# Time Series Analysis

## 6. Markov Chain Monte Carlo

Andrew Lesniewski

Baruch College
New York

Spring 2017

# Outline

1. Bayesian inference versus classical inference
2. Metropolis-Hastings algorithm

# Bayesian viewpoint on inference

- Let $x$ denote a single observation of a random variable (data point); this may be a number, a vector, or a more complex object.
- By $\theta$, we denote the parameters of the probability distribution $x$. As usual, we use the notation
$$x \sim p(x|\theta), \tag{1}$$
to indicate that $x$ is distributed according to the law $p(x|\theta)$.
- The objective of statistical inference is to predict the probability distribution of a new data point $x$, given a set $x_1, \ldots, x_N$ of $N$ observed data points.
- In the previous lectures we took the classical ("frequentist") point of view: in order to predict the PDF $p(x|\theta)$, we plug into $p(x|\theta)$ some optimized value of $\theta$ (for example the MLE estimate $\widehat{\theta}$ of $\theta$).

# Bayesian viewpoint on inference

- This classical paradigm to estimating $\theta$ is explicitly based on the assumption that all probabilities involved have an objective character, and the MLE estimator $\widehat{\theta}$ gives the best fit based on the observations.
- The objectivity assumption is important in scientific inquiry, where the reproducibility of results is a key criterion of their validity.
- Such an objectivity requirement is, however, impractical (or outright impossible) in finance, where market outcomes are unique and irreproducible. Probability distributions are based on the participants' views or are implied from option prices.
- A more general and flexible inference paradigm is *Bayesian inference*, which naturally integrates the concept of *subjective probability*.

# Bayesian viewpoint on inference

- Bayesian inference is based on Bayes' definition of conditional probability,

$$\mathrm{P}(E|H) = \frac{\mathrm{P}(E \cap H)}{\mathrm{P}(H)} \,. \tag{2}$$

- One can think about $H$ as a "hypothesis" and about $E$ as "evidence".
- Interchanging the roles of $E$ and $H$, and equating the two expressions for $\mathrm{P}(E \cap H)$, we obtain the following *Bayes' rule*:

$$\mathrm{P}(H|E) = \frac{\mathrm{P}(E|H)\mathrm{P}(H)}{\mathrm{P}(E)} \,. \tag{3}$$

# Bayesian viewpoint on inference

- Relation (3) is phrased as

$$\overset{\text{p(H I E)}}{\textit{posterior probability}} = \frac{\overset{\text{p(E I H)}}{\textit{liklihood}} \times \overset{\text{p(H)}}{\textit{prior probability}}}{\textit{evidence} \quad \text{p(E)}} . \tag{4}$$

  Since the denominator is a constant independent o $H$, this is often stated as a proportionality

$$\textit{posterior probability} \propto \textit{liklihood} \times \textit{prior probability} . \tag{5}$$

  prior: something you believe

- The prior probability P($H$) is a (subjective) probability that the hypothesis holds, while the likelihood P($E|H$) is the probability of $E$ occurring given $H$.
- The *prior distribution* is the (subjective ) probability distribution of the parameters, representing one's beliefs about them, before any evidence is observed. I is denoted by $p(\theta)$.

# Bayesian viewpoint on inference

- The likelihood function is the distribution of the observed data conditioned on its parameters, i.e. $p(x|\theta)$. We use the notation $\mathcal{L}(\theta|x) = p(x|\theta)$ to emphasize that it is a function of the parameters.
- The *marginal likelihood* (a.k.a. the *evidence*) is the distribution of the observed data, i.e. the marginal PDF over the parameters $\theta$,

$$p(x) = \int p(x|\theta)p(\theta)d\theta. \tag{6}$$

- The *posterior distribution* is the distribution of the parameters after taking into account the evidence. From Bayes' rule (3),

$$\begin{aligned} p(\theta|x) &= \frac{p(x|\theta)p(\theta)}{p(x)} \\ &\propto p(x|\theta)p(\theta). \end{aligned} \tag{7}$$

- This identity is the essence of Bayesian inference.

# Bayesian viewpoint on inference

- The *posterior predictive distribution* is the PDF of a new data point $x'$ given the posterior:

$$p(x'|x) = \int p(x'|\theta)p(\theta|x)d\theta. \tag{8}$$

p(x I theta)?

- According to the strict Bayesian paradigm one has to use the posterior predictive distribution to in order to predict the distribution of a new, unobserved data point.
- In other words, an entire probability distribution of the parameters $\theta$ should be found rather than a single value of $\theta$.
- As we have discussed earlier in these lectures, prediction in classical statistics usually involves finding an optimum point estimate of the parameters, namely the MLE estimator:

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} \mathcal{L}(\theta|x). \tag{9}$$

# Bayesian viewpoint on inference

- A half way approach is the *maximum a posteriori estimation* (MAP), in which the posterior distribution is maximized. Namely,

$$\begin{aligned}\widehat{\theta}_{MAP} &= \arg\max_{\theta} p(\theta|x) \\ &= \arg\max_{\theta} \frac{p(x|\theta)p(\theta)}{p(x)} \\ &= \arg\max_{\theta} p(x|\theta)p(\theta).\end{aligned} \tag{10}$$

- As an example, consider a linear regression model

$$y = \alpha + \beta x, \tag{11}$$

given the observations $(x_1, y_1), \ldots, (x_N, y_N)$. Assuming normality, the likelihood function of this data set is

theta contains alpha and beta, but prior only contains beta

$$p(x, y|\theta) = (2\pi\sigma^2)^{-N/2} \prod_{j=1}^{N} \exp\Big(-\frac{(y_j - \alpha - \beta x_j)^2}{2\sigma^2}\Big).$$

# Bayesian viewpoint on inference

- Maximizing this likelihood function leads to the standard estimators for the parameters $\alpha$ and $\beta$.
- On the other hand, we might want to express the view that the parameter $\beta$ should be close to zero. This can be done by selecting a Gaussian prior

$$p(\beta) \propto \exp\Big(-\frac{\lambda\beta^2}{2}\Big),$$

where $\lambda > 0$ is a hyperparameter.

- MAP estimation leads then to the problem of minimizing the following posterior function:

$$-\log p(\theta|x,y) = \frac{\cancel{N}}{2}\log\sigma^2 + \frac{1}{2\sigma^2}\sum_{j=1}^{N}(y_j - \alpha - \beta x_j)^2 + \frac{\lambda\beta^2}{2}.$$

This results in the *ridge regression*, where the parameter $\beta$ is shrunk towards zero.

# Bayesian viewpoint on inference

- Likewise, selecting the prior
$$p(\beta) \propto e^{-\lambda|\beta|},$$
with $\lambda > 0$, leads to the *lasso regression*.
- The only difference between the MLE and MAP estimates is the presence of the prior probability: the MAP estimate allows us to inject our prior beliefs regarding the values of $\theta$ into the calculation .
- The classical approach has the disadvantage that it does not account for any uncertainty in the values of the parameters, and thus tends to underestimate the variance of the predictive distribution.

MLE: 2nd derivative measure the uncertainty of MLE

# Conjugate priors

Beta distribution: alpha-> alpha+sum{xi}
beta-> beta-sum{xi}

- Full Bayesian inference involves calculating the constant factor (6) in the denominator (the evidence). This is, in general, a difficult task (it involves multi-dimensional integration), and it makes full Bayesian inference a hard problem.
- One way to address this issue is to, for a given likelihood function, express our prior beliefs in a way that allow us to carry out the integration in (6). This leads to the notion of a *conjugate prior*.
- If the posterior ~~distribution~~ likelihood function $p(\theta|x)$ is in the same family of ~~distributions~~ likelihood function as the prior ~~PDF~~ likelihood function $p(\theta)$, the prior and posterior are called conjugate distributions. In such a situation, the prior is called a conjugate prior for the likelihood function.

# Conjugate priors

- As an example, assume that the liklihood function is normal with known standard deviation $\sigma$, i.e. the only model parameter is the mean $\mu$.
- Taking the prior to be normal with (hyper)parameters $\mu_0, \sigma_0$, we find that the posterior is normal with mean

$$\mu' = \Big(\frac{\mu_0}{\sigma_0^2} + \frac{1}{\sigma^2}\sum_{j=1}^{N} x_j\Big) \Big/ \Big(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\Big), \tag{12}$$

and standard deviation

$$\sigma' = 1 \Big/ \sqrt{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}}\,. \tag{13}$$

if mu0->0, sigma0-> infinity, then sigma'->sigma0

# Conjugate priors

- Conjugate priors are known explicitly only in a limited number of cases. Even then, the calculations involved may be quite formidable.
- A practical approach to Bayesian computations is through Monte Carlo metods, and in particular through Markov chain Monte Carlo methods.

# Markov chains

- Consider a discrete time process $\theta_t$, $t = 0, 1, \ldots$, taking values in a state space $A$. $A$ can be continuous (such as $\mathbb{R}$ or $\mathbb{R}^n$) or discrete (such as $a^1, \ldots, a^n$).
- The process $\theta_t$ is called a *Markov chain*, if

$$\mathrm{P}(\theta_{t+1}|\theta_{1:t}) = \mathrm{P}(\theta_{t+1}|\theta_t), \tag{14}$$

  or, in other words, if $\theta_{t+1}$ is independent of $\theta_s$ with $s < t$.
- We will be considering Markov chains with *stationary transition probabilities*, i.e. such that $\mathrm{P}(\theta_{t+1}|\theta_t)$ is independent of $t$. This is, in particular, true if the process is strictly stationary.
- If the state space is finite, the transition probabilities can be encoded in the matrix $P \in \mathrm{Mat}_n(\mathbb{R})$, defined by

$$p_{ij} = \mathrm{P}(\theta_{t+1} = \overset{\text{j}}{a^j}|\theta_t = \overset{\text{i}}{a^i}). \tag{15}$$

From state i to state j

# Markov chains

- Notice that

$$\sum_{j=1}^{n} p_{ij} = 1. \tag{16}$$

  Matrices with non-negative elements satisfying the above identity are called *stochastic matrices*.

- The initial probability distribution $\pi_0$ is a vector defined as

$$\pi_{0,j} = \mathrm{P}(\theta_0 = a^j). \tag{17}$$

- The *marginal probability distribution* $\pi_t$ at time $t = 1, 2, \ldots$ has components

$$\pi_{t,i} = \sum_{j=1}^{n} p_{ji} \pi_{t-1,j},$$

  i.e.

$$\begin{aligned} \pi_t &= \pi_{t-1} P \\ &= \pi_0 P^t. \end{aligned} \tag{18}$$

# Equilibrium and detailed balance

- The marginal probability distribution $\pi$ is called an *equilibrium distribution*, if $\pi_t = \pi_{t-1}$, i.e.
$$\pi = \pi P. \tag{19}$$
- The notion of an equilibrium distribution is closely related to the *detailed balance condition*, namely
$$\pi_i p_{ij} = \pi_j p_{ji}. \tag{20}$$
A Markov chain for which there is a probability distribution $\pi$ satisfying the detailed balance condition is called *reversible*.
- This terminology is justified by the observation that (20) states that the absolute probabilities of moving from state $i$ to $j$ and from $j$ to $i$ are identical. In other words, the Markov chain remains invariant under reversing the time direction.
- We claim that *if $\pi$ satisfies the detailed balance condition, then it is an equilibrium distribution.*

# Equilibrium and detailed balance

- The proof is a straightforward calculation:

$$
\begin{aligned}
(\pi P)_j &= \sum_{i=1}^{n} \pi_i p_{ij} \\
&= \sum_{i=1}^{n} \pi_j p_{ji} \\
&= \pi_j \sum_{i=1}^{n} p_{ji} \\
&= \pi_j,
\end{aligned}
$$

  i.e. $\pi P = \pi$. QED
- It is, however, in general not true that an equilibrium distribution satisfies the detailed balance condition.
- The concepts above can be generalized to the case of a continuous state space by replacing the sums with suitable integrals.

# Convergence to equilibrium

- An important practical question is whether a given Markov chain has an equilibrium distribution and whether it is unique.
- We assume the the Markov chain is irreducible, i.e. it is possible to reach from one state to any other state with non-zero probability.
- Consider the limit

$$P_\infty = \lim_{t\to\infty} P^t. \tag{21}$$

limP^{t-1} P

Then $P_\infty P = P_\infty$, and all rows of $P_\infty$ are equal. We will denote this common value by $\pi$.

All rows converge, and sum of each row equal to 1

- In other words, $\pi$ represents an equilibrium distribution of the Markov chain. When is it unique?

# Convergence to equilibrium

- A state $\theta_j$ is called *ergodic*, if it is
  - (i) recurrent, i.e. the system returns to this state with probability 1, and the expected time of the return is finite; non-zero
  - (ii) aperiodic, i.e. the recurrence is non-periodic. e.g., not every 10mins

  If all states in an irreducible Markov chain are ergodic, then the chain is called ergodic.
- **Theorem 1**. *For an ergodic Markov chain, there is a unique probability vector $\pi$ such that $\pi P = \pi$.*
- **Theorem 2**. *For an ergodic Markov chain,*

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T} P^t = P_\infty. \tag{22}$$

  In other words, in an ergodic Markov chain, the time average of a state is equal to its average under the equilibrium probability distribution.
- The rate of convergence to the equilibrium distribution depends on the details of the transition matrix.

# Markov chain Monte Carlo

- *Markov chain Monte Carlo* (MCMC) methods are methodologies for generating samples from a probability distribution $p(\theta)$ using the techniques of Markov chains.
- The underlying idea is to construct a Markov chain that has the desired distribution $p(\theta)$ as its equilibrium distribution. The states of the chain are then used as samples from the distribution $p(\theta)$.
- In applications, the distribution $p(\theta)$ may not be given by an explicit formula, or may be otherwise hard to sample from.
- MCMC methods are used for calculating numerical approximations of multi-dimensional integrals, which arise in Bayesian statistics. This is accomplished by generating random samples from the prior or posterior distributions.
- Such applications may require integrations over hundreds or thousands of unknown parameters.

# Classic Monte Carlo method

- MCMC methods should be contrasted with the classic Monte Carlo methods used in many financial applications (such as valuation models).
- Classic Monte Carlo utilizes non-Markov chain simulated random variables.
- The basic input into classic Monte Carlo methods is an efficient algorithm for generating samples from the uniform distribution $U(0,1)$ on the interval $(0,1)$. These samples can then be transformed into samples from the standard normal distribution, as well as a number of other parametric distributions.
- Typically, the simulated random variables are statistically independent; any desired dependence is artificially introduced using e.g. the Cholesky decomposition of the covariance matrix.
- Simulation from any given distribution $p(\theta)$ may pose challenges as no simple transformation onto the uniform distribution is possible.

# Acceptance-rejection method

- A general technique for simulation of independent random variables from a given PDF $p(\theta)$ is the *acceptance-rejection method*.
- The PDF $p(\theta)$ is usually given in the form:

$$p(\theta) = \frac{1}{Z} f(\theta), \tag{23}$$

where $f(\theta)$ is an "unnormalized" density function, and $Z$ is the normalization constant,

$$Z = \int f(\theta) d\theta, \tag{24}$$

whose value may not be explicitly known.

- Let $q(\theta)$ be another PDF, called the *proposal distribution*, with the properties that: uniform, gaussian
  (i) there is an effective method for simulating from $q(\theta)$,
  (ii) there is a constant $M$ (explicitly known) such that for all $\theta$,

$$f(\theta) \leq Mq(\theta). \quad \text{Bounded} \tag{25}$$

# Acceptance-rejection method

- The acceptance-rejection method is a two step process. In order to generate a random number from $p(\theta)$,
  (i) generate a candidate $\eta \sim q(\theta)$, and $u \sim U(0,1)$ (the uniform distribution on $(0,1)$,
  (ii) if
  $$u \leq \frac{f(\eta)}{Mq(\eta)}, \tag{26}$$
  then accept $\eta$, else reject it and go back to (i).
- Drawing a graph, we see that the algorithm indeed generates samples from $p(\theta)$.
- It is helpful to chose the constant $M$ as close to the optimal as possible; large $M$ slows down the algorithm.
- Even with the optimal choice of $M$, the algorithm may lead to a large number of rejections and poor performance. This is especially true in the case the proposal PDF significantly differs from the target PDF.

# Markov chain Monte Carlo

- There are various approaches to MCMC. Two most popular approaches are:
  (i) Metropolis-Hastings algorithm,
  (ii) Gibbs sampler.
- These algorithms are clever extensions of the A-R algorithm by allowing the proposal distribution to depend on the state of a Markov chain.
- The Markov chain is selected so that the distribution $p(\theta)$ is its equilibrium distribution.

Gibbs sampler ref:
https://darrenjw.wordpress.com/2011/07/16/gibbs-sampler-in-various-languages-revisited/

# Formulation of the algorithm

- The objective of the algorithm is to generate samples from a target distribution $p(\theta)$. We assume that it has unnormalized density $f(\theta)$, i.e. $p(\theta) \sim f(\theta)$.
- Let $q(\eta|\theta)$ be a *proposal* PDF, with $\int q(\eta|\theta)d\eta = 1$. We interpret the proposal distribution by saying that if the chain is in the state $\theta$, then it generates random variables $\eta$. The proposal PDF is selected so that it is easy to sample from it.
- In the unlikely situation that $q(\theta|\eta)$ satisfies the detailed balance condition,

$$p(\theta)q(\eta|\theta) = p(\eta)q(\theta|\eta), \tag{27}$$

  then a sample $\eta$ from $q(\eta|\theta)$
- In general though, the detailed balance condition will be violated. For example, the process may be more likely to move from $\theta$ to $\eta$, i.e.

$$p(\theta)q(\eta|\theta) > p(\eta)q(\theta|\eta). \tag{28}$$

# Formulation of the algorithm

- This can be "corrected" by introducing the *probability of a move* $\alpha(\theta, \eta)$ from $\theta$ to $\eta$, so that the total transition probability is equal to

$$\widetilde{p}(\eta|\theta) = q(\eta|\theta)\alpha(\theta, \eta). \tag{29}$$

- In contrast to the A-R algorithm, if the move is not made, the system returns to the original state, and returns $\theta$ as the sample from $p(\theta)$.
- Let us now determine $\alpha(\theta, \eta)$. We will do it by requiring that $\widetilde{p}(\eta|\theta)$ satisfied the detailed balance condition:

$$p(\theta)\widetilde{p}(\eta|\theta) = p(\eta)\widetilde{p}(\theta|\eta),$$

or

$$\begin{aligned} p(\theta)q(\eta|\theta)\alpha(\theta, \eta) &= p(\eta)q(\theta|\eta)\alpha(\eta, \theta) \\ &= p(\eta)q(\theta|\eta). \end{aligned} \tag{30}$$

The last equality holds, because, under the scenario (28), it is natural to require that $\alpha(\eta, \theta) = 1$.

# Formulation of the algorithm

- This leads to the following value for $\alpha(\theta, \eta)$:

$$\alpha(\theta, \eta) = \frac{p(\eta)q(\theta|\eta)}{p(\theta)q(\eta|\theta)} .$$

- Notice that, importantly, the formula above does not depend on the normalizing factor in $p(\theta)$ and we can replace $p(\theta)$ by $f(\theta)$.
- The same analysis can be done if the sense of the inequality in (28) is reversed, leading to the definition of $\alpha(\eta, \theta)$.
- We summarize these calculations in the following manner. Let us fform the *Hastings ratio*:

$$r(\theta, \eta) = \frac{f(\eta)q(\theta|\eta)}{f(\theta)q(\eta|\theta)} , \tag{31}$$

and define

$$\alpha(\theta, \eta) = \min(r(\theta, \eta), 1). \tag{32}$$

# Formulation of the algorithm

- We are ready now ready to formulate the *Metropolis-Hastings (M-H) algorithm*.
- Choose an initial value $\theta_0$, and for $j = 1, 2, \ldots$, and proceeds as follows:
  (i) Generate a candidate $\eta \sim q(\eta|\theta_j)$ and $u \sim U(0, 1)$.
  (ii) Accept $\eta$ with probability $\alpha(\theta_j, \eta)$, i.e. accept if

$$u \leq \alpha(\theta_j, \eta), \tag{33}$$

  and reject otherwise.
  (iii) If accepted, set $\theta_{j+1} = \eta$, else, $\theta_{j+1} = \theta_j$.
- The simulated values are $\theta_1, \theta_2, \ldots$.
- An important special case is when the proposal distribution is symmetric, $q(\theta|\eta) = q(\eta|\theta)$, in which case

$$r(\theta, \eta) = \frac{f(\eta)}{f(\theta)} \,. \tag{34}$$

This case, historically preceding the general M-H algorithm is referred to as the *Metropolis algorithm*.

# Burn-in period

- The simulated values can be considered as drawn from the PDF $p(x)$ only after an initial transient period.
- The length of the burn-in period depends on the rate at which the probability distribution of the Markov chain converges to its stationary limit.
- The random variables generated by the M-H algorithm are not necessarily independent of each other. For many purposes, such as evaluztion of integrals this is not a major concern. Other methodologies, such as *Hamiltonian MCMC* address this issue.

# Choice of the proposal distribution

- A suitable choice of the proposal PDF is a key part of the implementation of the M-H algorithm. Typically, the proposal distribution is specified as a parameterized family of distributions, with tunable parameters.
- A possible choice of $q(\eta|\theta)$ is a suitable multivariate distribution $\varphi(\eta - \theta)$. In this case, the candidate is given by $\eta = \theta + \zeta$, where $\zeta \sim \varphi$. This variant of the algorithm is called the *random walk M-H*, because the candidate is chosen as the current state plus noise.
- If $\varphi$ is symmetric around the origin, $\varphi(-\zeta) = \varphi(\zeta)$, then $q(\eta|\theta)$ is symmetric, and so the M-H algorithm reduces to the Metropolis algorithm.
- Popular choices of $\varphi$ are:
  (i) multivariate normal distribution,
  (ii) multivariate $t$-distribution,
  (iii) multivariate uniform distribution.

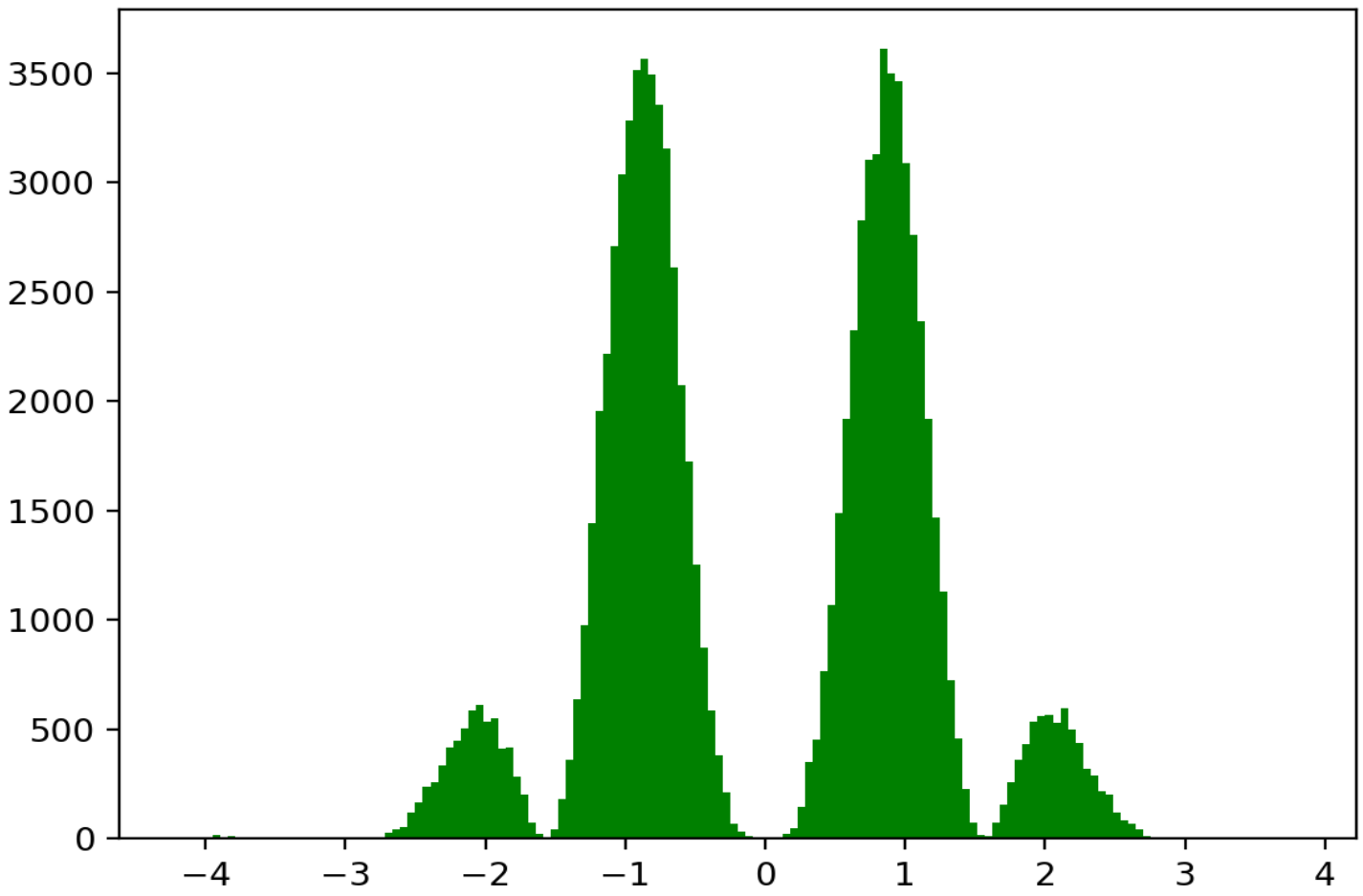
# Choice of the proposal distribution

- Another common choice is $q(\eta|\theta) = \varphi(\eta)$, i.e. the candidates are drawn independently from a fixed distribution, independently of the current state. The resulting proposal distribution is non-symmetric and the full M-H algorithm has to be used.
- Again, the distribution $\varphi$ can be specified as e.g. a (multivariate) normal distribution, $t$-distribution, gamma distribution, or uniform distribution.
- This version of the algorithm is known as the *independent M-H* algorithm.

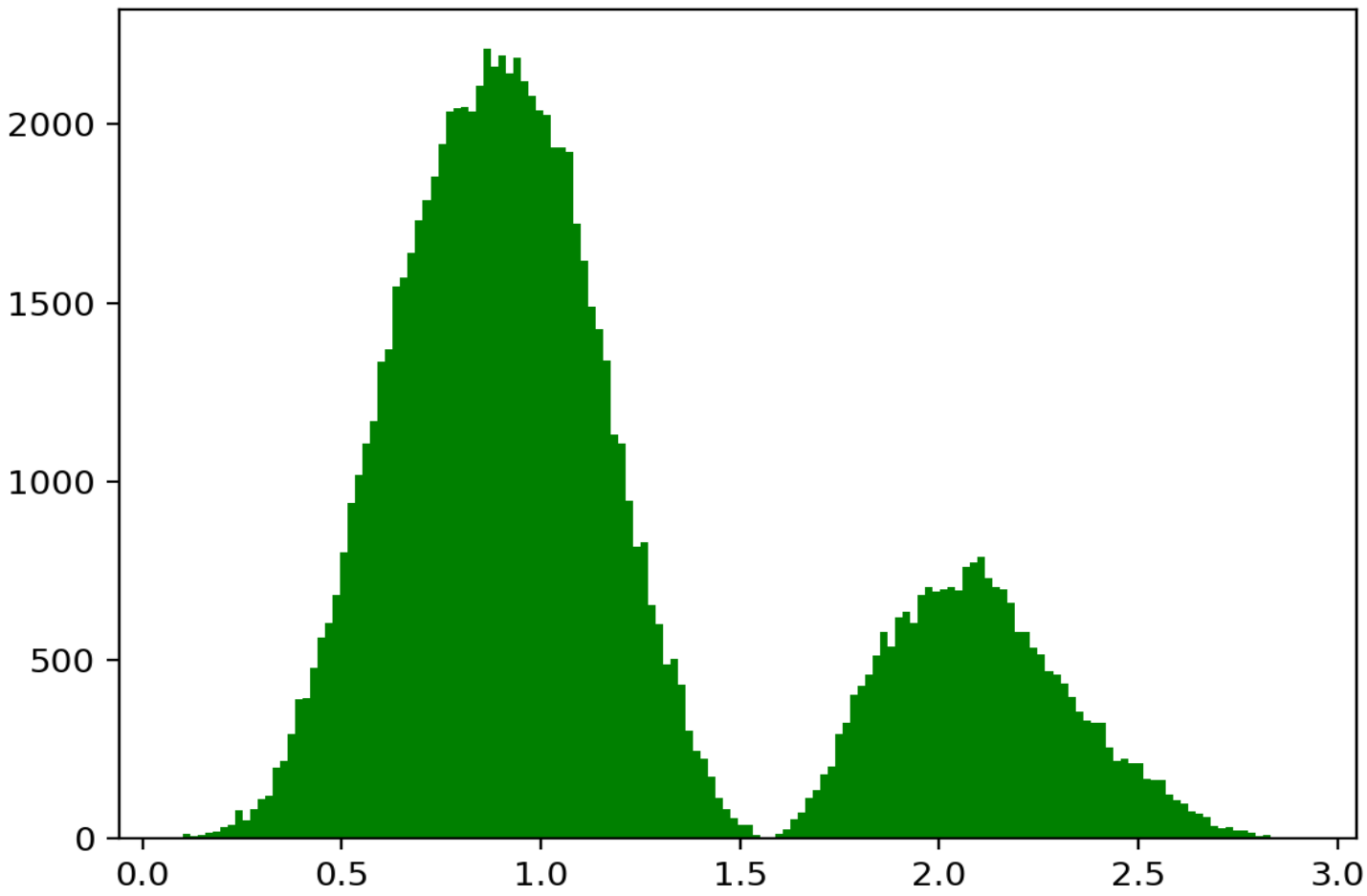
# Example: perturbed Gaussian distribution

- To illustrate the method, we consider the following amusing example taken from [2]. Let
$$f(\theta) = \sin^2(\theta)\sin^2(2\theta)\exp(-\theta^2/2) \tag{35}$$
be the (unnormalized) standard normal distribution perturbed by the factor $sin^2(\theta)\sin^2(2\theta)$.
- We choose the proposal to be the uniforrm PDF,
$$q(\eta|\theta) = \frac{1}{2a}\,1_{(\theta-a,\theta+a)}(\eta), \tag{36}$$
where $a > 0$ is a tunable parameter.
- The following graphs show the results of the simulations with two choices of the parameter $a$: $a = 1$ and $a = 0.1$, respectively. In both cases, $\theta_0 = 3.14$.
- Notice how the choice of $a$ affects the simulation.

# Example: perturbed Gaussian distribution



# Example: perturbed Gaussian distribution



# Example: perturbed Gaussian distribution

- Here is the code snippet used to generate these graphs in Python:

```
import numpy as np
import matplotlib.pyplot as plt

def target_density(x):
    fac = np.sin(x) * np.sin(2*x)
    return fac**2 * np.exp(-0.5*x**2)

def metropolis(theta, a):
    eta = np.random.uniform(theta - a, theta + a)
    u = np.random.uniform()
    if u > target_density(eta) / target_density(theta):
        eta = theta
    return eta

N = 100000
theta = np.zeros(N)
theta[0] = 3.14
for i in range(1, N):
    theta[i] = metropolis(theta[i-1], 1.0)

plt.hist(theta[1000:N], 150, color = 'g')
plt.show()
```

MCMC needs to play with the parameters

# References

*Handbook of Markov Chain Monte Carlo*, ed. by S. Brooks, G. L. Jones, and X.-L. Meng, CRC Press (2011).

Robert, C. P.: The Metropolis-Hastings algorithm, arXiv:1504.01896v3 [stat.CO] 27 Jan 2016.