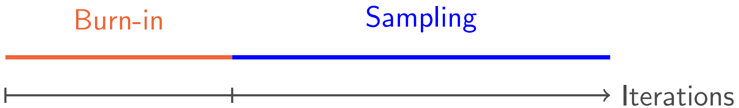
# The Convergence Challenge in MCMC

- **Ideal goal**: Assess whether MCMC chains have converged
- **Fundamental problem**:
  - In general, impossible to know for sure that there is no problem
  - But we can sometimes know for sure that there *is* a problem
- **Two phases of MCMC**:
  - Transient phase (burn-in): mixing time
  - Stationary phase: Monte Carlo estimation

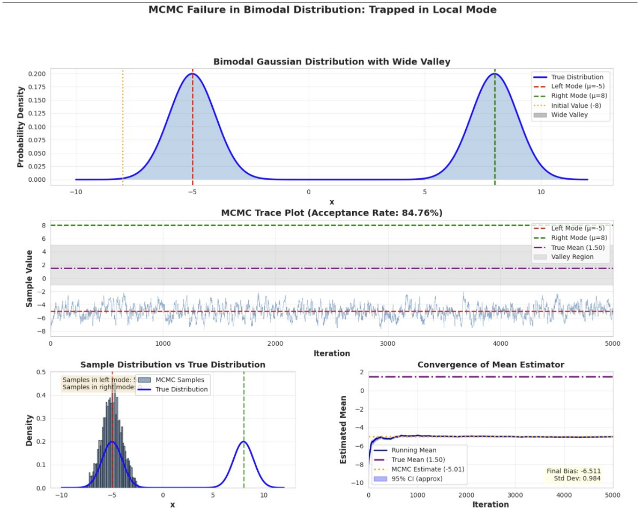


# Why Convergence Matters

**Non-converged chains:**

- Biased estimates
- Incorrect uncertainty quantification
- Missing important modes
- Unreliable inference

# Motivating example



# The Intuition Behind Gelman-Rubin

**Core Idea**

If MCMC chains have converged to the target distribution, then:

- Multiple chains started from different points should look similar
- Within-chain variance $\approx$ Between-chain variance

**Compare two sources of variance:**

1. Within-chain variance (W)
   How much each chain varies
2. Between-chain variance (B)
   How different chains are from each other

# Within-chain variance - $W$

Run $M$ chains. The sample mean of $M$ sample variances

$$W = \frac{1}{M}\sum_{m=1}^{M}\left[\frac{1}{T-1}\sum_{t=1}^{T}(X_{m,t}-\bar{X}_m)^2\right]$$

We have that the expected sample variance for one chain is

$$\mathbb{E}\left[\frac{1}{T-1}\sum_{t=1}^{T}(X_{m,t}-\bar{X}_m)^2\right] = \frac{T}{T-1}\left(\sigma^2 - Var(\bar{X}_{m,.})\right)$$

making the estimator unbiased only in the case $Var(\bar{X}_{m,.}) = \sigma^2/T$ (iid samples).
For MCMC samples, $Var(\bar{X}_{m,.})$ is typically larger than $\sigma^2/T$ due to autocorrelation, so $W$ underestimates $\sigma^2$.

# Between-chain variance - $B$

For the $M$ chains, we compute the variance of the chain means:

$$B = \frac{1}{M-1} \sum_{m=1}^{M} (\bar{X}_{m,.} - \bar{X}_{..})^2$$

where $\bar{X}_{..}$ is the mean across all chains. We have that

$$\mathbb{E}[B] = Var(\bar{X}_{m,.})$$

# Estimators for Target Variance

We have 2 estimators for the target variance $\sigma^2$:

$$W = \frac{1}{M}\sum_{m=1}^{M}\left[\frac{1}{T-1}\sum_{t=1}^{T}(X_{m,t} - \bar{X}_m)^2\right]$$

and

$$V = \frac{T-1}{T}W + B = \left(1 - \frac{1}{T}\right)W + B$$

$V$ weights the within-chain variance $W$ heavily when you have many samples, but adds between-chain variance $B$ to account for the fact that chains might not be fully mixed yet.

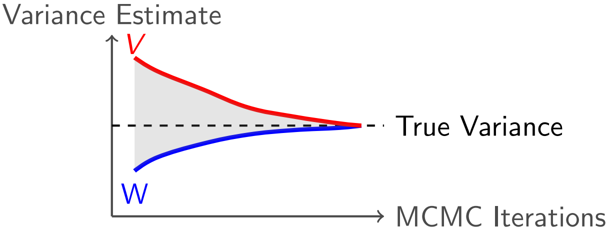
In case we start the chains from overdispersed initial values, we expect $B$ to be large, since chain means $\bar{X}_{m,.}$ are more spread out than they should be. Thus: $B$ overestimates $Var(\bar{X}_{m,.})$. This makes $V$ overestimate $\sigma^2$.

# The Gelman-Rubin Statistic

**Definition**

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

- Original recommendation: $\hat{R} < 1.1$ for convergence.
- More recent advice: $\hat{R} < 1.01$ (Vehtari et al., 2021)
- But what does $\hat{R}$ really mean?



# Connection to Effective Sample Size

## Key Approximation (Vats & Knudson, 2021)

$$\hat{R} \approx \sqrt{1 + \frac{M}{\text{ESS}}}$$

Where:

- $M$ = number of chains
- ESS = number of independent samples with the same standard error as a correlated sample.

**Implications:**

- $\hat{R} = 1.1 \Rightarrow \text{ESS} \approx 5M$ (5 independent samples per chain)
- $\hat{R} = 1.01 \Rightarrow \text{ESS} \approx 50M$ (50 independent samples per chain)

5 effective samples per chain is too small for reliable inference!

# Weaknesses of Gelman-Rubin

1. **Only detects lack of convergence**
2. **Cannot detect if all modes are found**
   - Only checks if chains agree with each other
   - All chains might miss the same modes
3. **Sensitive to initialization**
   - Chains starting in the same wrong place

# Convergence Assessment

### Use Multiple Diagnostics

1. **Gelman-Rubin statistic**: $\hat{R} < 1.01$
2. **Effective Sample Size**
3. **Trace plots**: Visual inspection
4. **Autocorrelation**: Check mixing quality

**Best Practices:**

- Use at least 4 chains (preferably more)
- Initialize chains from overdispersed starting points
- Run chains longer than you think necessary