:
    - run multiple chains, flatten out the topology in each but one
    - Flattening via $\exp\left(\frac{1}{T}E(\theta)\right)$,

      where $E(\theta)$ is the negative unnormalized log-posterior at position θ
    - Allow exchange between higher temperature chains and the lowest temperature



https://dictionary.helmholtz-uq.de/content/tempering.html

- Nested sampling
  - used for model comparison
  - allows for the evaluation of a Bayes factor via MCMC sampling

Take-away

- Literature:
  - https://jellis18.github.io/post/2018-01-02-mcmc-part1/
  - https://twiecki.io/blog/2015/11/10/mcmc-sampling/
  - "A student's guide to Bayesian statistics" (Ben Lambert)

- Popular python-based MCMC samplers
  - Emcee
  - PyMC
  - Sampyl
  - PTMCMC

  - PolyChord
  - dynesty

- Python analysis/plotting tools
  - acor
  - seaborn
  - corner
  - ChainConsumer