UCSB, Physics 129L, Computational Physics
Lecture notes, Week 8

Zihang Wang (UCSB), zihangwang@ucsb.edu

March 3, 2025

Contents

1 Monte Carlo Method and Markov Chain Monte Carlo (MCMC)

MCMC is a specialized Monte Carlo method used when direct sampling from a complex probability distribution is difficult. Instead of independent random samples, it constructs a Markov chain where each sample depends on the previous one, gradually converging to the desired distribution. The general procedure involves constructing a Markov chain whose stationary distribution is the target distribution.

1.1 Importance Sampling in Bayesian Inference

Consider the Bayes' formula when approximating the **posterior** using a **prior** distribution $p(\boldsymbol{\theta})$ and a **likelihood function** $p(\mathbf{y}|\boldsymbol{\theta})$,

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y})}. \tag{1}$$

We can see that $p(\boldsymbol{\theta}|\mathbf{y})$ is always normalized.

We would like to calculate physical quantities $h(\theta)$ using the posterior distribution with a fixed input dataset $\mathbf{y}$,

$$\mathbb{E}(h(\boldsymbol{\theta})) = \int h(\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}, \tag{2}$$

and use Bayes' formula, we have,

$$\mathbb{E}(h(\boldsymbol{\theta})) = \int h(\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta} = \frac{\int h(\boldsymbol{\theta})p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}}{\int d\boldsymbol{\theta}p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}. \tag{3}$$

Instead of sampling directly from $p(\mathbf{y}|\boldsymbol{\theta})$, which may be difficult, we sample from a simpler proposal distribution $q(\boldsymbol{\theta})$, using the **importance sampling** we discussed previously. The posterior expectation can be rewritten in terms of the weight function $w(\boldsymbol{\theta})$,

$$\mathbb{E}(h(\boldsymbol{\theta})) = \int h(\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta} = \frac{\int h(\boldsymbol{\theta})w(\boldsymbol{\theta})q(\boldsymbol{\theta})d\boldsymbol{\theta}}{\int d\boldsymbol{\theta}w(\boldsymbol{\theta})q(\boldsymbol{\theta})}, \quad w(\boldsymbol{\theta}) = \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{q(\boldsymbol{\theta})}. \tag{4}$$

We can approximate this with the Monte Carlo integration,

$$\mathbb{E}(h(\boldsymbol{\theta})) = \frac{\frac{1}{N}\sum_{i=1}^{N} h(\boldsymbol{\theta}_i)w(\boldsymbol{\theta}_i)}{\frac{1}{N}\sum_{i=1}^{N} w(\boldsymbol{\theta}_i)}, \tag{5}$$

where $\boldsymbol{\theta}_i$ is sampled from the proposal distribution $q(\boldsymbol{\theta}_i)$, $\boldsymbol{\theta}_i \sim q(\boldsymbol{\theta})$. One of the examples for the above expectation values is the posterior distribution: it can be approximated as the **sum of delta functions**,

$$p(\boldsymbol{\theta}'|\mathbf{y}) = \mathbb{E}(\delta(\boldsymbol{\theta} - \boldsymbol{\theta}')) = \int \delta(\boldsymbol{\theta} - \boldsymbol{\theta}')p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta} = \frac{\frac{1}{N}\sum_{i=1}^{N} w(\boldsymbol{\theta}_i)\delta(\boldsymbol{\theta} - \boldsymbol{\theta}')}{\frac{1}{N}\sum_{i=1}^{N} w(\boldsymbol{\theta}_i)}. \tag{6}$$

If some weights $w(\boldsymbol{\theta})$ are much larger than 1, it means that the proposal distribution $q(\boldsymbol{\theta})$ is poorly aligned with the target distribution $p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$. Ideal when $w(\boldsymbol{\theta})$. When the values $w(\boldsymbol{\theta}) \gg 1$, since some weights are very large while others are very small, the estimate is dominated by a few high-weight samples, leading to high variance in the estimate. This means we need many more samples to get accurate results, therefore, the **proposal function should be as close to posterior as possible**.

The marginal likelihood under a model M (e.g. Gaussian or Poisson),

$$p(\mathbf{y}) \to p(\mathbf{y}|M) = \int d\boldsymbol{\theta} p(\mathbf{y}|\boldsymbol{\theta}, M) p(\boldsymbol{\theta}, M), \tag{7}$$

is important when comparing different models. For example, if we want to select the best model among competing models M_1, M_2, we need the normalization constant for each model, and define the **Bayesian factor**,

$$\text{BF} = \frac{p(\mathbf{y}|M_1)}{p(\mathbf{y}|M_2)}, \tag{8}$$

and use for **hypothesis testing**: If BF > 1, model M_1 is preferred over model M_2 based on the data.

The marginal likelihood can also be used to calculate moments (moment generating function or "partition function"). However, as the dimension increases, the **marginal integral becomes harder to sample and the proposal function becomes harder to guess**.

1.2 Relative posterior distribution in Bayesian Inference

In various of cases, we do not have to calculate this marginal distribution directly **if we only care about relative probabilities, not absolute probabilities.** For example, if we want to know the ratio between two expectation values, we have,

$$\frac{\mathbb{E}(h(\boldsymbol{\theta}))}{\mathbb{E}(h(\boldsymbol{\theta'}))} = \frac{\frac{1}{N}\sum_{i=1}^{N} h(\boldsymbol{\theta}_i) w(\boldsymbol{\theta}_i)}{\frac{1}{N}\sum_{i=1}^{N} w(\boldsymbol{\theta}_i)} \cdot \frac{\frac{1}{N}\sum_{i=1}^{N} w(\boldsymbol{\theta}_i)}{\frac{1}{N}\sum_{i=1}^{N} h(\boldsymbol{\theta}'_i) w(\boldsymbol{\theta}'_i)} = \frac{\frac{1}{N}\sum_{i=1}^{N} h(\boldsymbol{\theta}_i) w(\boldsymbol{\theta}_i)}{\frac{1}{N}\sum_{i=1}^{N} h(\boldsymbol{\theta}'_i) w(\boldsymbol{\theta}'_i)}. \tag{9}$$

We note that the marginal likelihood cancels since they only depends on the data, $p(\mathbf{y}) \sim \sum_{i=1}^{N} w(\boldsymbol{\theta}_i), \boldsymbol{\theta}_i \sim q(\boldsymbol{\theta})$.

For example, the posterior distribution can be approximated as the sum of delta functions,

$$p(\boldsymbol{\theta'}|\mathbf{y}) = \int \delta(\boldsymbol{\theta} - \boldsymbol{\theta'}) p(\boldsymbol{\theta}|\mathbf{y}) p(\boldsymbol{\theta}) d\boldsymbol{\theta} \sim \frac{1}{N} \sum_{i=1}^{N} w(\boldsymbol{\theta}_i) \delta(\boldsymbol{\theta} - \boldsymbol{\theta'}), \tag{10}$$

and the **histogram** will converge to the above posterior as sample size grows larger. **However**, even with the relative posterior distribution, we have to estimate a **global proposal function** $q(\boldsymbol{\theta})$ such that it is a good approximation of the posterior distribution so that the weights are approximately 1. When the degree's of freedom is large and/or posterior distribution is complex, we **may not have a good estimation on the global proposal function**. As we discussed previously, this might result in insufficient sampling problem. Instead of proposing $q(\boldsymbol{\theta})$ directly, we estimate the following **conditional proposal function** $q(\boldsymbol{\theta}|\boldsymbol{\theta'})$, where the proposal is adapted based on the current sample guess

$\boldsymbol{\theta'}$. We should note that the s**ample in this case is not the measurements y, but the hyper parameters $\boldsymbol{\theta}$.**

In other words, we convert the task in finding a good proposal function $q(\boldsymbol{\theta})$ into a stochastic sequence $\{\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}\}$ of guided search via the conditional proposal function $q(\boldsymbol{\theta}|\boldsymbol{\theta'})$, and hope the **stationary distribution** of this sequence is the posterior distribution we are looking for. This is the idea behind the **Markov Chain Monte Carlo (MCMC)**, where the conditional proposal function only depends on the current guess.

1.3 Metropolis-Hastings Algorithm

A widely used MCMC method is the **Metropolis-Hastings algorithm**. Given the current state $\boldsymbol{\theta}_t$, one proposes a new candidate $\boldsymbol{\theta}$ from a **Markov chain** with a proposal distribution $q(\boldsymbol{\theta}|\boldsymbol{\theta}_t)$ since only based on the current state $\boldsymbol{\theta}_t$.

The candidate is accepted with probability,

$$r(\boldsymbol{\theta}, \boldsymbol{\theta}_t) = \min\left\{1, \frac{p(\mathbf{y}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})\, q(\boldsymbol{\theta}_t|\boldsymbol{\theta})}{p(\mathbf{y}|\boldsymbol{\theta}_t)\, p(\boldsymbol{\theta}_t)\, q(\boldsymbol{\theta}|\boldsymbol{\theta}_t)}\right\}. \tag{11}$$

If the proposed $\boldsymbol{\theta}$ is accepted, then $\boldsymbol{\theta}_t \to \boldsymbol{\theta}$; otherwise, the chain remains at $\boldsymbol{\theta}_t$.

Recall the master equation we discussed previously,

$$\frac{dp(\boldsymbol{\theta}_t)}{dt} = Q_{(\boldsymbol{\theta_t},\boldsymbol{\theta})} p(\boldsymbol{\theta}) - Q_{(\boldsymbol{\theta},\boldsymbol{\theta_t})}\, p(\boldsymbol{\theta}_t), \quad Q_{(\boldsymbol{\theta},\boldsymbol{\theta_t})} = p(\mathbf{y}|\boldsymbol{\theta}_t)\, q(\boldsymbol{\theta}|\boldsymbol{\theta}_t), \quad Q_{(\boldsymbol{\theta_t},\boldsymbol{\theta})} = p(\mathbf{y}|\boldsymbol{\theta})\, q(\boldsymbol{\theta}_t|\boldsymbol{\theta}), \tag{12}$$

and in this case, we have the following to conditions. When the rate is larger than zero, probability increases as we move from $\theta_t \to \theta$ over a dt, the proposed state is always accepted because it moves toward higher-probability regions,

$$p(\boldsymbol{\theta}) - p(\boldsymbol{\theta}_t) \sim \frac{dp(\boldsymbol{\theta}_t)}{dt} > 0 \to \frac{Q_{(\boldsymbol{\theta_t},\boldsymbol{\theta})} p(\boldsymbol{\theta})}{Q_{(\boldsymbol{\theta},\boldsymbol{\theta_t})}\, p(\boldsymbol{\theta}_t)} > 1, \tag{13}$$

A special case is the **random walk Metropolis** algorithm, where the **proposal distribution is symmetric**, $q(\boldsymbol{\theta}|\boldsymbol{\theta}_t) = q(\boldsymbol{\theta}_t|\boldsymbol{\theta})$. On example will be,

$$\boldsymbol{\theta} = \boldsymbol{\theta}_t + \epsilon(\boldsymbol{\theta}_t), \quad \epsilon(\boldsymbol{\theta}_t) \sim \mathcal{N}(0, \mathbf{C}), \tag{14}$$

where $\mathbf{C}$ is a **correlation matrix**.

In this scenario, the acceptance ratio simplifies to,

$$r(\boldsymbol{\theta}, \boldsymbol{\theta}_t) = \min\left\{1, \frac{p(\mathbf{y}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})}{p(\mathbf{y}|\boldsymbol{\theta}_t)\, p(\boldsymbol{\theta}_t)}\right\}. \tag{15}$$

While the above ratio does not depends on the conditional proposal function $q(\boldsymbol{\theta}|\boldsymbol{\theta}_t)$, the generation of new step $\boldsymbol{\theta}_t \to \boldsymbol{\theta}$ does follow the conditional proposal function $q(\boldsymbol{\theta}|\boldsymbol{\theta}_t)$.

When the **detailed balance** condition is achieved,

$$\frac{p(\mathbf{y}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})}{p(\mathbf{y}|\boldsymbol{\theta}_t)\, p(\boldsymbol{\theta}_t)} = 1, \tag{16}$$

$\boldsymbol{\theta}$ follows the posterior distribution,

$$\boldsymbol{\theta} \sim p(\boldsymbol{\theta}|\mathbf{y}) \approx p(\mathbf{y}|\boldsymbol{\theta})\, p(\boldsymbol{\theta}). \tag{17}$$

Rather than drawing independent samples from a global proposal function $q(\boldsymbol{\theta})$, MCMC constructs a Markov chain with a conditional proposal function $q(\boldsymbol{\theta}|\boldsymbol{\theta_t})$ whose equilibrium (stationary) distribution is the posterior.

1.4 Gibbs Sampling

When the posterior distribution is **multidimensional** and the conditional probability for each parameter are tractable, **Gibbs sampling** can be employed. Suppose the parameter is $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_d)$, and the joint distribution is given by,

$$p(\mathbf{y}, \boldsymbol{\theta}) = p(\theta_1, \ldots, \theta_d, \mathbf{y}) = p(\mathbf{y}|\boldsymbol{\theta})\, p(\boldsymbol{\theta}). \tag{18}$$

Gibbs sampling **updates each component sequentially** by sampling from the corresponding full conditional distribution (fix all $\boldsymbol{\theta}$ other than $\boldsymbol{\theta}_i$ in the above joint distribution $p(\mathbf{y}, \boldsymbol{\theta})$,

$$\theta_{i,t+1} \sim p\Big(\theta_i \,\Big|\, \theta_{1,t+1}, \ldots, \theta_{i-1,t+1}, \theta_{i+1,t}, \ldots, \theta_{d,i+1}, \mathbf{y}\Big), \tag{19}$$

where θ_i is the i-th component of $\boldsymbol{\theta}$, and $\theta_{i,t+1}$ is the estimation at step $t+1$. $\theta_{j<i,t+1}, \cdots$ are components that have been updated at previous step $t+1$, and $\theta_{j<i,t}, \cdots$ are components that have not yet been updated at the current step t for $i = 1, \ldots, d$. By cycling through all parameters, the sampler gradually builds up a sequence of samples that converges to the posterior distribution. In this case, the acceptance is always one.

The joint distribution $p(\mathbf{y}, \boldsymbol{\theta})$ becomes the posterior distribution $p(\boldsymbol{\theta}|\mathbf{y})$ after the Markov chain has converged to a stationary distribution, following the Bayes' rule,

$$\boldsymbol{\theta} \sim p(\boldsymbol{\theta}|\mathbf{y}) \approx p(\mathbf{y}, \boldsymbol{\theta}). \tag{20}$$

1.5 Hamiltonian Monte Carlo (HMC)

For high-dimensional or complex posterior distributions, **Hamiltonian Monte Carlo** (HMC) is very important. HMC introduces **auxiliary momentum variables** and **uses the Hamiltonian dynamics to propose new states**. The incorporation of gradient information allows the sampler to make longer, informed moves in the parameter space while maintaining a high acceptance rate. This method is particularly advantageous in Bayesian inference tasks.

Hamiltonian Monte Carlo (HMC) introduces an auxiliary momentum variable $\mathbf{p}$ in addition to the position variable $\boldsymbol{\theta}$ (the parameters of interest). They form **a state in phase space**. The Hamiltonian $H(\boldsymbol{\theta}, \mathbf{p})$ represents the total energy of the system, defined as:

$$H(\boldsymbol{\theta}, \mathbf{p}) = U(\boldsymbol{\theta}) + K(\mathbf{p}), \tag{21}$$

where $U(\boldsymbol{\theta})$ is the **potential energy** as the **negative log posterior distribution**,

$$\begin{aligned} U(\boldsymbol{\theta}) &= -\log(p(\boldsymbol{\theta}|\mathbf{y})) = -\log\left(\frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y})}\right) \\ &= -\log\left[p(\mathbf{y}|\boldsymbol{\theta})\right] - \log[p(\boldsymbol{\theta})] + \log(p(\mathbf{y})) = -\log\left[p(\mathbf{y}|\boldsymbol{\theta})\right] - \log[p(\boldsymbol{\theta})], \end{aligned} \tag{22}$$

where we can neglect the constant contribution from the marginal likelihood function $\log(p(\mathbf{y}))$. $K(\mathbf{p}) = \frac{1}{2}\mathbf{p}^T M^{-1}\mathbf{p}$ is the **auxiliary kinetic energy**, and $\mathbf{p}$ is usually modeled as a **Gaussian**.

The evolution of the system is governed by Hamilton's equations:

$$\frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}t} = \frac{\partial H}{\partial \mathbf{p}}, \quad \frac{\mathrm{d}\mathbf{p}}{\mathrm{d}t} = -\frac{\partial H}{\partial \boldsymbol{\theta}}. \tag{23}$$

These equations describe how position and momentum change over time, simulating a physical system. This comes with an very important assumption: the total Hamiltonian is conserved when we evolve the system.

To numerically solve Hamilton's equations, HMC uses a **leapfrog integrator**, which approximates the trajectory and mostly preserves reversibility and phase-space volume. The updates are performed iteratively with **a step size** ϵ and a **number of steps** L. The product $t = \epsilon L$ gives the total time of evolution.

After simulating this trajectory in phase space over $t = \epsilon L$, in an ideal case where the integrator perfectly conserves the Hamiltonian, **we always accept** since the following always holds true,

$$\Delta H = H(\boldsymbol{\theta}_{\text{current}}, \mathbf{p}_{\text{current}}) - H(\boldsymbol{\theta}_{\text{evolved}}, \mathbf{p}_{\text{evolved}}) = 0. \tag{24}$$

However, the leapfrog integrator slightly change the Hamiltonian (increase or decrease) after time t. If we always accept when it decreases the energy and always reject when it increases in energy,

$$\text{always accept: } \Delta H < 0, \quad \text{always reject: } \Delta H > 0, \tag{25}$$

the Hamiltonian will be bounded in the long run, but eventually approach zero $H \to 0$.

On the other hand, if we always accept when it increases the energy and always reject when it decreases in energy,

$$\text{always reject: } \Delta H < 0, \quad \text{always accept: } \Delta H > 0, \tag{26}$$

the Hamiltonian will be unbounded in the long run, leading to numerical instability.

Therefore, to balance the two contribution, we have,

$$\text{always accept: } \Delta H < 0, \quad \text{accept with probability } \exp(-\Delta H)\text{: } \Delta H > 0. \tag{27}$$

In other words, after simulating this trajectory in phase space over $t = \epsilon L$, the new state is accepted or rejected based on the **Metropolis criterion**,

$$P_{\text{accept}} = \min\left(1, e^{H(\boldsymbol{\theta}_{\text{current}}, \mathbf{p}_{\text{current}}) - H(\boldsymbol{\theta}_{\text{evolved}}, \mathbf{p}_{\text{evolved}})}\right), \tag{28}$$

which corrects the numerical error in Hamiltonian accumulated by the leapfrog integrator.

Once the decision to accept or reject the proposed state is made, **the momentum is discarded**, and **a new momentum p** is sampled independently from a **Gaussian distribution** for the next step. In general, it can be distributions other than Gaussian. We update the parameter to the evolved one $\boldsymbol{\theta} \to \boldsymbol{\theta}_{\text{evolved}}$.

The idea of the HMC is that, with sufficiently large number of steps, when tracing out the auxiliary momentum, the system will achieve an equilibrium (stationary) posterior distribution, sampled from the exponential of the potential energy,

$$\boldsymbol{\theta} \sim \exp(-U(\boldsymbol{\theta})) = p(\boldsymbol{\theta}|\mathbf{y}) \approx p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}). \tag{29}$$

1.6 General MCMC Procedure

- **Define the posterior** $p(\boldsymbol{\theta}|\mathbf{y}) = p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})/p(\mathbf{y})$. We want to sample the posterior, up to a normalization constant $p(\mathbf{y})$ (hard to obtain).

- **Proposal Mechanism** Since we do not know the normalization constant $p(\mathbf{y})$ but we still want to sample from the posterior distribution, we can either find a global proposal function $q(\boldsymbol{\theta})$ that directly sample $p(\mathbf{y})$, or we can define the conditional proposal distribution $q(\boldsymbol{\theta}|\boldsymbol{\theta}_t)$, and iteratively approach to the stationary distribution: **Drawing samples from the stationary distribution is identical as drawing samples from the posterior distribution.**

  Here are how new sample candidates are proposed,

  - **Random walk Metropolis:** $\boldsymbol{\theta} = \boldsymbol{\theta}_t + \epsilon(\boldsymbol{\theta}_t)$, such that $\epsilon(\boldsymbol{\theta}_t) \sim \mathcal{N}(0, \mathbf{C})$, where $\mathbf{C}$ is the correlation matrix.
  - **Gibbs Sampling:**

$$\theta_{i,t+1} \sim p\Big(\theta_i \,\Big|\, \theta_{1,t+1}, \ldots, \theta_{i-1,t+1}, \theta_{i+1,t}, \ldots, \theta_{d,i+1}, \mathbf{y}\Big), \tag{30}$$

  - **Hamiltonian Monte Carlo:**

$$H(\boldsymbol{\theta}, \mathbf{p}) = U(\boldsymbol{\theta}) + K(\mathbf{p}), \tag{31}$$

- **Initialize the Markov Chain**
  - Start with an arbitrary initial state $\boldsymbol{\theta}_0$.
  - It does not need to be from the posterior distribution.

- **Multiple Markov chains** are often initialized from different starting points $\boldsymbol{\theta}_0, \boldsymbol{\theta}_0', \cdots$, in order to confirm convergence.

- **Iterate the MCMC Algorithm**

  1. **Generate a Proposal $\boldsymbol{\theta}$** using the above proposal mechanism.
  2. **Compute the Acceptance Probability, if any.**
     For the **Metropolis-Hastings algorithm**:

$$r(\boldsymbol{\theta}, \boldsymbol{\theta}_t) = \min\left\{1, \frac{p(\mathbf{y}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})\, q(\boldsymbol{\theta}_t|\boldsymbol{\theta})}{p(\mathbf{y}|\boldsymbol{\theta}_t)\, p(\boldsymbol{\theta}_t)\, q(\boldsymbol{\theta}|\boldsymbol{\theta}_t)}\right\}. \tag{32}$$

  3. **Accept or Reject x':**
     - Draw a random $u \sim \text{Uniform}(0, 1)$.
     - If $u \leq r(\boldsymbol{\theta}, \boldsymbol{\theta}_t)$, accept $\boldsymbol{\theta}$ and set $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}$.
     - Otherwise, reject $\boldsymbol{\theta}$ and set $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t$.

- **Repeat for Many Iterations**
  - The Markov chain gradually converges to the posterior distribution.

- **Discard Burn-in Samples**
  - **The early samples may not be from the stationary distribution**.
  - Discard an initial portion (e.g., first 10–20% of samples).

- **Analyze the Samples**
  - Compute estimates, histograms, expectation values, etc.
  - Check **convergence diagnostics** (e.g., **Gelman-Rubin statistic, autocorrelation**).

    The **Gelman-Rubin diagnostic** assesses the **convergence of multiple Markov chains** by comparing the between-chain and within-chain variances. Given M chains, each of length N, we have,

$$B = \frac{N}{M-1}\sum_{m=1}^{M}(\bar{\boldsymbol{\theta}}_m - \bar{\boldsymbol{\theta}})^2, \quad \text{(Between-chain variance)}$$

$$W = \frac{1}{M}\sum_{m=1}^{M}\sum_{n=1}^{N}\left(\frac{1}{N}(\boldsymbol{\theta}_{m,n} - \boldsymbol{\theta}_m)^2\right), \quad \text{(Within-chain variance)}$$

    where $\bar{\boldsymbol{\theta}}_m$ is the mean of chain m, $\bar{\boldsymbol{\theta}}$ is the mean across all chains, $\boldsymbol{\theta}_{m,n}$ is the n-th sample from chain m.

The estimated variance of the posterior distribution is,

$$\hat{V} = \frac{N-1}{N}W + \frac{B}{N} \tag{33}$$

The **potential scale reduction factor** (PSRF) is,

$$\hat{R} = \sqrt{\frac{\hat{V}}{W}} \tag{34}$$

For convergence, we expect $\hat{R} \approx 1$. Values significantly greater than 1 indicate non-convergence.

Autocorrelation measures how correlated a sample $\boldsymbol{\theta}_t$ is with its lagged values $\boldsymbol{\theta}_{t+k}$ in a single Markov chain. The autocorrelation function at lag k is:

$$\rho_k = \frac{\mathbb{E}[(\boldsymbol{\theta}_t - \mu)(\boldsymbol{\theta}_{t+k} - \mu)]}{\sigma^2} \tag{35}$$

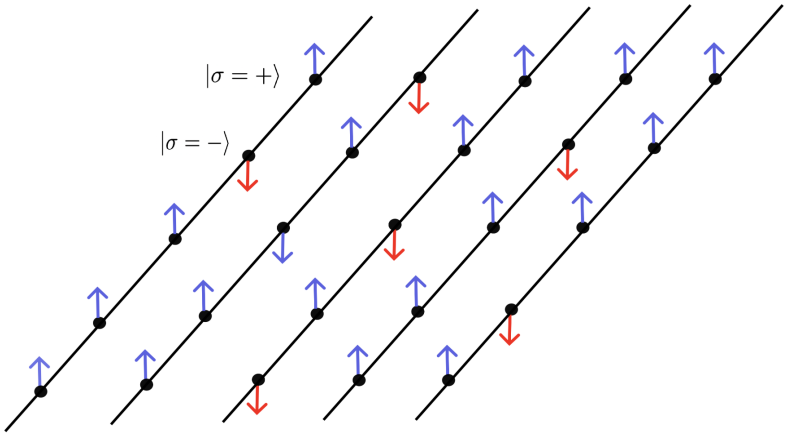
where, $\mu = \mathbb{E}[X_t]$ is the mean of the chain $\sigma^2 = \text{Var}(\boldsymbol{\theta}_t)$ is the variance of the chain.

High autocorrelation suggests poor mixing, requiring longer runs or thinning of the chain.

2 2D Classical Ising Model (Lattice Ising Model with MCMC

The classical Ising model is a mathematical model used in statistical mechanics to study phase transitions in ferromagnetic materials. It was first proposed by the physicist Ernst Ising in 1925 as a simplified representation of the magnetic behavior of certain materials, such as iron. The Ising model has since become a fundamental tool in statistical mechanics and condensed matter physics.

The Ising model is typically defined on a lattice (or grid), which is a regular arrangement of discrete points in space. In this demonstration, we will be focusing on the 2D model. The structure is given below.



2.1 2D Lattice generation and Ising Hamiltonian

We can treat spins lattice as random variables, denoted as $\{S_{x,y}\}$. Let's define a function to initialize a random spin configuration. When working with a numpy array, we can naturally assign its dimensional index to be its special location, with value -1 or 1.

Depending on the arrangement of spins, the Hamiltonian for the classical Ising model is often given by the following expression:

$$H = -J \sum_{\langle i,j \rangle} S_i S_j - B \sum_i S_i \tag{36}$$

where J is the coupling constant, representing the strength of the interaction between neighboring spins. $\langle i, j \rangle$ denotes a sum over pairs of nearest-neighbor spins, S_i represents the spins, and B represents an external magnetic field. You can see this is the classical Ising model: the Hamiltonian has no exchange interactions.

2.2 Statistical Description of the Ising Model

Let's look at various thermodynamic properties and statistical measures. The fundamental quantity is the partition function,

$$Z = \sum_{i,j} e^{-\beta H} \tag{37}$$

where $\beta = \frac{1}{k_B T}$. For the Ising model, we have:

$$Z = e^{-\beta\left(-J \sum_{\langle i,j \rangle} S_i S_j - B \sum_i S_i\right)}. \tag{38}$$

In particular, the joint distribution is a probability mass function (since we are working on a finite system), and it has the form:

$$P(S) = \frac{1}{Z} e^{-\beta H}, \tag{39}$$

where $S = \{S_1, S_2, \cdots, S_L\}$, and each S defines a unique energy. As one can see, it is very hard to draw samples from this multivariable distribution. It is easy to see that the total number of unique spin configurations exponentially grows with the system's dimension, 2^{L^2}, making computations difficult for large system sizes $L > 20$.

2.3 Gibbs Sampler on Ising Model

Calculating the exact partition function and obtaining the exact Boltzmann weight for a spin configuration becomes hard as we increase the system size. Let's consider a conditional distribution function: a variable S_i given the rest of the spins,

$$P(S_i \mid S_{\text{rest}}) = \frac{e^{\beta\left(-2JS_i \sum_{\langle j \rangle} S_j - B \sum_j S_j - BS_i\right)}}{1 + e^{\beta\left(-2JS_i \sum_{\langle j \rangle} S_j - B \sum_j S_j - BS_i\right)}}. \tag{40}$$

A Gibbs sampling process involves drawing a sample from the distribution of that variable while fixing the values of all other variables.

For each variable in the distribution, sample a new value from its conditional distribution given the current values of all other variables (e.g. in our case, for each spin site).

2.4 Magnetization, Landau Theory, and Phase Transition of the Ising Model

Magnetization is a measure of the average magnetic moment per spin in a given direction. In the context of the Ising model, the magnetic moment is represented by the sum of the spins. The **magnetization** M is defined as,

$$M = \frac{1}{N} \sum_i S_i \tag{41}$$

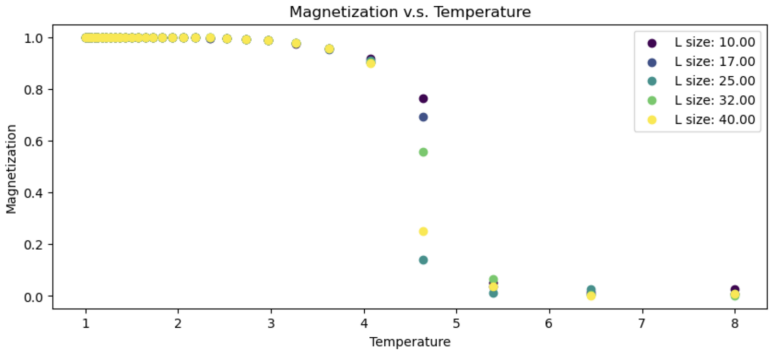
where $N = L^2$ is the total number of spins. In the absence of an external magnetic field ($B = 0$), the Ising model exhibits a spontaneous magnetization at low temperatures, where most spins tend to align in the same direction. In the Ising model, the magnetization serves as an **order parameter** that indicates the presence of a magnetic order in the system. The order parameter is often used to characterize the different phases of the model, especially in the study of phase transitions.

Landau theory, in the context of phase transitions, is a phenomenological approach that describes the free energy of a system near a critical point. It provides a framework for understanding symmetry-breaking mechanisms and the emergence of order parameters. Near a second-order phase transition, the

Landau free energy can often be written as a Taylor expansion in terms of an order parameter, the magnetization,

$$F = F_0 + a(T - T_c)M^2 + bM^4 + \dots \tag{42}$$

where F_0 is the free energy at the critical temperature (T_c), a and b are phenomenological coefficients, and T is the temperature.



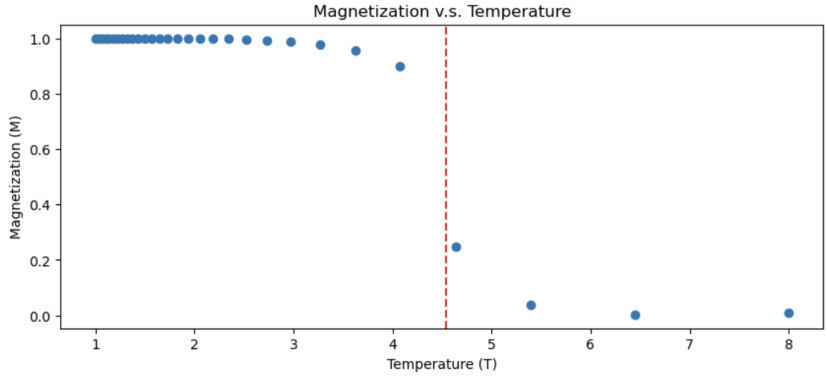
2.5 magnetization at different temperature

We can also look at the MC convergence at different lattice size.

In 2D Ising model, the critical temperature can be calculated via the following,

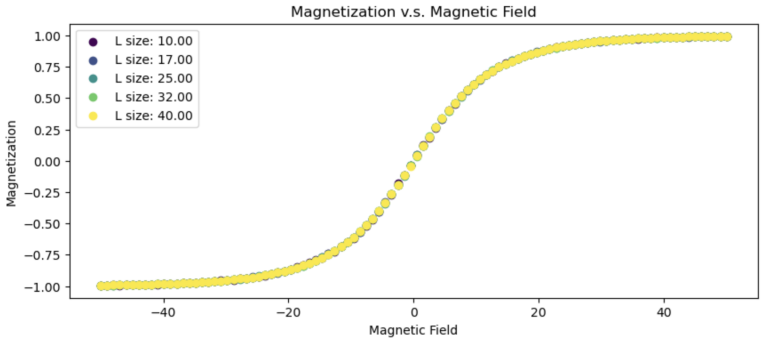
$$T_C = \frac{2J}{\log(1+\sqrt{2})}, \tag{43}$$

and it is labeled as the red vertical dash-line in the following figure,



2.6 Classical Magnetic Field dependence: magnetization of the 2D Ising model

We can look at the MC convergence at different lattice size.

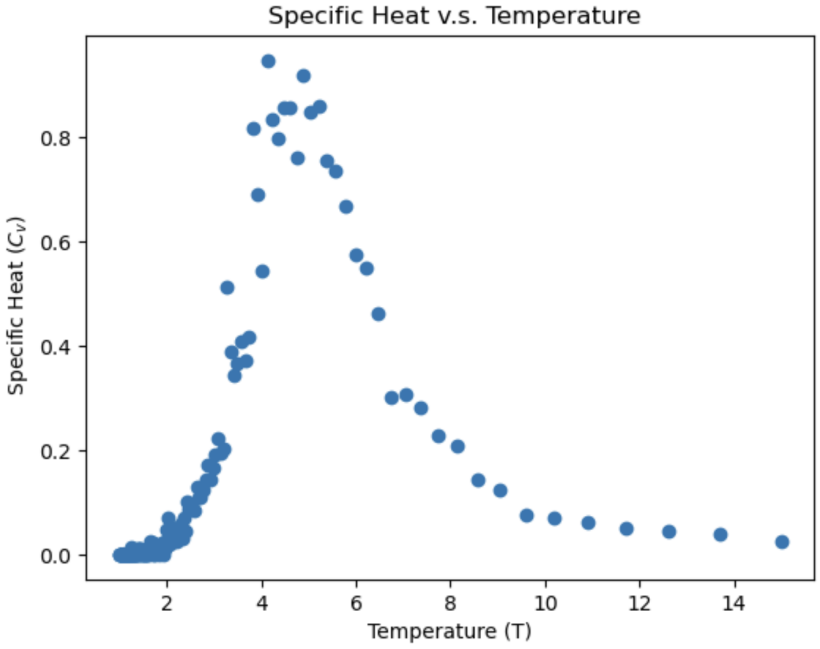


2.7 Specific Heat of the 2D Ising Model

The specific heat (C_v) measures the amount of heat energy required to change the temperature of a substance by a unit temperature. It can be related to the variance of the energy in the 2D Ising model as:

$$C_v = \frac{\beta^2}{N} \left(\langle E^2 \rangle - \langle E \rangle^2 \right), \tag{44}$$

where N is the total number of spins.
We can also look at the temperature dependence,



2.8 Magnetic Susceptibility of the 2D Ising model

Similarly, the magnetic susceptibility quantifies the material's ability to become magnetized per unit change in the external magnetic field strength,

$$\chi = \frac{\beta}{N} \left(\langle M^2 \rangle - \langle M \rangle^2 \right). \tag{45}$$

In the following figure, we have the temperature dependence on magnetic field susceptibility.