Approximate Inference and MCMC

Advanced Statistical Inference

Simone Rossi

Approximate Inference

1. Consider a posterior distribution

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) = \frac{p(\boldsymbol{y} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{y})}.$$

   Explain why the denominator

$$p(\boldsymbol{y}) = \int p(\boldsymbol{y} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}$$

   is often the main obstacle to exact Bayesian inference. In your answer, distinguish between knowing the posterior only up to proportionality and knowing the fully normalized posterior.

2. In Bayesian linear regression with Gaussian likelihood and Gaussian prior, the posterior can be computed analytically. Explain why this is an example of conjugacy. Then give two examples of models discussed in the course for which exact inference is generally not available.

3. Suppose we approximate a one-dimensional posterior on a grid of equally spaced points $\theta_1, \ldots, \theta_K$ with spacing Δ. Let

$$\widetilde{p}_k = p(\boldsymbol{y} \mid \theta_k)p(\theta_k).$$

   Show how to construct an approximation of:

   - the normalized posterior probability at each grid point;
   - the posterior probability mass assigned to the region around θ_k;
   - the marginal likelihood $p(\boldsymbol{y})$.

4. Why does grid approximation become impractical in high dimensions? If we use K grid points per dimension for a parameter in $\mathbb{R}^D$, how many total grid locations must be evaluated?

Monte Carlo Methods

1. Let $x_1, \ldots, x_N \sim p(x)$ independently and consider the Monte Carlo estimator

$$\hat{I}_N = \frac{1}{N} \sum_{i=1}^{N} f(x_i)$$

for the expectation $\mathbb{E}_{p(x)}[f(x)]$. Show that $\hat{I}_N$ is an unbiased estimator of the target quantity.

2. The variance of the Monte Carlo estimator is

$$\mathrm{Var}(\hat{I}_N) = \frac{\mathrm{Var}(f(x))}{N}.$$

How many times more samples are needed to reduce the standard deviation of the estimator by a factor of 10? What does this imply about the computational cost of high-accuracy Monte Carlo estimates?

3. To estimate π, sample points (x_i, y_i) uniformly from the square $[-1, 1]^2$ and define

$$f(x_i, y_i) = \mathbb{I}(x_i^2 + y_i^2 \leq 1).$$

Write a Monte Carlo estimator of π using these samples. Then suppose that, out of $N = 5000$ samples, 3927 fall inside the unit disk. Compute the estimate.

4. Consider the one-dimensional unnormalized density

$$p^*(\theta) = \exp\left(-\frac{(\theta - 1)^2}{2}\right).$$

You want to estimate

$$\mathbb{E}_{p(\theta)}[\theta^2]$$

using Monte Carlo with the samples

$$\theta^{(1)} = 0, \quad \theta^{(2)} = 1, \quad \theta^{(3)} = 2, \quad \theta^{(4)} = 1.$$

- Write the Monte Carlo estimator.
- Compute its numerical value from the four samples.
- Compare it to the exact value for a Gaussian $\mathcal{N}(1, 1)$.

5. In rejection sampling, assume we know an unnormalized target density $p^*(x)$ and a proposal density $q(x)$ such that

$$p^*(x) \leq Cq(x)$$

for all x. Explain the role of the constant C. Why does choosing C too large make the algorithm inefficient?

6. Consider the target

$$p^*(x) = \exp\Big(-x^2/2\Big)$$

and proposal $q(x) = \mathcal{U}(-3, 3)$.

- Find the smallest constant C such that $p^*(x) \leq Cq(x)$ for all $x \in [-3, 3]$.
- Using the approximation $\mathbb{P}(\text{accept}) \approx 1/C$, estimate the acceptance probability.
- Briefly explain why rejection sampling becomes much less attractive in high dimensions.

Metropolis-Hastings

1. Let the target distribution be known up to a constant:

$$p(\theta \mid \boldsymbol{y}) \propto p(\boldsymbol{y} \mid \theta)p(\theta).$$

Starting from the Metropolis-Hastings acceptance ratio,

$$\alpha(\theta, \theta') = \frac{p(\theta' \mid \boldsymbol{y})\, q(\theta \mid \theta')}{p(\theta \mid \boldsymbol{y})\, q(\theta' \mid \theta)},$$

show that the normalizing constant of the posterior cancels out.

2. Suppose the proposal is symmetric, so that

$$q(\theta' \mid \theta) = q(\theta \mid \theta').$$

Simplify the acceptance ratio and explain the resulting decision rule in words.

3. Consider a target density with unnormalized log-density

$$\log p^*(\theta) = -\frac{\theta^2}{2}.$$

The current state is $\theta^{(t)} = 1$ and the proposed state is $\theta' = 2$.

- Compute the Metropolis acceptance ratio for a symmetric proposal.
- Compute the corresponding acceptance probability.
- If the uniform random draw is $u = 0.5$, state whether the proposal is accepted.

4. Consider a random-walk Metropolis sampler with symmetric Gaussian proposal for the target

$$\log p^*(\theta) = -\frac{\theta^2}{2}.$$

The current state is $\theta^{(t)} = 0.5$ and the proposal is $\theta' = -1$.

- Compute the log acceptance ratio.
- Compute the acceptance probability.
- If $u = 0.8$, determine whether the proposal is accepted.

5. A Metropolis-Hastings chain produces the sequence

$$0.2,\ 0.4,\ 0.4,\ 1.1,\ 1.1,\ 1.1,\ 0.7.$$

How many proposals were rejected? Why do repeated values naturally appear in MCMC output but not in i.i.d. sampling?

6. Explain the purpose of each of the following practical tools for MCMC:

- burn-in;
- multiple chains with different initializations;
- trace plots;
- the potential scale reduction factor $\hat{R}$.

Hamiltonian Monte Carlo

1. In Hamiltonian Monte Carlo, the Hamiltonian is defined as

$$\mathcal{H}(\boldsymbol{\theta}, \boldsymbol{\rho}) = \mathcal{U}(\boldsymbol{\theta}) + \mathcal{K}(\boldsymbol{\rho}),$$

with

$$\mathcal{U}(\boldsymbol{\theta}) = -\log p(\boldsymbol{\theta} \mid \boldsymbol{y}), \qquad \mathcal{K}(\boldsymbol{\rho}) = \frac{1}{2}\boldsymbol{\rho}^\top \boldsymbol{M}^{-1}\boldsymbol{\rho}.$$

Explain the role of the position variable $\boldsymbol{\theta}$, the momentum variable $\boldsymbol{\rho}$, and the mass matrix $\boldsymbol{M}$.

2. Show why the normalization constant of the posterior is not needed to compute the gradient of the potential energy $\nabla_{\boldsymbol{\theta}}\mathcal{U}(\boldsymbol{\theta})$.

3. Write the three leapfrog updates used in HMC for step size ϵ. Why is the leapfrog integrator preferred over a naive Euler discretization in this context?

4. Consider a one-dimensional HMC system with mass $M = 1$, potential

$$\mathcal{U}(\theta) = \frac{\theta^2}{2},$$

current state $\theta_0 = 1$, momentum $\rho_0 = 0$, step size $\epsilon = 0.1$, and one leapfrog step.

- Compute the updated momentum after the first half-step.
- Compute the updated position.
- Compute the final momentum after the second half-step.

- Compute the Hamiltonian at the start and at the end of the leapfrog step.

5. HMC often mixes faster than a random-walk Metropolis sampler on correlated or high-dimensional targets. Explain why using gradient information can lead to proposals that travel farther while still keeping a high acceptance probability.

6. The two main tuning parameters of HMC are the step size ϵ and the number of leapfrog steps L.

    - What is the effect of choosing ϵ too large?
    - What is the effect of choosing ϵ too small?
    - What can go wrong if L is much too small or much too large?