Foundations of Data Science

Lecture 6: Markov chain Monte Carlo

Prof. Gilles Louppe
g.louppe@uliege.be

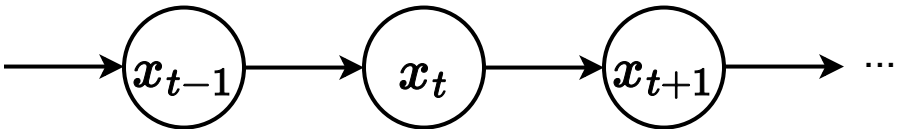
LIÈGE université

How to sample from a distribution $p(x)$ for which we can only evaluate an unnormalized version $\tilde{p}(x)$ of its density?

Use cases:

- Generating data points x from a generative model $p(x)$.
- Computing expectations $\mathbb{E}_{p(x)}[f(x)]$ or integrals $\int f(x)p(x)dx$.
- Computing posterior distributions $p(z \mid x)$ in Bayesian inference.

Markov chain Monte Carlo



Markov chains

A Markov chain is a sequence of random variables $X_1, X_2, \ldots$ that assumes the conditional independence of X_{t+1} and $X_{1:t-1}$ given X_t.

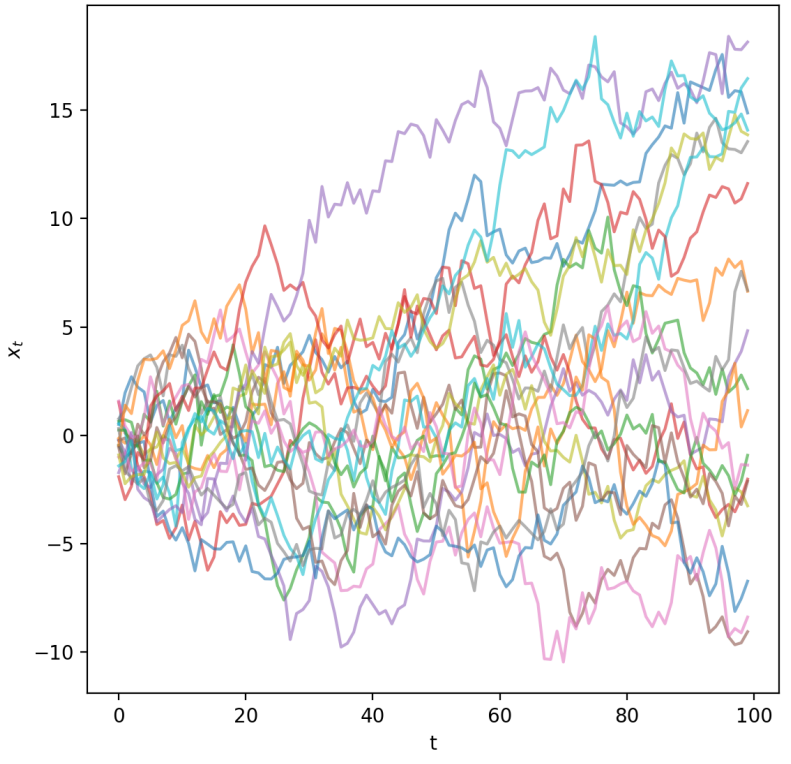
It follows that the joint distribution of $X_1, \ldots, X_T$ can be factorized as

$$p(x_1, \ldots, x_T) = p(x_1) \prod_{t=1}^{T-1} p(x_{t+1} \mid x_t),$$

where $p(x_1)$ is the **initial distribution** and $p(x_{t+1} \mid x_t)$ is the **transition model**.

When the transition model does not depend on t, i.e., $p(x_{t+1} \mid x_t) = p(x' \mid x)$ for all t, the Markov chain is said to be **time-homogeneous**.

Example: Gaussian random walk



Gaussian random walk with $p(x_1) = \mathcal{N}(x_1|0, 1)$ and transition model $p(x_{t+1} \mid x_t) = \mathcal{N}(x_{t+1}|x_t, \sigma^2)$.

Stationary distributions

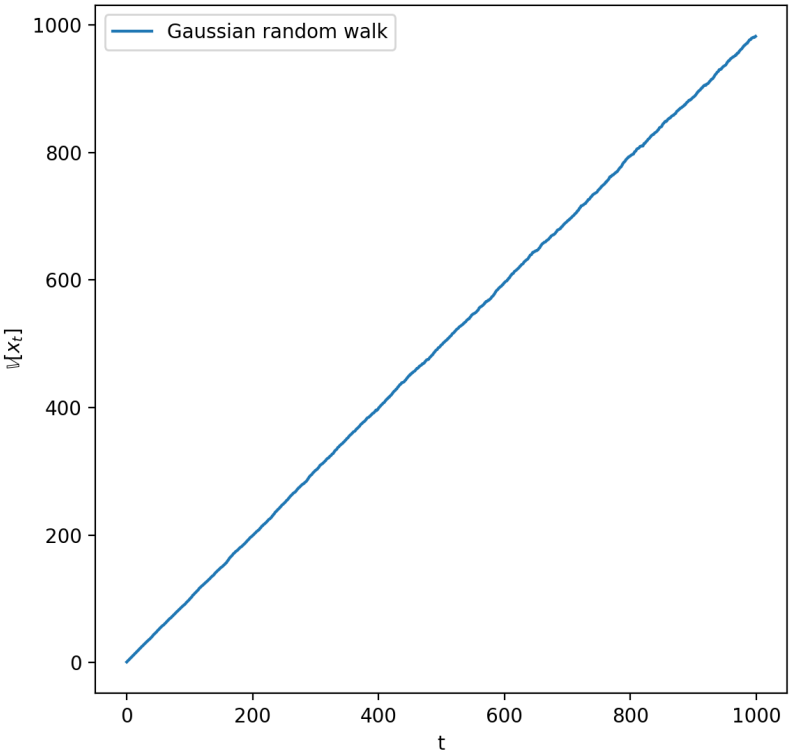
A stationary distribution π is a distribution that remains unchanged when passed through the transition model.

Formally, $\pi(x)$ is stationary for the transition model $p(x' \mid x)$ if

$$\pi(x') = \int \pi(x) p(x' \mid x) dx.$$

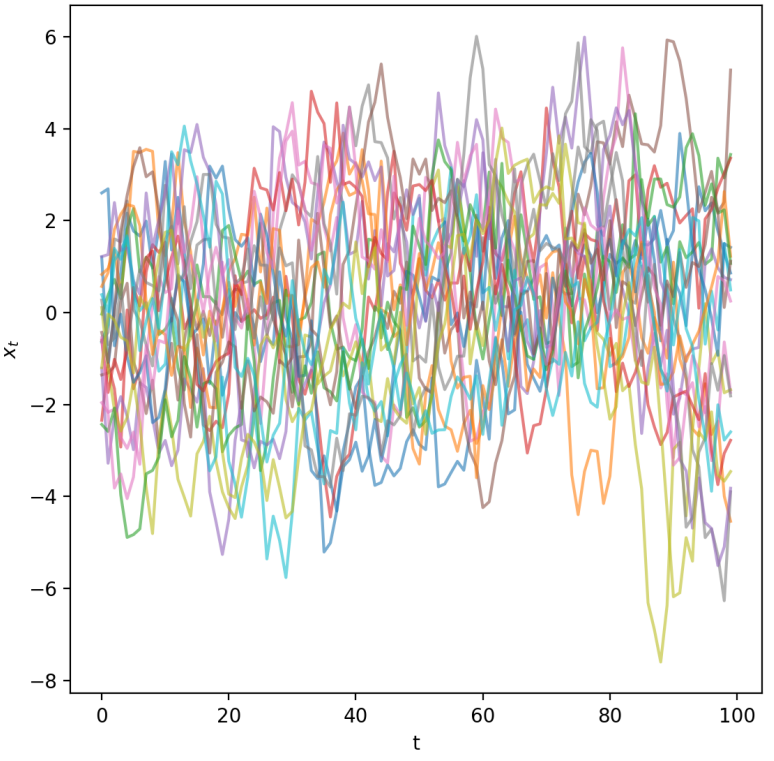
Therefore, if $x_t \sim \pi(x)$, then $x_{t+1} \sim \pi(x)$.

Example: Gaussian random walk (continued)



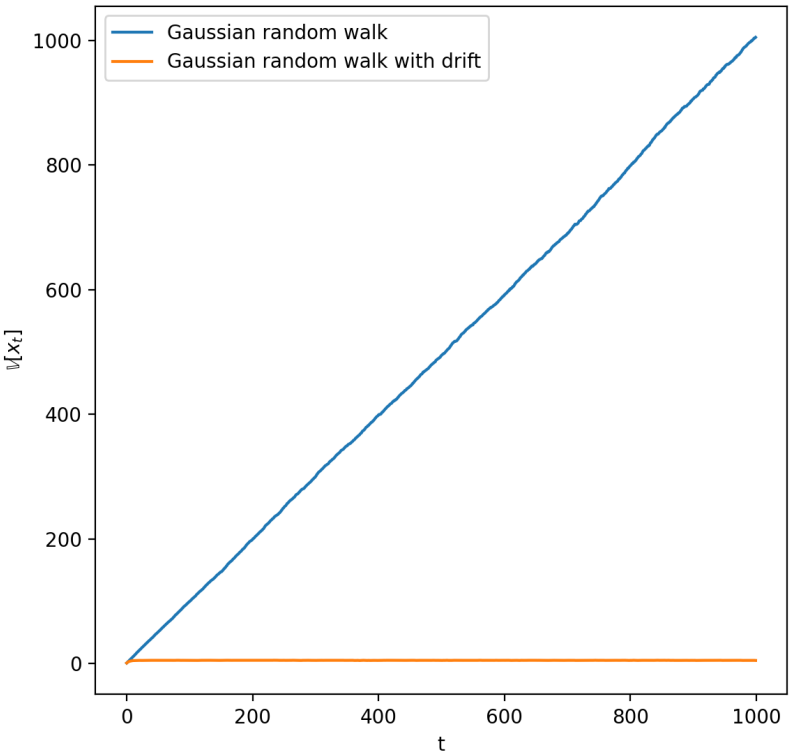
The Gaussian random walk does not have a stationary distribution since its variance increases indefinitely over time.

Example: Gaussian random walk with drift



Gaussian random walk with $p(x_1) = \mathcal{N}(x_1|0,1)$ and transition model

$$p(x_{t+1} \mid x_t) = \mathcal{N}(x_{t+1}|x_t - \kappa(x_t - \mu), \sigma^2).$$



The Gaussian random walk with drift has a stationary distribution

$$\pi(x) = \mathcal{N}(x|\mu, \frac{\sigma^2}{2\kappa - \kappa^2}).$$

Regularity conditions

A markov chain is said to be:

- **irreducible** if any state x' can be reached from any state x, i.e., for any x, x', there exists t such that $p(x_t = x' \mid x_1 = x) > 0$.
- **aperiodic** if there is no deterministic cycle, i.e., for all x, the greatest common divisor of the set $\{t : p(x_t = x \mid x_1 = x) > 0\}$ is 1.
- for discrete state spaces, **positive recurrent** if the expected return time to any state is finite; for continuous state spaces, **Harris recurrent** if the chain returns to any set A with positive measure infinitely often with probability 1.

Basic limit theorem for Markov chains

Theorem. If a time-homogeneous Markov chain is irreducible, aperiodic, and (positive/Harris) recurrent, then it satisfies the following properties:

- (Existence and uniqueness) There exists a **unique** stationary distribution π;
- (Convergence) For any initial distribution of $p(x_1)$, $p(x_t) \to \pi(x)$ as $t \to \infty$;
- (Ergodic theorem) For any function f with $\mathbb{E}_\pi[|f(x)|] < \infty$,

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} f(x_t) = \mathbb{E}_\pi[f(x)]$$

  almost surely.

Proof. (Omitted and beyond scope.)

This result implies that if a Markov chain satisfies the regularity conditions, then it has a unique stationary distribution and converges to it regardless of the initial distribution.

However, the theorem does not provide a method to construct a Markov chain with a desired stationary distribution, nor does it specify the rate of convergence to the stationary distribution.

Detailed balance

A transition model $p(x' \mid x)$ satisfies detailed balance with respect to a distribution π if

$$\pi(x)p(x' \mid x) = \pi(x')p(x \mid x')$$

for all x, x'.

Proposition. If a transition model $p(x' \mid x)$ satisfies detailed balance with respect to a distribution π, then π is a stationary distribution of $p(x' \mid x)$.

Proof. Integrating both sides of the detailed balance equation over x, we get

$$\begin{aligned}\int \pi(x)p(x' \mid x)dx &= \int \pi(x')p(x \mid x')dx \\ &= \pi(x') \int p(x \mid x')dx \\ &= \pi(x').\end{aligned}$$

The detailed balance condition is a **sufficient** but not necessary condition for stationarity.

However, it is often easier to verify than the stationarity condition, especially when constructing Markov chains with a desired stationary distribution.

Metropolis-Hastings

The Metropolis-Hastings algorithm is a procedure for sampling from a target distribution $\pi(x)$, given only an unnormalized version $\tilde{\pi}(x)$ of its density.

The core idea is to construct a Markov chain whose stationary distribution matches the target distribution $\pi(x)$.

The Metropolis-Hastings algorithm proceeds as follows:

1. Given the current state x_t, propose a new state $x' \sim q(x' \mid x_t)$ using a proposal distribution $q(x' \mid x_t)$.
2. Compute the acceptance ratio

$$\alpha(x' \mid x_t) = \min\left(\frac{\tilde{\pi}(x')q(x_t \mid x')}{\tilde{\pi}(x_t)q(x' \mid x_t)}, 1\right).$$

3. Accept the proposed state x' with probability $\alpha(x' \mid x_t)$. If accepted, set $x_{t+1} = x'$. Otherwise, set $x_{t+1} = x_t$.
4. Repeat steps 1-3 for a large number of iterations.

This process generates a sequence of samples $x_1, x_2, \ldots$ that form a Markov chain.

Proposition. The Metropolis-Hastings algorithm constructs a Markov chain whose stationary distribution is $\pi(x)$.

Proof. The transition model of the Markov chain is given by

$$p(x' \mid x) = \begin{cases} q(x' \mid x)\alpha(x' \mid x) & \text{if } x' \neq x, \\ r(x) & \text{if } x' = x, \end{cases}$$

where $r(x) = 1 - \int q(x' \mid x)\alpha(x' \mid x)dx'$ is the probability of rejecting the proposal and staying at the current state.

To show that $\pi(x)$ is stationary, we verify the detailed balance condition for $x' \neq x$:

$$\begin{aligned}
\tilde{\pi}(x)p(x' \mid x) &= \tilde{\pi}(x)q(x' \mid x)\alpha(x' \mid x) \\
&= \tilde{\pi}(x)q(x' \mid x) \min\left(\frac{\tilde{\pi}(x')q(x \mid x')}{\tilde{\pi}(x)q(x' \mid x)}, 1\right) \\
&= \min(\tilde{\pi}(x)q(x' \mid x), \tilde{\pi}(x')q(x \mid x')) \\
&= \tilde{\pi}(x')q(x \mid x') \min\left(\frac{\tilde{\pi}(x)q(x' \mid x)}{\tilde{\pi}(x')q(x \mid x')}, 1\right) \\
&= \tilde{\pi}(x')p(x \mid x').
\end{aligned}$$

Since $\pi(x) = \frac{\tilde{\pi}(x)}{Z}$ for some normalization constant Z, the detailed balance condition also holds for $\pi(x)$. Hence, $\pi(x)$ is a stationary distribution of the Markov chain.

For $x' = x$, detailed balance holds trivially.

Proposition. The Markov chain constructed by the Metropolis-Hastings algorithm satisfies the regularity conditions under appropriate conditions on q.

Sketch of proof:

- Aperiodicity: The chain can stay at the same state when proposals are rejected, so $p(x|x) > 0$. This breaks any periodic cycles.
- Irreducibility: If the proposal q has appropriate support, then for any x, x' where $\pi(x), \pi(x') > 0$, there exists a sequence of proposals with positive probability connecting them.
- Recurrence: If π is a proper probability distribution, the chain is positive/Harris recurrent. (Proof beyond scope)

These conditions ensure convergence to $\pi(x)$ by the basic limit theorem.

Proposal distributions

The efficiency of the Metropolis-Hastings algorithm heavily depends on the choice of the proposal distribution $q(x' \mid x)$.

Common choices include:

- **Random walk**: $q(x' \mid x) = \mathcal{N}(x'|x, \sigma^2 I)$ for continuous state spaces. Simple and general, but efficiency depends on tuning σ (optimal acceptance rate: ~20-40%).
- **Langevin**: $q(x' \mid x) = \mathcal{N}(x'|x + \frac{\epsilon^2}{2}\nabla \log \tilde{\pi}(x), \epsilon^2 I)$, which uses gradient information to guide proposals toward high-probability regions. Requires differentiable $\tilde{\pi}(x)$.
- **Independence**: $q(x' \mid x) = q(x')$, independent of current state. Can be efficient if $q(x')$ closely approximates $\pi(x)$.

Beyond Metropolis-Hastings

While Metropolis-Hastings is foundational, several extensions improve efficiency:

- **Hamiltonian Monte Carlo (HMC)**: Uses Hamiltonian dynamics to propose new states, enabling larger, informed jumps through parameter space. Requires gradient $\nabla \log \pi(x)$.
- **No-U-Turn Sampler (NUTS)**: An adaptive variant of HMC that automatically tunes trajectory length, eliminating manual parameter tuning. Default sampler in Stan and PyMC.
- **Gibbs sampling**: Updates variables one at a time from conditional distributions. A special case of MH with acceptance probability = 1 when conditionals are tractable.

These advanced methods are widely used in practice and build on MH principles!

Run the first part of `nb06b-mcmc.ipynb` to see Metropolis-Hastings in action.

Bayesian inference with MCMC

Bayesian inference with MCMC

In Bayesian inference, we are often interested in computing the posterior distribution

$$p(z, \theta \mid x) = \frac{p(x \mid z, \theta) p(z, \theta)}{p(x)}$$

of latent variables z and parameters θ given observed data x.

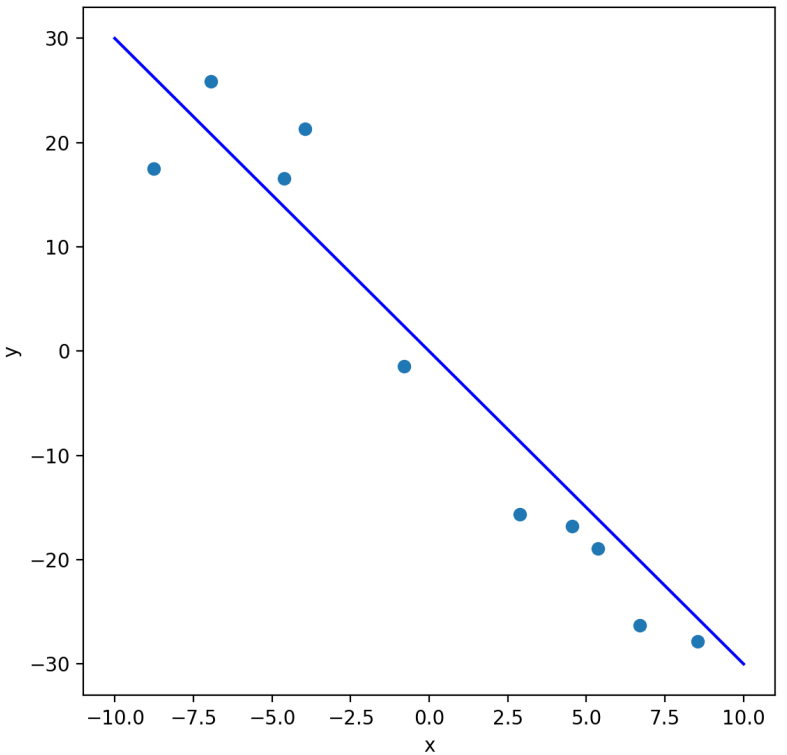
The posterior density is often **intractable** due to the marginal density $p(x) = \int p(x \mid z, \theta) p(z, \theta) dz d\theta$.

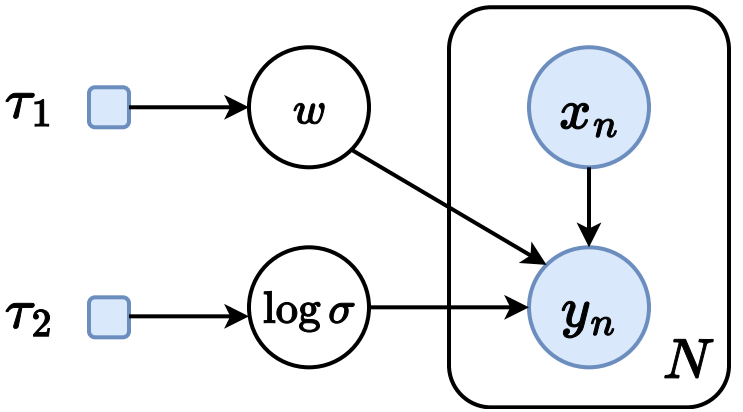
MCMC methods, such as Metropolis-Hastings, can be used to generate samples from the posterior distribution $p(z, \theta \mid x)$ even when the marginal density $p(x)$ is intractable.

Since MCMC methods only require an unnormalized density $\tilde{p}$, **we can use the joint tractable density** $p(x, z, \theta) = p(x \mid z, \theta)p(z, \theta)$ as the target distribution, ignoring the marginalization constant $p(x)$.

Example: Bayesian linear regression

Consider a data set of N observations $x = \{(x_i, y_i)\}_{i=1}^{N}$, where $x_i \in \mathbb{R}^d$ are input features and $y_i \in \mathbb{R}$ are target values.





We model the conditional distribution of y given x using a Bayesian linear regression model

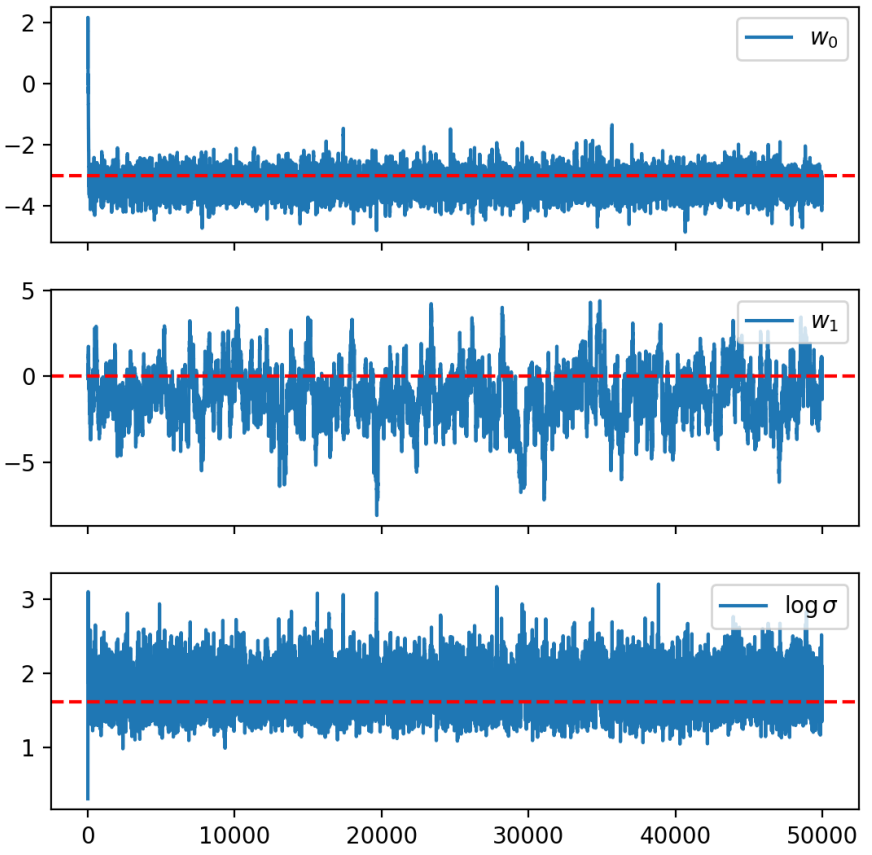
$$p(y \mid x, w, \log \sigma) p(w) p(\log \sigma),$$

where $p(y \mid x, w, \log \sigma) = \mathcal{N}(y \mid w^T x, \sigma^2)$ is the likelihood, $p(w) = \mathcal{N}(w \mid 0, \tau_1^2 I)$ is the prior over weights w, $p(\log \sigma) = \mathcal{N}(\log \sigma \mid 0, \tau_2^2)$ is the prior over the log-noise standard deviation, and τ_1, τ_2 are hyperparameters.

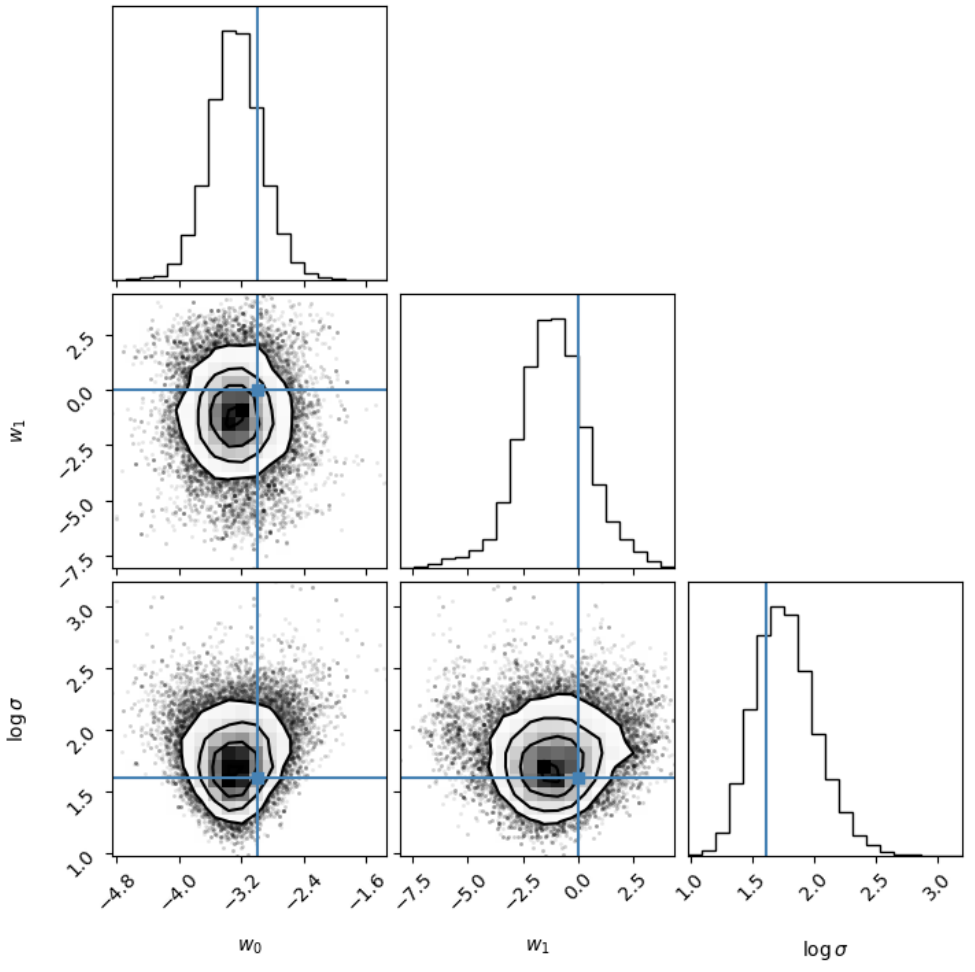
The target distribution is the posterior $p(w, \log \sigma \mid \{x_i, y_i\}_{i=1}^N)$, which is proportional to the joint density

$$
\begin{aligned}
&p(\{y_i\}_{i=1}^N \mid \{x_i\}_{i=1}^N, w, \log \sigma) p(w) p(\log \sigma) \\
&= \left(\prod_{i=1}^N p(y_i \mid x_i, w, \log \sigma) \right) p(w) p(\log \sigma).
\end{aligned}
$$

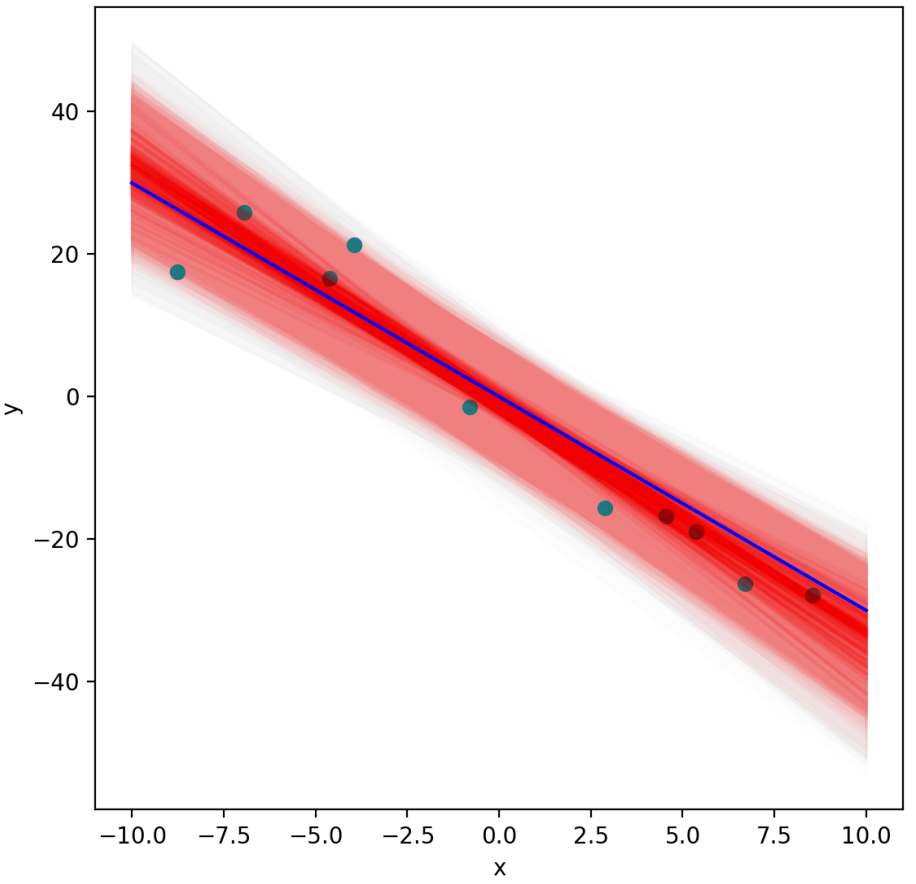
We can use Metropolis-Hastings to sample from this posterior distribution using the joint density as the unnormalized target density.



Markov chain traces for w_0, w_1, and $\log \sigma$ using Metropolis-Hastings.



Posterior samples visualized using a corner plot.



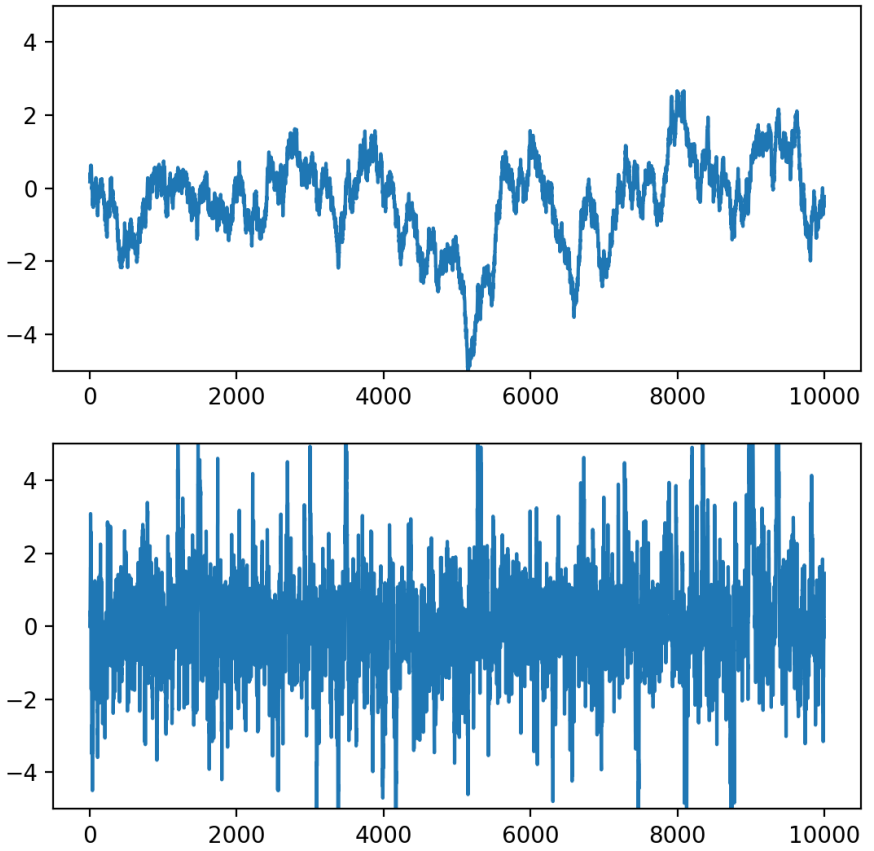
Posterior predictive distribution $p(y \mid x, \{x_i, y_i\}_{i=1}^N)$.

Diagnostics for MCMC

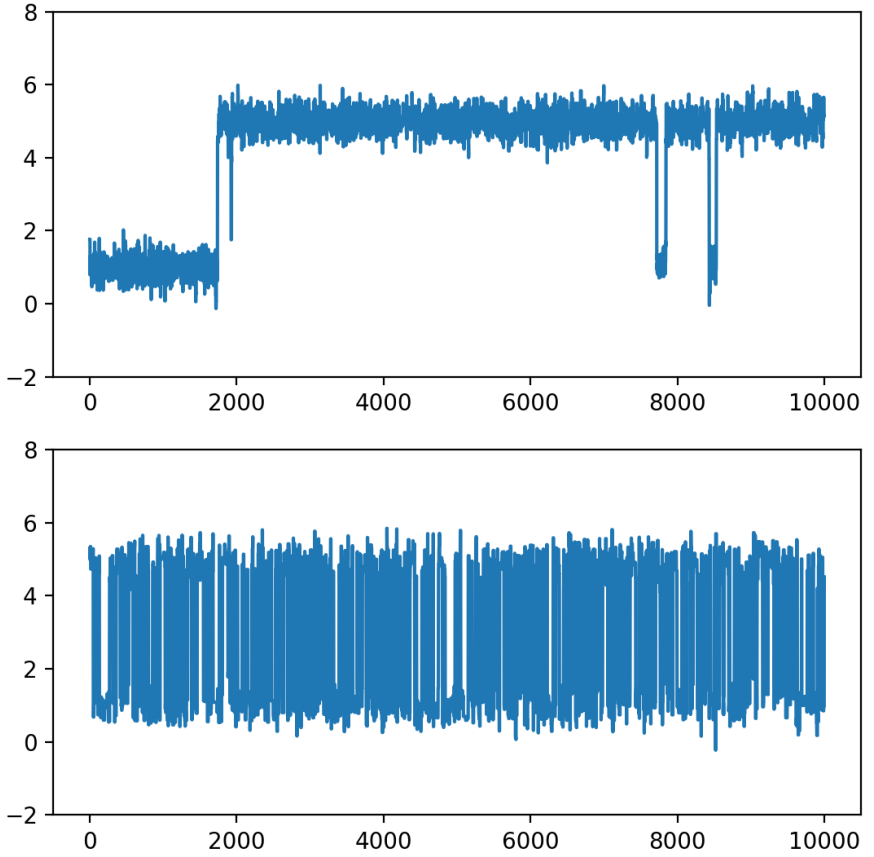
MCMC algorithms produce only approximate samples from the target distribution. It is important to assess the quality of these samples using diagnostics in terms of **autocorrelation**, **mixing** and **convergence**.

Trace plots of the Markov chains can reveal mixing, autocorrelation, and convergence issues:

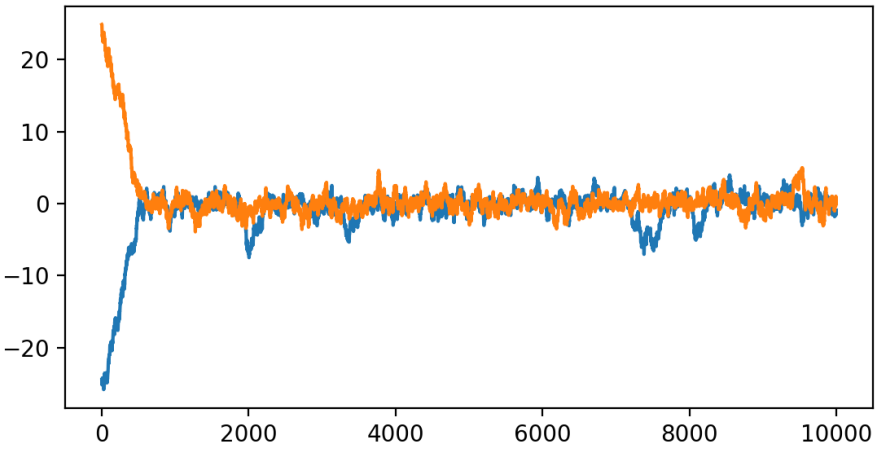
- High autocorrelation is indicated by chains that exhibit strong correlations between successive samples.
- Poor mixing is indicated by chains that explore the state space slowly or get stuck in certain regions.
- When running multiple chains, lack of convergence is indicated by chains that do not overlap or converge to the same distribution.



High vs. low autocorrelation. For highly autocorrelated chains, mixing is poor and effective sample size is low.



Poor vs. good mixing for a multimodal target. Poor mixing can be due to small proposal steps or energy barriers between modes.



Two chains run on the same target but initialized differently.
Traces show convergence to a same distribution.

Acceptance rate

The first quantitative diagnostic to monitor is the **acceptance rate**, defined as the fraction of proposed samples that are accepted.

- A very low acceptance rate (e.g., < 10%) indicates that the proposal distribution is making too large jumps, leading to many rejections.
- A very high acceptance rate (e.g., > 80%) suggests that the proposal distribution is making too small jumps, resulting in slow exploration of the state space.
- An optimal acceptance rate often lies between 20% and 40%, depending on the dimensionality of the target distribution.

Effective Sample Size

Quantitatively, the **effective sample size**

$$\mathrm{ESS} = \frac{T}{1 + 2\sum_{k=1}^{\infty} \rho_k}$$

accounts for autocorrelation in MCMC samples, where T is the number of samples in the chain and

$$\rho_k = \frac{\mathrm{Cov}(X_t, X_{t+k})}{\mathrm{Var}(X_t)}$$

is the **autocorrelation at lag** k computed over the samples.

A higher ESS indicates more independent samples and better mixing.

Integrated Autocorrelation Time

The **integrated autocorrelation time**

$$\tau_{\text{int}} = 1 + 2\sum_{k=1}^{\infty} \rho_k$$

quantifies the number of correlated samples equivalent to one independent sample.

A smaller τ_{int} indicates better mixing and more efficient sampling.

Note that $\text{ESS} = \frac{T}{\tau_{\text{int}}}$.

Gelman-Rubin Statistic

The **Gelman-Rubin statistic** $\hat{R}$ serves as a convergence diagnostic when running M chains in parallel, each with T samples (after burn-in). Let

$$W = \frac{1}{M} \sum_{m=1}^{M} s_m^2$$

be the within-chain variance, where s_m^2 is the sample variance of chain m, and

$$B = \frac{T}{M-1} \sum_{m=1}^{M} (\bar{x}_m - \bar{x})^2$$

be the between-chain variance, where $\bar{x}_m$ is the sample mean of chain m and $\bar{x}$ is the overall mean across chains. Then,

$$\hat{R} = \sqrt{\frac{\frac{T-1}{T} W + \frac{1}{T} B}{W}}.$$

Typically,

- values of $\hat{R}$ approximately 1 indicate convergence across chains;
- values greater than 1.1 suggest lack of convergence.

The end.