MCMC Convergence Diagnostics - Mathematical Formulation and Hand Calculations

1. Introduction to MCMC Convergence

MCMC algorithms generate samples from a target distribution, but we need to ensure the chains have converged to the stationary distribution. Key diagnostics include:

1. **R-hat (Potential Scale Reduction Factor)**: Compares between-chain and within-chain variance
2. **Effective Sample Size (ESS)**: Accounts for autocorrelation in the chain

2. R-hat Statistic

The R-hat statistic is defined as:

$$\hat{R}^2 = (S-1)/S + (1/S)(B/W)$$

where:

- S = number of samples per chain (after warmup)
- B = between-chain variance
- W = within-chain variance

Computing B and W:

Within-chain variance:

$$W = (1/M) \sum_{i=1}^{M} s^2_i$$

where s^2_i is the variance of chain i:

$$s^2_i = (1/(S-1)) \sum_{j=1}^{S} (\theta_{ij} - \bar{\theta_i})^2$$

Between-chain variance:

$$B = (S/(M-1)) \sum_{i=1}^{M} (\bar{\theta_i} - \bar{\theta})^2$$

where:

- M = number of chains

- θ_{ij} = j-th sample from chain i
- $\bar{\theta}_i$ = mean of chain i
- $\bar{\theta}$ = overall mean across all chains

Interpretation:

- $\hat{R}$ = 1: Chains have converged
- $\hat{R}$ > 1: Chains have not fully converged
- Typical threshold: $\hat{R}$ < 1.1 (or < 1.01 for critical applications)

3. Effective Sample Size

The ESS accounts for autocorrelation:

```
ESS = S / (1 + 2Σₜ₌₁^∞ ρₜ)
```

where ρ_t is the autocorrelation at lag t:

```
ρₜ = Cor(θᵢ, θᵢ₊ₜ)
```

Computing autocorrelation:

```
ρₜ = (1/(S-t)) Σᵢ₌₁^(S-t) (θᵢ - θ̄)(θᵢ₊ₜ - θ̄) / σ²
```

4. Hand Calculation Example

Let's compute $\hat{R}$ and ESS for a simple example with:

- M = 2 chains
- S = 5 samples per chain
- Single parameter θ

Data:

```
Chain 1: [1.2, 1.4, 1.1, 1.3, 1.2]
Chain 2: [1.8, 1.6, 1.9, 1.7, 1.5]
```

Step 1: Compute chain means

$$
\begin{aligned}
\bar{\theta}_1 &= (1.2 + 1.4 + 1.1 + 1.3 + 1.2) / 5 = 1.24 \\
\bar{\theta}_2 &= (1.8 + 1.6 + 1.9 + 1.7 + 1.5) / 5 = 1.70 \\
\bar{\theta} &= (1.24 + 1.70) / 2 = 1.47
\end{aligned}
$$

Step 2: Compute within-chain variances

$$
\begin{aligned}
s^2_1 &= [(1.2-1.24)^2 + (1.4-1.24)^2 + (1.1-1.24)^2 + (1.3-1.24)^2 + (1.2-1.24)^2] / 4 \\
&= [0.0016 + 0.0256 + 0.0196 + 0.0036 + 0.0016] / 4 \\
&= 0.013
\end{aligned}
$$

$$
\begin{aligned}
s^2_2 &= [(1.8-1.70)^2 + (1.6-1.70)^2 + (1.9-1.70)^2 + (1.7-1.70)^2 + (1.5-1.70)^2] / 4 \\
&= [0.01 + 0.01 + 0.04 + 0.00 + 0.04] / 4 \\
&= 0.0225
\end{aligned}
$$

Step 3: Compute W and B

$$
W = (s^2_1 + s^2_2) / 2 = (0.013 + 0.0225) / 2 = 0.01775
$$

$$
\begin{aligned}
B &= S/(M-1) \times \Sigma_i(\bar{\theta}_i - \bar{\theta})^2 \\
&= 5/1 \times [(1.24-1.47)^2 + (1.70-1.47)^2] \\
&= 5 \times [0.0529 + 0.0529] \\
&= 5 \times 0.1058 \\
&= 0.529
\end{aligned}
$$

Step 4: Compute R̂

$$
\begin{aligned}
\hat{R}^2 &= (S-1)/S + (1/S)(B/W) \\
&= 4/5 + (1/5)(0.529/0.01775) \\
&= 0.8 + 0.2 \times 29.8 \\
&= 0.8 + 5.96 \\
&= 6.76
\end{aligned}
$$

$$
\hat{R} = \sqrt{6.76} = 2.60
$$

This indicates poor convergence ($\hat{R} > 1.1$).

Step 5: Compute ESS (for one chain)

For Chain 1, compute autocorrelations:

```
Lag 0: ρ₀ = 1.0 (always)
Lag 1: ρ₁ = [(1.2-1.24)(1.4-1.24) + (1.4-1.24)(1.1-1.24) + ...] / (4 × 0.013)
```

Simplified example with $\rho_1 = 0.3$, $\rho_2 = 0.1$, $\rho_3 = 0$:

```
ESS = 5 / (1 + 2(0.3 + 0.1))
    = 5 / (1 + 0.8)
    = 5 / 1.8
    = 2.78
```

This means our 5 samples have the statistical power of about 2.78 independent samples.

5. Bimodal Distribution Example

For the mixture distribution:

```
p(x) = 0.5 × N(x|-3, 4) + 0.5 × N(x|1, 2)
```

The theoretical moments are:

```
E[x] = 0.5 × (-3) + 0.5 × 1 = -1
E[x²] = 0.5 × (9 + 4) + 0.5 × (1 + 2) = 8
Var[x] = E[x²] - (E[x])² = 8 - 1 = 7
```

With multiple chains, we can assess whether all chains explore both modes.

6. Change Point Detection Model

Model Structure:

```
c ~ Uniform(1, N)
λ₁ ~ Gamma(α, β)
λ₂ ~ Gamma(α, β)
xᵢ ~ Poisson(λ₁) if i ≤ c
xᵢ ~ Poisson(λ₂) if i > c
```

Log Joint Distribution:

```
log p(x, λ₁, λ₂, c) = Σᵢ₌₁ᶜ log p(xᵢ|λ₁) + Σᵢ₌ᶜ₊₁ᴺ log p(xᵢ|λ₂)
                    + log p(λ₁) + log p(λ₂) + log p(c)
```

Gibbs Sampling Updates:

Update λ_1:

$$\lambda_1 \mid x, \lambda_2, c \sim \text{Gamma}(\alpha + \Sigma_{i=1}^{c} x_i, \beta + c)$$

Update λ_2:

$$\lambda_2 \mid x, \lambda_1, c \sim \text{Gamma}(\alpha + \Sigma_{i=c+1}^{N} x_i, \beta + N - c)$$

Update c:

$$p(c \mid x, \lambda_1, \lambda_2) \propto \exp[\Sigma_{i=1}^{c} x_i \log(\lambda_1) - c\times\lambda_1 + \Sigma_{i=c+1}^{N} x_i \log(\lambda_2) - (N-c)\times\lambda_2]$$

Small Example:

Data: [3, 2, 1, 5, 6, 7] ($N = 6$) Parameters: $\alpha = 1$, $\beta = 1$

Initial values:

- $c^{(0)} = 3$
- $\lambda_1^{(0)} = 2$
- $\lambda_2^{(0)} = 5$

Iteration 1:

1. Update λ_1:

   $$\alpha' = 1 + (3 + 2 + 1) = 7$$
   $$\beta' = 1 + 3 = 4$$
   $$\lambda_1^{(1)} \sim \text{Gamma}(7, 4)$$

2. Update λ_2:

   $$\alpha' = 1 + (5 + 6 + 7) = 19$$
   $$\beta' = 1 + 3 = 4$$
   $$\lambda_2^{(1)} \sim \text{Gamma}(19, 4)$$

3. Update c: Compute probabilities for each possible c and sample.

This process continues until convergence, assessed using $\hat{R}$ and ESS.