EC 709: Markov Chain Monte Carlo Methods

Liang Zhong[1]

Boston University

samzl@bu.edu

Novemeber 2023

[1] This slides is based on Guillaume Pouliot's Lecture notes on his Website

Overview

Table of Contents

Motivation: Integration

1. Calculate expectations: $E[f(\theta)]$ with respect to a probability distribution p
   - $\Rightarrow$ $\int f(\theta)p(\theta)d\theta$, but the integral might be intractable or hard to compute
2. Many point estimators are defined as extreme (M-estimators):

$$\hat{\theta} = argmin_\theta \sum_{i=1}^{n} m(\theta, y_i)$$

   - $\Rightarrow$ The objective function $m(\theta, y_i)$ could involve an integral over latent variables
   - e.g., $m(\theta, y_i) = -logp(y_i|\theta) = -log(\int p(y, u|\theta)du)$, that also could be intractable

- Approximating integrals by "sampling instead of summing"

$$\int f(\theta)p(\theta)d\theta \approx \frac{1}{N}\sum_{i=1}^{N} f(\theta_i), \theta_i \sim p$$

⋆ Needs to sample from p

Motivation: Inference

1. Hypothesis testing, the p-value $= P(\text{test statistic} \in \text{Rejection region}|H_0)$:
   - Hard to compute when the distribution of the test statistic is not normal or chi-square

   $\Rightarrow$ $\int 1\{\theta \in \text{Rejection region}\}p(\theta)d\theta$, where θ is the test statistic, and p is the distribution

   $\Rightarrow$ Similar to the integration issue
2. The confidence interval might be hard to compute: the variance formula is too complex

$\Rightarrow$ Consider Bayesian Inference:

E.g., θ is a parameter we want to do inference on, and p is the posterior distribution

$\Leftarrow$ By the Bernstein von Mises theorem, the posterior delivers frequentist large sample inference

$\star$ Also Needs to sample from p

Importance Sampling

- It might be difficult to sample p directly, but we have access to some easy-to-sample q which does not vanish on the support of p
- Then the approximation:

$$\int f(\theta)p(\theta)d\theta \equiv \int f(\theta)\omega(\theta)q(\theta)d\theta \approx \frac{1}{N}\sum_{i=1}^{N} f(\theta_i)\omega(\theta_i), \theta_i \sim q$$

- where $\omega(\theta) = p(\theta)/q(\theta)$, called importance weight, might be more tractable
- Disadvantage: Only worked for the integration issue

The Key question

- Key question: If you have a distribution p which you can only evaluate, i.e. you know $p(\theta)$ for each $\theta \in \Theta$, how can you sample $\theta \sim p$?
  - For a given $\theta \in \Theta$, I can get the real number $p(\theta)$, e.g. $p(9.34) = 0.0124$, but I want samples $X_1, ..., X_n \sim p$
- we want to build a sampler

Review of Metropolis-Hastings (MH)

Algorithm 1 Metropolis-Hastings

```
for i = 0 to N − 1 do
    Sample u ∼ U[0, 1]
    Sample θ* ∼ q(θ* | θ^(i))
    if u < [p(θ*) q(θ^(i) | θ*)] / [p(θ^(i)) q(θ* | θ^(i))] then
        θ^(i+1) = θ*
    else
        θ^(i+1) = θ^(i)
    end if
end for
```

for $i = 0$ to $N - 1$ **do**

  Sample $u \sim U[0, 1]$

  Sample $\theta^* \sim q(\theta^* \mid \theta^{(i)})$

  if $u < \frac{p(\theta^*)q(\theta^{(i)}|\theta^*)}{p(\theta^{(i)})q(\theta^*|\theta^{(i)})}$ **then**

    $\theta^{(i+1)} = \theta^*$

  else

    $\theta^{(i+1)} = \theta^{(i)}$

  end if

end for

Acceptance Rate

- The acceptance rate is:

$$\alpha = min\{\frac{p(\theta^\star)q(\theta^{(i)}|\theta^\star)}{p(\theta^{(i)})q(\theta^\star|\theta^{(i)})}, 1\}$$

- Ratio of p:
  1. If $p(\theta^\star) > p(\theta^{(i)})$, more likely to accept the new draw
  2. If $p(\theta^\star) < p(\theta^{(i)})$, frequency of draw $\theta^\star$ vs keep $\theta^{(i)}$ is proportional to the ratio of their evaluations
- Ratio of q: corrects for the frequency of proposal

$\Rightarrow$ decreases/increases the acceptance probability of values which are overproposed/underproposed.

Main advantage of MCMC in Bayesian Inference

- The acceptance rate is:

$$\alpha = min\{\frac{p(\theta^\star)q(\theta^{(i)}|\theta^\star)}{p(\theta^{(i)})q(\theta^\star|\theta^{(i)})}, 1\}$$

$\star$ When conducting Bayesian inference, the normalization constant is not required:

- Recall: $P(\theta|data) \propto P(data|\theta)\pi(\theta)$
  - $P(\theta|data)$ is the posterior, also the p in MCMC; $P(data|\theta)$ is the likelihood; $\pi(\theta)$ is the prior
  - Sometimes the normalizing constant $P(data)$ is hard to calculate, and it is canceled out in $\frac{P(\theta^\star|data)}{P(\theta^{(i)}|data)}$

  ! If the prior is uniform, can sampling directly from the likelihood

q is very Important

- The acceptance rate is:

$$\alpha = min\{\frac{p(\theta^\star)q(\theta^{(i)}|\theta^\star)}{p(\theta^{(i)})q(\theta^\star|\theta^{(i)})}, 1\}$$

- q needs to be chosen properly
  - It won't be informative if you always reject
  - Extreme case, if $q = p$, we always accept
  - Evaluate q is computationally costly, so want a symmetric proposal $q(\theta|\theta') = q(\theta'|\theta)$

  $\Rightarrow$ it vanishes from the acceptance proposal

- How to choose q?

Choice of the proposal distribution q

- People often called it "more art than science"
- In particular, for common proposals of the form $q(\theta^\star|\theta)$ proposing symmetrically around the mean θ
  - large variance means exploring more, but a lot of rejection
  - ⋆ low variance means a lot of acceptance but very little exploration, i.e., high autocorrelation and little information gain

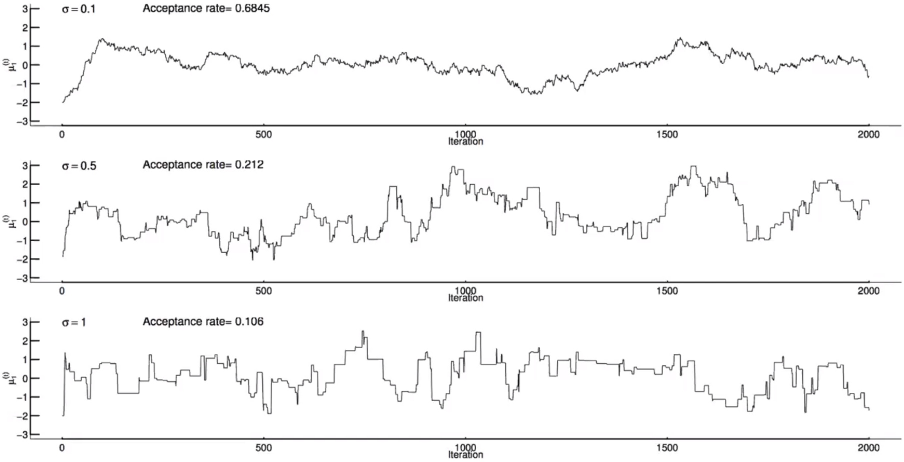
  E.g., picking the σ^2 in $q(\theta^\star|\theta) = N(\theta, \sigma^2)$
- What would be an appropriate number σ?

A illustration Example

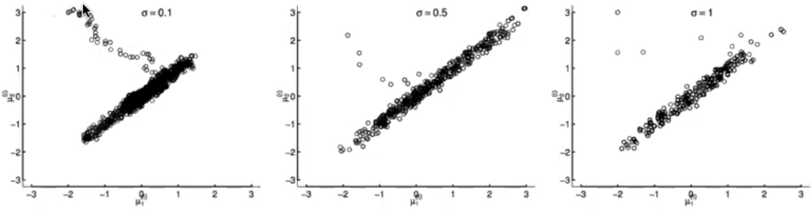
- Sample from Bivariate Normal Distribution:
  - $Y = (Y_1, Y_2)\prime \sim N(0, \Sigma)$
  - $corr(Y_1, Y_2) = 0.99$
- Proposal distribution:

$$q(Y, Y') \sim exp(-\frac{1}{2\sigma^2}|Y - Y'|^2)$$

Trace plot with different σ



Burn-in period with different σ



- Only the moderate variance performs the best
- Roberts, Gelman, and Gilks (1997) analyzes a stylized example and show the optimal acceptance rate in the model is about 0.234

Table of Contents

What if it is hard to evaluate P?

- MCMC was designed for the case that we can evaluate $p(\theta)$
- Many applications doesn't even able to evaluate $p(\theta)$, or we don't care about the whole sample
  - What if we only care about the mode $argmax_{\theta} p(\theta)$
  - What if we only have access to an unbiased estimate $\hat{p}$ of p
  - What if we can generate synthetic data from the parametrized probability model of interest, but cannot write down the likelihood
- Any modification of Metropolis-Hastings to accommodate all those situations to maintain the core idea?

Optimization

- What if we only care about a point estimate

$$\theta_{max} = argmax_\theta p(\theta)$$

  - In practice we often minimizing some objective function $g(\theta) \geq 0$
  - can be converted into the maximum of the probability above as $p(\theta) \propto exp(-g(\theta))$

- Can grid search, take a look at $\{\theta_1, ..., \theta_M\}$ to calculate $\{p(\theta_1), ..., p(\theta_M)\}$ and pick the max

  - spends a lot of time in low-density regions

- We would like to keep the chain close to the optimum, How about concentrating the distribution gradually?

$\Leftarrow$ "Exaggerate" the optimum once we are confident the chain is not "too far" from the maximum, i.e., once we have plausibly reached stationary, and do so gradually

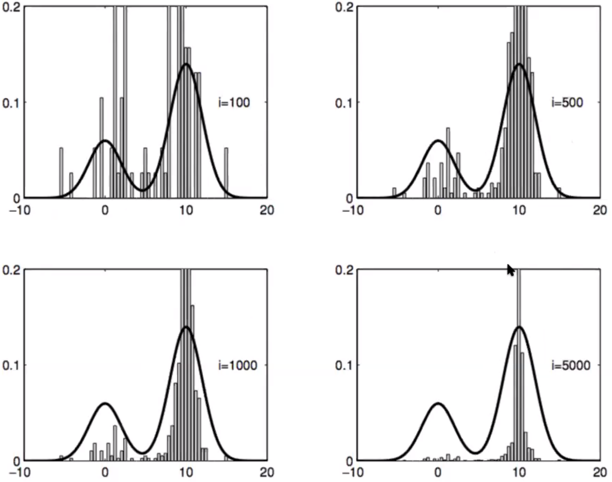
Simulated Annealing

Algorithm 4 Simulated Annealing

for $i = 0$ to $N - 1$ **do**
  Sample $u \sim U[0, 1]$
  Sample $\theta^* \sim q(\theta^* \mid \theta^i)$
  if $u < \frac{p^{1/T_i}(\theta^*)q(\theta^{(i)}|\theta^*)}{p^{1/T_i}(\theta^{(i)})q(\theta^*|\theta^{(i)})}$ **then**
    $\theta^{(i+1)} = \theta^*$
  else
    $\theta^{(i+1)} = \theta^{(i)}$
  end if
end for

Note that the target, $p_i(\theta) \propto p^{1/T_i}(\theta)$, which gradually concentrates around its optimum as $i \to \infty$ and $T_i \to 0$ (called cooling schedule, often use $1/log(i)$)

Performance of Simulated Annealing



In general, no guarantee of achieving global convergence. (Some convergence results for delicate chosen T_i, but converge slower than-grid search)

Can only approximate the target

- Sometimes, evaluating the posterior or likelihood corresponding to our economic model will require evaluating an expensive numerical integral

E.g., For the $exp(-g(\theta))$ in the above example, $g(\theta)$ itself requires integration, and maybe we still want the whole distribution rather than the mode

$\Rightarrow$ Target distribution evaluated by importance sampling

  - Impossible to integrate exactly
  - Easy to propose an unbiased estimate

- More generally, given any proposed $\theta^\star$, you don't have $\pi(\theta^\star)$, but you do have $\hat{\pi}(\theta^\star) \geq 0$ and a guarantee that

$$E(\hat{\pi}(\theta^\star)) = \pi(\theta^\star)$$

Two options available

1. Intuitively, if the estimate $\hat{\pi}(\theta^{\star})$ is very accurate, the resulting draws should approximate draws from the target distribution

$\Leftarrow$ This approach is called the Markov Chain Within Metropolis (MCWM)

- If the approximate is not good, can we still obtain an exact sampling? Yes!

2. Idea: treat draws of the unbiased estimate as auxiliary variables in an M-H algorithm

$\Leftarrow$ Captures the uncertainty in the posterior evaluation and draw from the joint distribution

$\Leftarrow$ Treat the posterior as the (pseudo) marginal of a (pseudo) joint distribution

Pseudo MCMC

Algorithm 5 Pseudo-Marginal MCMC

```
for k = 1 to K do
    Sample u ~ U[0, 1]
    Sample θ* ~ q(θ* | θ^(k))
    Sample π̂*, an unbiased estimate of π(θ*)
    if u < [π̂* q(θ^(k) | θ*)] / [π̂^(k) q(θ* | θ^(k))] then
        θ^(k+1) = θ* and π̂^(k+1) = π̂*
    else
        θ^(k+1) = θ^(k) and π̂^(k+1) = π̂^(k)
    end if
end for
```

Why is it drawing from the true π? Let's look at the "acceptance ratio" to see what are we sampling from

Decompose the acceptance ratio

- Denote $\omega^{(k)} = \frac{\hat{\pi}^{(k)}}{\pi(\theta^{(k)})}$ as an auxiliary variable

$\Rightarrow$ $\frac{\hat{\pi}^\star q(\theta^{(k)}|\theta^\star)}{\hat{\pi}^{(k)} q(\theta^\star|\theta^{(k)})}$

$$= \frac{\frac{\hat{\pi}^\star}{\pi(\theta^\star)} \pi(\theta^\star) q(\theta^{(k)}|\theta^\star)}{\frac{\hat{\pi}^{(k)}}{\pi(\theta^{(k)})} \pi(\theta^{(k)}) q(\theta^\star|\theta^{(k)}))} = \frac{\omega^\star \pi(\theta^\star) q(\theta^{(k)}|\theta^\star)}{\omega^{(k)} \pi(\theta^{(k)}) q(\theta^\star|\theta^{(k)})}$$

$$= \frac{\omega^\star \pi(\theta^\star) p(\omega^\star|\theta^\star)}{\omega^{(k)} \pi(\theta^{(k)}) p(\omega^{(k)}|\theta^{(k)})} \times \frac{p(\omega^{(k)}|\theta^{(k)}) q(\theta^{(k)}|\theta^\star)}{p(\omega^\star|\theta^\star) q(\theta^\star|\theta^{(k)})}$$

- Can recognize acceptance ratio for (ω, θ) with proposal $p(\omega^\star|\theta^\star) q(\theta^\star|\theta^{(k)})$ and target $\omega^\star \pi(\theta^\star) p(\omega^\star|\theta^\star)$

About the target distribution

- In practice, want to draw from the θ marginal of the (ω, θ) joint
- How does one do that in practice? Just ignore the ω
- What are we sampling from when we marginalize ω

$$\int_{\omega} \omega\pi(\theta)p(\omega|\theta)d\omega = \pi(\theta)\int_{\omega} \omega p(\omega|\theta)d\omega = \pi(\theta)E(\frac{\hat{\pi}(\theta)}{\pi(\theta)}) = \pi(\theta)$$

- The sampling is exact!

Only access generative model

- In economics, we often have complicated model
  - The likelihood would be involved
  - Easy to simulate but hard to write down the closed form

E.g., Structural economics models sometimes explicitly model agents as sequentially taking decisions, as well as the distribution of innovations

  - It can be quite difficult, or impossible, to work out their likelihood, let alone evaluate them
  - However, it can be very easy to generate from them once you have fixed the parameters

- Specially, consider test statistic/data Y generated from a generative model g parameterized in θ and taking as argument a random element z with known distribution:

$$Y_\theta = g(\theta, z), \theta \in \Theta, z \sim F_z$$

- Idea: If we generat Y_θ for θ close to the true parameter θ_0, then Y_θ and observed Y should be close to each other since $Y \sim Y_{\theta_0}$

Approximate Bayesian Computation (ABC)

Algorithm 6 ABC

for $i = 0$ to $N-1$ **do**
  Sample $u \sim U[0,1]$
  Sample $\theta^* \sim q(\theta^* \mid \theta^i)$
  Sample $z \sim F_z$
  Compute $Y_{\theta^*} = g(\theta^*, z)$
  if $u < \mathbf{1}\left\{d\left(\hat{Y}, Y_{\theta^*}\right) < \epsilon\right\} \frac{q(\theta^{(i)}|\theta^*)}{q(\theta^*|\theta^{(i)})}$ **then**
    $\theta^{(i+1)} = \theta^*$
  else
    $\theta^{(i+1)} = \theta^{(i)}$
  end if
end for

- This is the Bayesian equivalent of indirect inference

⇐ JJ has a nice paper to illustrate their connections

Technical Note on ABC

- Need to specify a distance between the true data Y and synthetic data Y_θ
- Same distances as in indirect inference can be used, e.g. difference between moments as in the simulated method of moments for p moments $m_1, ..., m_p$:

$$d(Y, Y_\theta) = |m_1(Y) - m_1(Y_\theta), ..., m_p(Y) - m_p(Y_\theta)|_2$$

- As in indirect inference, choosing moments/ more general auxiliary model/ pseudolikelihood can be difficult
- Would prefer other nonparametric approaches (wouldn't go into details)
  - Bernton et al. (2017) use Wasserstein distance as d

  $\Leftarrow$ Combines adaptive proposal and a shrinking ϵ
  - Kaji et al. (2020) use a neural network classifier as d

Other questions for MCMC

1. Rarely precise guidelines for practitioners

   E.g., How to choose the stopping criteria?

   $\Leftarrow$ Involves distance between probabilities, seems much harder than in optimization

   $\Leftarrow$ The number of iterations reported in the literature spans many orders of magnitude (dozens, millions, trillions)

2. Can we use parallelization to boost the speed?

$\Leftarrow$ Since MCMC methods are iterative, they are not obvious to parallelize

3. How to construct the unbiased MCMC estimator without concerned about the burn-in periods?

- Most of them can be (partially) solved by introducing "coupling"
- Take a look at Pierre E. Jacob's Website if you are interested in some recent advance on these topics

Thank You!