Bayesian Inference and Modern MCMC Sampling

Leader: Samaroo, Kai J S.
Writer: Guanzhong Yang, Samuel Patel, Tan, Nathan
March 2026

1. READING NOTES

1.1 Fundamental Ideas of Bayesian School

- **Core Assumption:**
  Treat parameters as random variables.
- **Prior Distribution $\pi(\theta)$:**
  The distribution specified for parameters before observing data.
- **Posterior Distribution $\pi(\theta|\mathcal{D})$:**
  Update the understanding of parameters by combining prior and likelihood.
- **Likelihood Function $\mathcal{L}(\theta) = \mathrm{P}(\mathcal{D}|\theta)$:**
  The probability of samples appearing given parameters, where θ is the parameter and $\mathcal{D}$ is the sample set.
- **Bayesian Formula $\pi(\theta|\mathcal{D}) = \frac{\pi(\theta)\cdot\mathcal{L}(\theta)}{\pi(\mathcal{D})}$:**
  Where $\pi(\mathcal{D}) = \int \pi(\theta)\cdot\mathcal{L}(\theta)\mathrm{d}\theta$. When the posterior distribution is difficult to calculate directly, approximation methods are used.

1.2 Monte Carlo Method

- **Key Concepts:**

$$\mathrm{E}_{\pi}[\mathrm{f}(\mathrm{X})] \approx \frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{f}(\mathrm{X}_{\mathrm{i}}), \text{ where } \mathrm{X}_{\mathrm{i}} \sim \pi(\mathrm{x}).$$

For a quantity that is difficult to calculate directly analytically (such as complex integrals, expected values of distributions, probability values), randomly draw a large number of samples from the target distribution, and use the statistical characteristics (frequency, arithmetic mean) of the samples to approximate the true value. The larger the sample size n, the higher the approximation accuracy.

In the sampling process, we generally need to know the CDF of the target distribution. This is because computers essentially randomly draw a sample u from the uniform distribution $\mathrm{U}(0,1)$(computers can only directly generate uniform distribution samples), and then perform the inverse transformation of the CDF on u to get $\mathrm{x} = \mathrm{F}^{-1}(\mathrm{u})$.

However, if the CDF is difficult to calculate, we introduce the following alternative methods.

1.3 Rejection Sampling and Importance Sampling

- **Key Concepts of Rejection Sampling:**

When it is impossible to sample directly from $\pi(x)$, an easy-to-sample "proposal distribution" $q(x)$ is used for assistance. Samples from $\pi(x)$ are obtained by accepting or rejecting samples drawn from $q(x)$.

- The steps for this method:

(1) Find a constant M such that $Mq(x) \geq \pi(x)$ for all x.
(2) Sample x from $q(x)$ and u from the uniform distribution $U(0,1)$.
(3) If $U \leq \frac{\pi(x)}{Mq(x)}$, accept x; otherwise, reject it.

We can even make a simple derivation. Let $D(s|A)$ be the density that sample s is a sample of $\pi(x)$ given acceptance:

$$D(s|A) = \frac{p(A|s)q(x)}{P(A)} = \frac{\frac{\pi(s)}{Mq(s)}q(s)}{P(A)} = \frac{\frac{\pi(s)}{M}}{P(A)} = \frac{\frac{\pi(s)}{M}}{\int \frac{\pi(s)}{Mq(s)}q(s)ds} = \frac{\frac{\pi(s)}{M}}{\frac{1}{M}\int \pi(s)ds} = \pi(s)$$

- **Key Concepts of Importance Sampling:**

When it is impossible to sample directly from $\pi(x)$, an easy-to-sample "proposal distribution" $q(x)$ is used for assistance.

$$E_\pi[f(X)] = \int f(x)\pi(x)dx = \int f(x)\pi(x)dx = \int f(x)\frac{\pi(x)}{q(x)}q(x)dx = E_q[f(X)\frac{\pi(X)}{q(X)}]$$

Its advantage is that we can use importance sampling to reduce variance, because $Var_\pi[f(X)] < Var_q[f(X)\frac{\pi(X)}{q(X)}]$ when q(x) is high where $|f(X)\pi(X)|$ is high.

1.4 Markov Chain Monte Carlo and Metropolis-Hastings (MCMC & MH Algorithm)

Rejection Sampling has some shortcomings, such as low computational efficiency. Especially in high-dimensional spaces, the overlapping area between $\pi(x)$ and $Mq(x)$ will decrease exponentially, leading to an extremely low sampling acceptance rate.

Therefore, we adopt the Markov Chain Monte Carlo method. We use the random walk of the Markov chain to reach the steady state $\pi(x)$, and take the subsequent samples as valid samples of $\pi(x)$. That is, $\lim_{n\to\infty} P_n(x) = \pi(x)$.

Note: For a Markov chain to have a steady state, it needs to satisfy the following three conditions:
(1) Irreducibility: Any two states in the Markov chain can reach each other.
(2) Aperiodicity: There is no fixed period for the chain to return to any state.
(3) Recurrence: Starting from any state i, the probability that the chain will return to i within a finite number of steps is 1.

- The steps for the MH Algorithm:

(1) Starting from the current state x_n, sample a candidate state x* according to the proposal distribution Q(x→x*)(a common proposal distribution is N(x_n, σ^2)).
(2) Calculate the acceptance rate of the candidate state as $\alpha(x_n \to x^*) = \min(1, \frac{\pi(x^*)Q(x^*\to x_n)}{\pi(x_n)Q(x_n\to x^*)})$.
(3) Sample a random number u from the uniform distribution $U(0,1)$. If $u < \alpha(x_n \to x^*)$, accept the move and set $x_{n+1} = x*$; otherwise, set $x_{n+1} = x_n$.
(4) Repeat the above steps until a sufficiently long state sequence is generated. After the chain converges, discard the first N steps, and the remaining states are valid samples conforming to $\pi(x)$.

- Why does the MH algorithm always reach a steady state?

A necessary and sufficient condition for $\pi(x)$ to be the steady-state distribution of the chain is $\pi(y) = \int \pi(x)P(x \to y)$.

If the detailed balance condition ($\pi(x)P(x \to y) = \pi(y)P(y \to x)$) is satisfied, the above steady-state equation must be satisfied. The MH algorithm satisfies the detailed balance condition.

$$P(x \to y) = \min(1, \frac{\pi(y)Q(y \to x)}{\pi(x)Q(x \to y)})Q(x \to y) = \frac{1}{\pi(x)}\min(\pi(x)Q(x \to y), \pi(y)Q(y \to x))$$

$$P(y \to x) = \min(1, \frac{\pi(x)Q(x \to y)}{\pi(y)Q(y \to x)})Q(y \to x) = \frac{1}{\pi(y)}\min(\pi(y)Q(y \to x), \pi(x)Q(x \to y))$$

$$\pi(x)P(x \to y) = \pi(y)P(y \to x)$$

- Problems of the MH Algorithm:

(1) MCMC is essentially a Markov chain, and the samples are correlated rather than independent. The samples are approximately independent only if the proposal distribution is sufficiently good, the burn-in period is long enough, and the thinning interval is large enough.

(2) Regarding the variance of the proposal distribution:

If the variance is too small, the step size is too short. Although the acceptance rate is high, the chain moves slowly and converges extremely slowly.

If the variance is too large, the step size is too long, and it easily jumps to low-probability regions, resulting in an extremely low acceptance rate and the chain being almost stationary.

1.5 Hamiltonian Monte Carlo (HMC) Sampling

MCMC explores the parameter space through random walk. However, in the case of high-dimensional, strongly correlated, and narrow-peaked distributions, the acceptance rate is low, which is likely to cause the problem of wandering without making progress.

HMC replaces "random walk" with "physical dynamics", allowing samples to glide along the energy surface instead of jumping randomly. It regards the Bayesian posterior as a physical potential energy field and then introduces momentum variables to form a Hamiltonian system.

- Steps of the HMC Algorithm:

(1) Artificially set a potential energy for each $\pi(\mathbf{x})$:

$$U(\mathbf{x}) = -\ln(\pi(\mathbf{x})).$$

In this way, where $\pi(\mathbf{x})$ is large, U(x) is small (a valley), which is easier to reach after movement.

(2) Assign an initial velocity to the particle in $\pi(\mathbf{x})$, and the kinetic energy r is randomly sampled from a multivariate normal distribution: $K(r) = \frac{1}{2} r^T M^{-1} r,\ r \sim N(0, M)$

(3) Consider the initial velocity, gradient and step size to track the final position of the particle:

$$r(t + \frac{\epsilon}{2}) = r(t) - \frac{\epsilon}{2} \nabla_x(-\ln \pi)$$

$$x(t + \epsilon) = x(t) + \epsilon M^{-1} r(t + \frac{\epsilon}{2})$$

$$r(t + \epsilon) = r(t + \frac{\epsilon}{2}) - \frac{\epsilon}{2} \nabla_x(-\ln \pi)$$

Take a small step ϵ each time and repeat for L times to obtain (x^*, r^*). The new proposal is $(x^*, -r^*)$.

This is because at this time,

$$Q((x, r) \to (x^*, -r^*)) = Q((x^*, -r^*) \to (x, r)),$$

so the leapfrog method is reversible and the detailed balance is achieved. Moreover,

$$\pi(\mathbf{x}, \mathbf{r}) = \pi(\mathbf{x})\pi(\mathbf{r}),\ \pi(\mathbf{r}) = \pi(-\mathbf{r}),$$

so adding a negative sign has no effect on the result.

(4) If the total energy is almost unchanged, the trajectory is reasonable and the proposal is accepted:

$$\alpha(\mathbf{x} \rightarrow \mathbf{x}^*) = \min(1, \frac{\pi(\mathbf{x}^*, -\mathbf{r}^*)}{\pi(\mathbf{x}, \mathbf{r})})$$

(5) Detailed balance can be achieved, and $\pi(\mathbf{x})$ is the steady-state distribution of the chain.

1.6 NUTS

A major challenge of HMC is how to select ϵ and L. For this reason, we take a different approach and invent the automatic HMC without parameter tuning -- NUTS (No-U-Turn Sampler).

Through NUTS, the particle moves forward by itself and stops automatically when it is "about to turn back and circle".

Then, from all the points passed, a new sample is randomly selected proportionally according to the occurrence probability of each data point.

- Steps of the NUTS Algorithm:

(1) Artificially set a potential energy for each $\pi(\mathbf{x})$, assign an initial velocity to the particle in $\pi(\mathbf{x})$, and the kinetic energy r is randomly sampled from a multivariate normal distribution.
(2) For the two ends of the chain, move the front end in the direction of r with a 50% probability, and move the back end in the direction of $-\mathrm{r}$ with a 50% probability. When moving, take 1 step for the first time, 2 steps for the second time, 4 steps for the third time, and so on.
(3) Stop immediately if the particle starts to turn back and form a U-shape: stop when $(x^+ - x^-)\cdot r^+ \leq 0$ or $(x^+ - x^-)\cdot r^- \leq 0$, where x^- is the rearmost point; x^+ is the foremost point; r^- is the momentum at the rearmost point; and r^+ is the momentum at the foremost point.
(4) Randomly select a new sample proportionally according to the occurrence probability of each data point.
We do not use the acceptance rate α here because we have filtered out the points with low acceptance rates in the step of "randomly selecting a new sample proportionally".

1.7 Variational Inference

Since the true posterior probability distribution is difficult to solve, we can use a simple and manageable distribution to approximate the true posterior probability distribution, and optimize this approximation by minimizing the gap between them.

Assume a simple distribution $q(\theta)$. We use KL divergence to represent the gap between the assumed distribution and the true distribution:

$$\begin{aligned} KL(q(\theta)||\pi(\theta|\mathcal{D})) &= -\int q(\theta)\ \ln[\frac{\pi(\theta|\mathcal{D})}{q(\theta)}]\,d\theta \\ &= \int q(\theta)\ \ln[q(\theta)]\,d\theta - \int q(\theta)\ \ln[\pi(\theta|\mathcal{D})]\,d\theta \\ &= E_q[\ln[q(\theta)]] - E_q[\ln[\pi(\theta|\mathcal{D})]] \\ &= E_q[\ln[q(\theta)]] - E_q[\ln[\pi(\theta)\cdot\mathcal{L}(\theta)]] + E_q[\ln[\pi(\mathcal{D})]] \end{aligned}$$

We note that:
For $E_q[\ln[q(\theta)]]$, since $q(\theta)$ is the distribution we assume, its form is known.
For $E_q[\ln[\pi(\theta)\cdot\mathcal{L}(\theta)]]$, since both the prior function and the likelihood function are known, it can be calculated.
For $E_q[\ln[\pi(\mathcal{D})]]$, since it is independent of θ, it can be regarded as a constant.
We usually define the Evidence Lower Bound (ELBO) as $E_q[\ln[\pi(\theta)\cdot\mathcal{L}(\theta)]] - E_q[\ln[q(\theta)]]$. After rearrangement, we get:

$$\text{Constant} = E_q[\ln[\pi(\mathcal{D})]] = \text{ELBO} + KL(q(\theta)||\pi(\theta|\mathcal{D}))$$

Therefore, to minimize the KL divergence, we only need to maximize the ELBO.