Probabilistic Methods (PM - 330725)

TOPIC 4: Monte Carlo Methods

Lecture 8

May 2025

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Escola Politècnica Superior d'Enginyeria de Manresa

Roadmap

Learning Outcomes

By the end of this topic, you will be able to:

- Be able to develop simulations using random sampling to estimate integrals and solve probabilistic problems.
- Be able to use Monte Carlo techniques to approximate complex integrals and understand their convergence properties.
- Be able to employ Monte Carlo methods for optimization problems, including finding global minima and maxima.
- Be able to apply simulations to model and analyze uncertainty in various real-world contexts.

Monte Carlo sampling

- Monte Carlo methods are a broad class of computational algorithms that rely on repeated random sampling to obtain numerical results.
- The underlying concept behind these methods is the use of randomness for solving problems that might in principle be deterministic. Monte Carlo methods are often used in physical and mathematical problems.
- it has a few distinct advantages in cases where it would be difficult or even impossible to use alternative approaches.

Monte Carlo sampling: Intro

- If $\theta^{(1)}, \ldots, \theta^{(m)}$ is an iid sequence from $p(\theta \mid y)$, then

$$\bar{\theta} = \frac{1}{m} \sum_{i=1}^{m} \theta^{(i)} \longrightarrow \mathbb{E}[\theta \mid y],$$

$$\bar{g}(\theta) = \frac{1}{m} \sum_{i=1}^{m} g\big(\theta^{(i)}\big) \longrightarrow \mathbb{E}\big[g(\theta) \mid y\big],$$

  for some function $g(\theta)$ of interest.

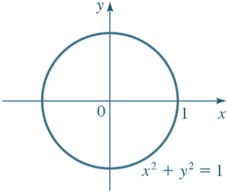
- Central limit theorem

$$\bar{\theta}_{1:m} \approx \mathcal{N}\left(\mathbb{E}[\theta \mid y], \frac{Var(\theta|y)}{m}\right) \quad \text{for large } m.$$

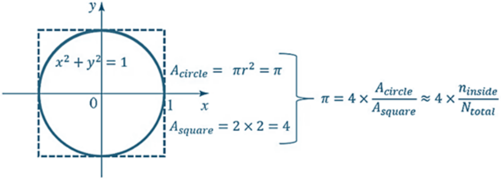
Example: Estimating the value of PI using a random number

The task is to find the value of PI (π)

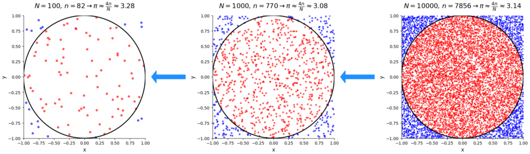
Let's start out with the definition of the unit circle (i.e. a circle of radius 1)



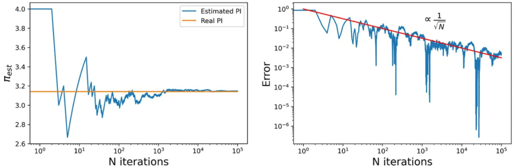
This means that if we start out with two random numbers in the range [-1,1],



Example: Estimating the value of PI using a random number

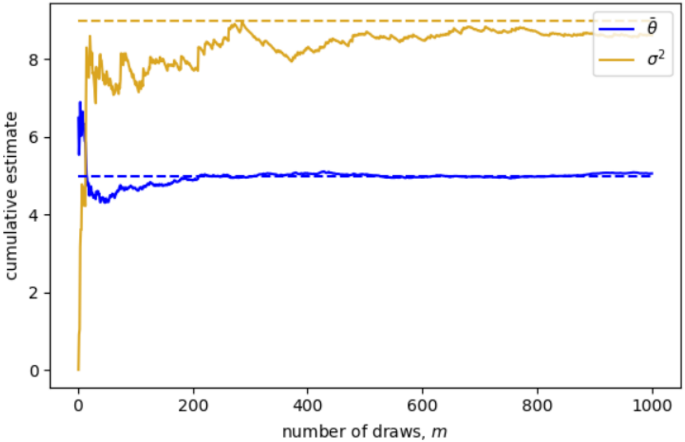


By generating N=100.000 random pairs of x,y coordinates, we can then check how our approximation of PI improves with the number of iterations also, note the logarithmic x-axis



(See **Examples_Lecture8.ipynb**)

Monte Carlo sampling - convergence



- $\bar{\theta}$ (blue) and σ^2 (gold) converge to their true values

(See **Example_Lecture3.ipynb**)

Direct Sampling (Direct Inversion of CDFs)

Direct Solution

$$\hat{x} \leftarrow F^{-1}(\xi)$$

Sampling Procedure:

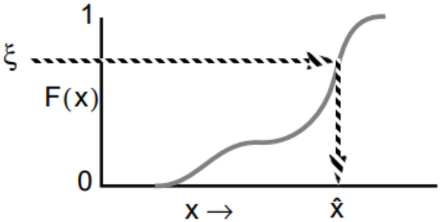
- Generate $\xi \sim U(0, 1)$
- Determine $\hat{x}$ such that $F(\hat{x}) = \xi$

Advantages

- Straightforward mathematics & coding
- "High-level"approach

Disadvantages

- Often involves complicated functions
- In some cases, $F(x)$ cannot be inverted (e.g., Klein–Nishina formulae)



Use when the inverse CDF is costly or impossible. Choose a bounding density $g(x)$ and constant c such that

·

$$c \cdot g(x) \geq f(x) \text{for all x,}$$

·

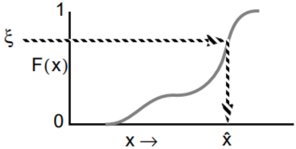
$$g(x) \text{ is easy to sample PDF.}$$

Sampling Procedure: · Sample $\hat{x}$ from $g(x)$:
· $\hat{x} \leftarrow G^{-1}(\xi_1)$.

Draw $\xi_2 \sim U(0, 1)$, test:

$$\xi_2 \, c \, g(\hat{x}) \leq f(\hat{x}).$$

- If true, accept $\hat{x}$, done.
- If false, reject $\hat{x}$ and repeat.



Advantages

- Simple computer operations

Disadvantages

- Low Level Approach, sometimes hard to reach it and understand

Gibbs sampling

- **Sampling from multivariate distributions**, $p(X_1, \ldots, X_p)$.
- Typically a posterior distribution: $p(\theta_1, \ldots, \theta_p \mid y)$.
- Requirement: Easily sampled full conditional distributions:
  - $p(\theta_1 \mid \theta_2, \theta_3, \ldots, \theta_p, y)$
  - $p(\theta_2 \mid \theta_1, \theta_3, \ldots, \theta_p, y)$
  - $\vdots$
  - $p(\theta_p \mid \theta_1, \theta_2, \ldots, \theta_{p-1}, y)$
- Gibbs sampling is a special case of **Metropolis–Hastings**.
- Metropolis-Hastings is a Markov Chain Monte Carlo (MCMC) algorithm.

The Gibbs sampling algorithm

Algorithm 1 Gibbs sampling

Require: Initial values $\theta_2^{(0)}, \ldots, \theta_p^{(0)}$, number of draws m

for $i = 1$ to m **do**

$$\theta_1^{(i)} \sim p\big(\theta_1 \mid \theta_2^{(i-1)}, \theta_3^{(i-1)}, \ldots, \theta_p^{(i-1)}, y\big)$$

$$\theta_2^{(i)} \sim p\big(\theta_2 \mid \theta_1^{(i)}, \theta_3^{(i-1)}, \ldots, \theta_p^{(i-1)}, y\big)$$

$$\vdots$$

$$\theta_p^{(i)} \sim p\big(\theta_p \mid \theta_1^{(i)}, \theta_2^{(i)}, \ldots, \theta_{p-1}^{(i)}, y\big)$$

end for

Output: m (autocorrelated) draws $\{\theta^{(i)}\}$ converging in distribution to the joint posterior $p(\theta_1, \ldots, \theta_p \mid y)$

Gibbs sampling draws converge to the posterior

- Gibbs draws $\theta^{(1)}, \ldots, \theta^{(m)}$ are dependent, but

$$\bar{\theta} = \frac{1}{m}\sum_{t=1}^{m}\theta^{(t)} \longrightarrow \mathbb{E}[\theta \mid y],$$

$$\bar{g}(\theta) = \frac{1}{m}\sum_{t=1}^{m} g\left(\theta^{(t)}\right) \longrightarrow \mathbb{E}\left[g(\theta) \mid y\right].$$

- $\theta^{(1)}, \ldots, \theta^{(m)}$ **converges in distribution** to the joint posterior $p(\theta \mid y)$.
- For each component j, the subsequence $\theta_j^{(1)}, \ldots, \theta_j^{(m)}$ converges to the marginal posterior of θ_j.
- **Central limit theorem:**

$$\bar{\theta}_{1:m} \approx \mathcal{N}\left(\mathbb{E}[\theta \mid y],\ Var(\bar{\theta})\right) \quad \text{for large } m.$$

Direct sampling vs Gibbs sampling

- Dependent draws $\longrightarrow$ less efficient than iid sampling.
- iid samples:

$$\mathrm{Var}(\bar{\theta}) = \frac{\sigma^2}{m}, \quad \sigma^2 = \mathrm{Var}(\theta \mid y).$$

- Autocorrelated samples:

$$\mathrm{Var}(\bar{\theta}) = \frac{\sigma^2}{m}\left(1 + 2\sum_{k=1}^{\infty} \rho_k\right),$$

  where ρ_k is the autocorrelation at lag k.
- Inefficiency factor:

$$\mathrm{IF} = 1 + 2\sum_{k=1}^{\infty} \rho_k \approx 1 + 2\sum_{k=1}^{K} \rho_k.$$

- Effective sample size (ESS):

$$\mathrm{ESS} = \frac{m}{\mathrm{IF}}.$$

Gibbs sampling from bivariate normal

Joint distribution

$$\begin{pmatrix}\theta_1\\ \theta_2\end{pmatrix} \sim \mathcal{N}_2\left(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix},\ \begin{pmatrix}\sigma_1^2 & \rho\,\sigma_1\sigma_2\\ \rho\,\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}\right)$$

Algorithm 2 Gibbs sampling from a bivariate normal

Require: Initial value $\theta_2^{(0)}$, number of draws m

1: **for** $i = 1$ to m **do**

2: $\quad \theta_1^{(i)} \sim \mathcal{N}\left(\mu_1 + \rho\dfrac{\sigma_1}{\sigma_2}(\theta_2^{(i-1)} - \mu_2),\ \sigma_1^2(1-\rho)^2\right)$

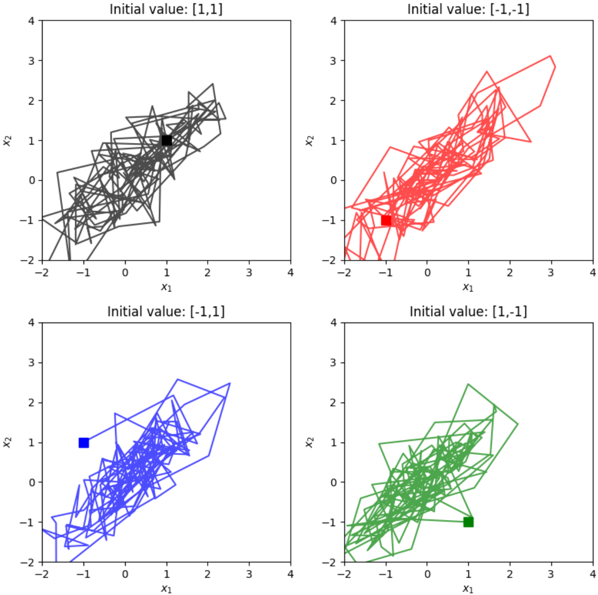
3: $\quad \theta_2^{(i)} \sim \mathcal{N}\left(\mu_2 + \rho\dfrac{\sigma_2}{\sigma_1}(\theta_1^{(i)} - \mu_1),\ \sigma_2^2(1-\rho)^2\right)$

4: **end for**

Ensure: m (autocorrelated) draws $\{\theta^{(i)}\}$ converging to $\mathcal{N}_2(\mu, \Sigma)$, where

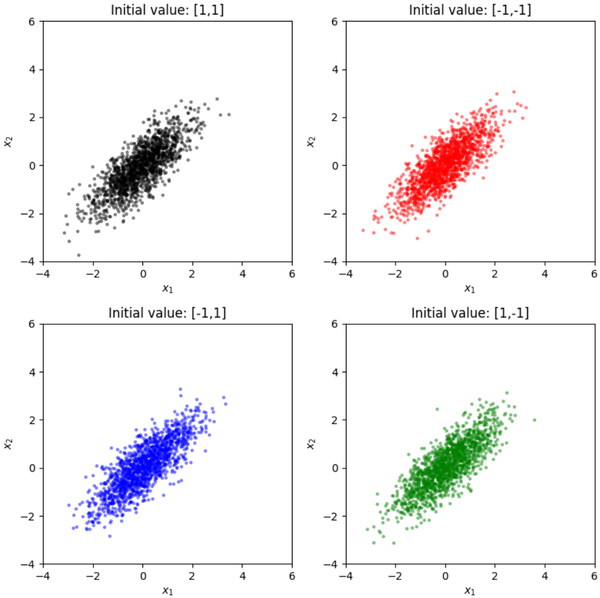
$$\mu = \begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \quad \Sigma = \begin{pmatrix}\sigma_1^2 & \rho\,\sigma_1\sigma_2\\ \rho\,\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}.$$

Gibbs sampling from bivariate normal

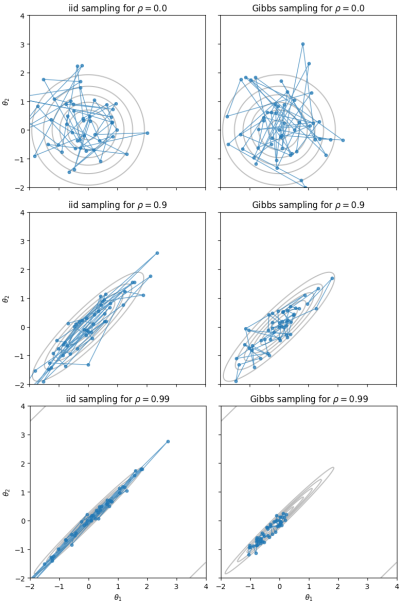


(See **Examples_Lecture8.ipynb**)

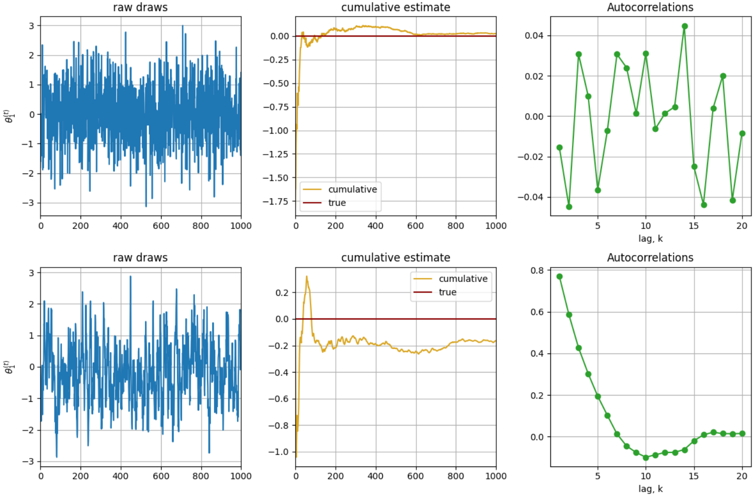
Gibbs sampling from bivariate normal



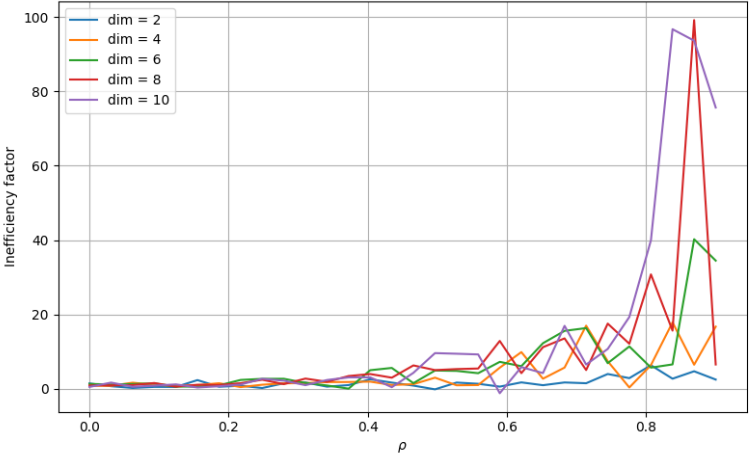
Gibbs sampling from bivariate normal



Direct vs Gibbs sampling, bivariate normal $\rho = 0 : 9$



Gibbs is inefficient when parameters are correlated



- Inefficiency grows rapidly in ρ, especially as dimension increases.
- Reflects how high correlation induces strong autocorrelation in Gibbs chains.

Normal model with conditionally conjugate prior

- **Normal model** with conditionally conjugate prior

$$\mu \sim \mathcal{N}(\mu_0,\ \tau_0^2), \quad \sigma^2 \sim \text{Inv-}\chi^2(\nu_0,\ \sigma_0^2).$$

- **Full conditional posteriors**

$$\mu \mid \sigma^2, x \ \sim\ \mathcal{N}(\mu_n,\ \tau_n^2),$$

$$\sigma^2 \mid \mu, x \ \sim\ \text{Inv-}\chi^2\left(\nu_n,\ \frac{\nu_0\,\sigma_0^2 + \sum_{i=1}^n (x_i - \mu)^2}{n + \nu_0}\right),$$

with $\mu_n,\ \tau_n^2$ defined exactly as in the known-σ^2 case.

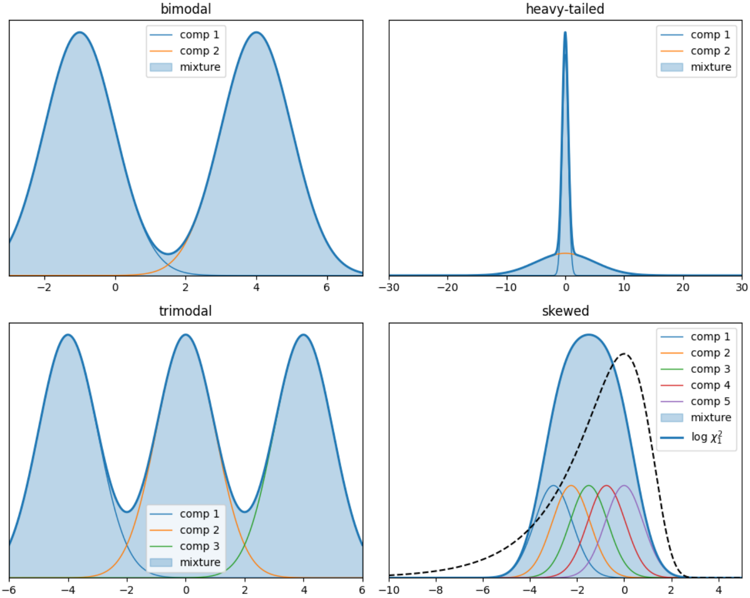
Data augmentation – Mixture distributions

- Let $\mathcal{N}(x \mid \mu, \sigma^2)$ denote the PDF of $x \sim \mathcal{N}(\mu, \sigma^2)$.
- Two-component mixture of normals [MoN(2)]:

$$p(x) = \omega\,\mathcal{N}(x \mid \mu_1, \sigma_1^2) + (1-\omega)\,\mathcal{N}(x \mid \mu_2, \sigma_2^2).$$

- Simulate from a MoN(2):
  - Simulate a membership indicator $Z \in \{1, 2\}$: $Z \sim \mathrm{Bernoulli}(\omega)$.
  - If $Z = 1$, simulate x from $\mathcal{N}(\mu_1, \sigma_1^2)$.
  - If $Z = 2$, simulate x from $\mathcal{N}(\mu_2, \sigma_2^2)$.

Illustration of mixture of normals



- K-component mixture of normals

$$p(x) = \sum_{k=1}^{K} \omega_k \, \mathcal{N}(x \mid \mu_k, \sigma_k^2) \quad \Big(\sum_k \omega_k = 1,\ \omega_k \geq 0\Big).$$

- **Indicators:** $Z_i = k$ if observation x_i comes from component k.

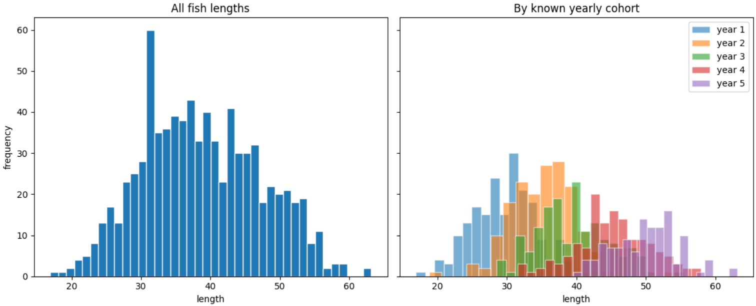
Algorithm 3 Simulating data from a mixture of normals

Require: Number of observations n, weights $\omega_{1:K}$, means $\mu_{1:K}$, variances $\sigma^2_{1:K}$.

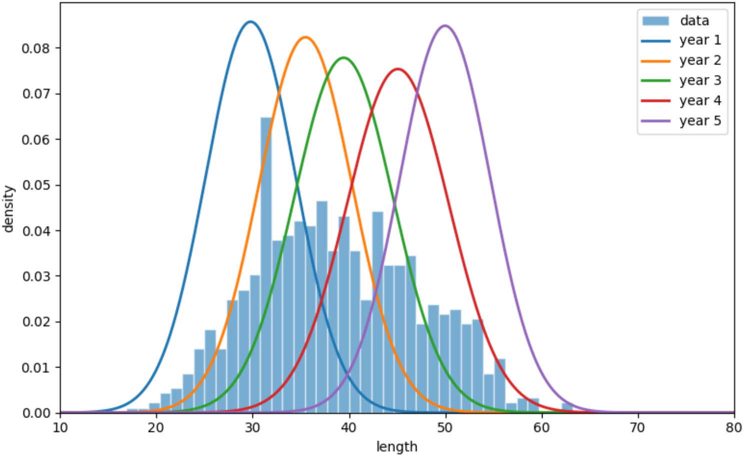
1: **for** $i = 1$ to n **do**
2: Draw component $z_i \sim \mathrm{Categorical}(\omega_1, \ldots, \omega_K)$
3: Draw observation $x_i \mid z_i \sim \mathcal{N}(\mu_{z_i}, \sigma^2_{z_i})$
4: **end for**

Ensure: iid sample $x = (x_1, \ldots, x_n)$ from the mixture

Fish length data with known yearly cohorts



Fish length data - fit with known yearly cohorts



Likelihood for a mixture and data augmentation

- The **likelihood** is a product of sums. Messy to work with.
- **Assume** that we know where each observation comes from:

$$z_i = k \quad \text{if } x_i \text{ came from mixture component } k.$$

- Given $z_1, \ldots, z_n$ it is easy to estimate the means $\mu_1, \ldots, \mu_K$, the variances $\sigma_1^2, \ldots, \sigma_K^2$ and the mixture proportions $\omega_1, \ldots, \omega_K$: just split the data into K groups according to $z_1, \ldots, z_n$.
- But we do not know $z_1, \ldots, z_n$!
- **Data augmentation**: add $z_1, \ldots, z_n$ as latent variables and update them in a separate Gibbs step.

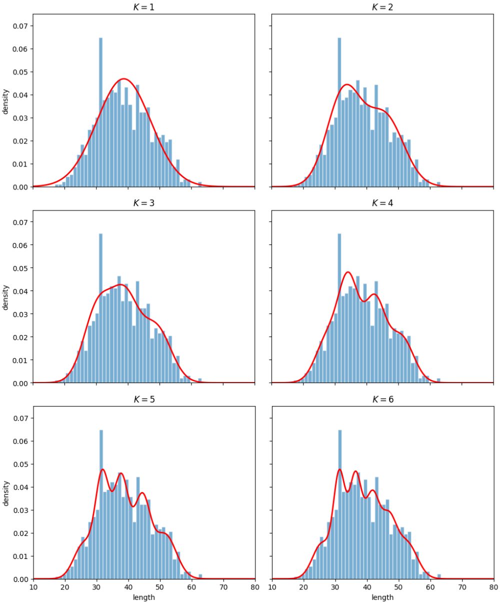
Gibbs sampling for mixture distributions

Algorithm 4 Mixture-of-Normals Gibbs sampler

Require: Observations $x_{1:n}$, number of components K, hyperparameters $(\alpha_{1:K}, \mu_{0,k}, \tau^2_{0,k}, \nu_{0,k}, \sigma^2_{0,k})$

for $j = 1$ to m **do**
    `// Update component parameters`
    for $k = 1$ to K **do**
        Let $x_k = \{ x_i : z_i^{(j-1)} = k \}$
        Draw $(\sigma_k^2)^{(j)} \sim \mathrm{Scaled{-}Inv{-}ffl}^2(\nu_{n,k}, \sigma^2_{n,k})$
        Draw $\mu_k^{(j)} \sim \mathcal{N}(\mu_{n,k},\ \tau^2_{n,k} \mid (\sigma_k^2)^{(j)},\ x_k)$
    end for
    `// Update mixture weights`
    Let $n_k = |x_k|$ for each k
    Draw $\omega^{(j)} \sim \mathrm{Dirichlet}(\alpha_1 + n_1, \ldots, \alpha_K + n_K)$
    `// Update allocation indicators`
    for $i = 1$ to n **do**
        for $k = 1$ to K **do**
            $\tilde{\omega}_k \propto \omega_k^{(j)}\ \mathcal{N}(x_i \mid \mu_k^{(j)}, (\sigma_k^2)^{(j)})$
        end for
        Normalize $\tilde{\omega}_{1:K}$ so they sum to 1
        Draw $z_i^{(j)} \sim \mathrm{Categorical}(\tilde{\omega}_{1:K})$
    end for
end for

Fish length data - mixture of normals fit



Markov chains

- Let $\mathcal{S} = \{s_1, s_2, \ldots, s_k\}$ be a finite set of **states**.
  - Weather: $\mathcal{S} = \{\text{sunny}, \text{rain}\}$
  - School grades: $\mathcal{S} = \{A, B, C, D, E, F\}$
- Markov chain is a stochastic process $\{X_t\}_{t=1}^{T}$ with state transitions

$$p_{ij} = \Pr(X_{t+1} = s_j \mid X_t = s_i)$$

- School grades:

$$X_1 = C,\ X_2 = C,\ X_3 = B,\ X_4 = A,\ X_5 = B$$

- Transition matrix for weather example:

$$P = \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix} = \begin{pmatrix} 0.9 & 0.1 \\ 0.7 & 0.3 \end{pmatrix}$$

Stationary distribution

- *h-step transition probabilities*

$$P_{ij}^{(h)} = \Pr(X_{t+h} = s_j \mid X_t = s_i)$$

- *h-step transition matrix by* **matrix power**

$$P^{(h)} = P^h$$

- **Unique** equilibrium distribution $\pi = (\pi_1, \ldots, \pi_k)$ if chain is
  - irreducible (possible to get to any state from any state)
  - aperiodic (does not get stuck in predictable cycles)
  - positive recurrent (expected time of returning is finite)
- **Limiting** long-run distribution

$$P^t \to \begin{pmatrix} \pi \\ \pi \\ \vdots \\ \pi \end{pmatrix} = \begin{pmatrix} \pi_1 & \pi_2 & \cdots & \pi_k \\ \pi_1 & \pi_2 & \cdots & \pi_k \\ \vdots & \vdots & & \vdots \\ \pi_1 & \pi_2 & \cdots & \pi_k \end{pmatrix} \quad \text{as } t \to \infty$$

Stationary distribution, cont.

- Limiting long-run distribution (unconditional probabilities)

$$P^t \to \begin{pmatrix} \pi \\ \pi \\ \vdots \\ \pi \end{pmatrix} = \begin{pmatrix} \pi_1 & \pi_2 & \cdots & \pi_k \\ \pi_1 & \pi_2 & \cdots & \pi_k \\ \vdots & \vdots & & \vdots \\ \pi_1 & \pi_2 & \cdots & \pi_k \end{pmatrix} \quad \text{as } t \to \infty$$

- Stationary distribution $\quad \pi = \pi P$
- Weather example:

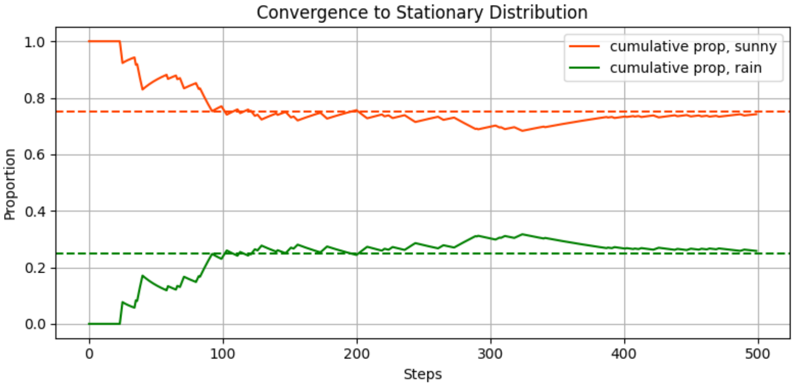
$$P = \begin{pmatrix} 0.9 & 0.1 \\ 0.3 & 0.7 \end{pmatrix}, \quad P^2 = \begin{pmatrix} 0.84 & 0.16 \\ 0.42 & 0.58 \end{pmatrix}$$

$$P^5 = \begin{pmatrix} 0.77 & 0.23 \\ 0.69 & 0.31 \end{pmatrix}, \quad P^{100} = \begin{pmatrix} 0.75 & 0.25 \\ 0.75 & 0.25 \end{pmatrix}$$

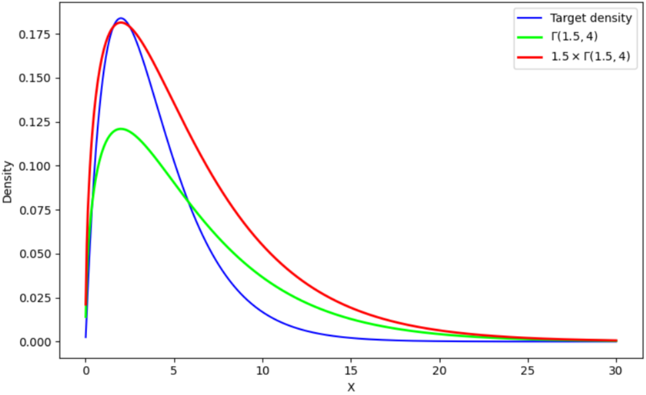
$$\pi = (0.75,\ 0.25)$$

The basic MCMC idea

- Simulate from discrete distribution $p(x)$ when $x \in \{s_1, \ldots, s_k\}$
- MCMC: simulate a Markov Chain with a stationary distribution that is exactly $p(x)$. Often continuous in our case.
- How to set up the transition matrix P?
  Metropolis-Hastings!



Rejection sampling



Initialize $\theta^{(0)}$ and iterate for $i = 1, 2, \ldots$

1. **Sample proposal**:

$$\theta_p \mid \theta^{(i-1)} \sim \mathcal{N}(\theta^{(i-1)}, c \cdot \Sigma)$$

2. Compute the acceptance probability

$$\alpha = \min\left(1, \frac{p(\theta_p \mid y)}{p(\theta^{(i-1)} \mid y)}\right)$$

3. With probability α set $\theta^{(i)} = \theta_p$, and $\theta^{(i)} = \theta^{(i-1)}$ otherwise.

Random walk Metropolis

- Assumption: we can compute $p(\theta_p \mid \mathbf{y})$ for any θ.
- Proportionality constants in posterior cancel out in:

$$\alpha = \min\left(1,\ \frac{p(\theta_p \mid \mathbf{y})}{p(\theta^{(i-1)} \mid \mathbf{y})}\right)$$

- In particular:

$$\frac{p(\theta_p \mid \mathbf{y})}{p(\theta^{(i-1)} \mid \mathbf{y})} = \frac{\frac{p(\mathbf{y}|\theta_p)p(\theta_p)}{p(\mathbf{y})}}{\frac{p(\mathbf{y}|\theta^{(i-1)})p(\theta^{(i-1)})}{p(\mathbf{y})}} = \frac{p(\mathbf{y} \mid \theta_p)p(\theta_p)}{p(\mathbf{y} \mid \theta^{(i-1)})p(\theta^{(i-1)})}$$

- **Proportional form of posterior is enough!**

$$\alpha = \min\left(1,\ \frac{p(\mathbf{y} \mid \theta_p)\, p(\theta_p)}{p(\mathbf{y} \mid \theta^{(i-1)})\, p(\theta^{(i-1)})}\right)$$

Random walk Metropolis

- Common choices of Σ in proposal $\mathcal{N}(\theta^{(i-1)}, c \cdot \Sigma)$:
  - $\Sigma = I$ (proposes 'off the cigar')
  - $\Sigma = J_{\hat{\theta},\mathbf{y}}^{-1}$ (propose 'along the cigar')
  - **Adaptive**. Start with $\Sigma = I$. Update Σ from initial run.
- Set c so average acceptance probability is 25–30%.
- **Good proposal**:
  - **Easy to sample**
  - **Easy to compute α**
  - Proposals should take reasonably large steps in θ-space
  - Proposals should not be reject too often.

- Generalization when the proposal density is not symmetric.

Initialize $\theta^{(0)}$ and iterate for $i = 1, 2, \ldots$

1. **Sample proposal**: $\theta_p \sim q(\cdot \mid \theta^{(i-1)})$
2. **Compute the acceptance probability**

$$\alpha = \min\left(1, \frac{p(\mathbf{y} \mid \theta_p)p(\theta_p)\, q(\theta^{(i-1)} \mid \theta_p)}{p(\mathbf{y} \mid \theta^{(i-1)})p(\theta^{(i-1)})\, q(\theta_p \mid \theta^{(i-1)})}\right)$$

3. **With probability α set** $\theta^{(i)} = \theta_p$, and $\theta^{(i)} = \theta^{(i-1)}$ otherwise.

The independence sampler

- **Independence sampler**:

$$q\left(\theta_p \mid \theta^{(i-1)}\right) = q(\theta_p)$$

- **Proposal** is **independent of previous draw**.
- Example:

$$\theta_p \sim t_\nu\left(\hat{\theta}, J^{-1}_{\hat{\theta},\mathbf{y}}\right),$$

  where $\hat{\theta}$ and $J_{\hat{\theta},\mathbf{y}}$ are computed by numerical optimization.
- Can be very **efficient**, but has a tendency to **get stuck**.
- Make sure that $q(\theta_p)$ has **heavier tails** than $p(\theta \mid \mathbf{y})$.

Metropolis-Hastings within Gibbs

- **Gibbs sampling** from $p(\theta_1, \theta_2, \theta_3 \mid \mathbf{y})$
  - Sample $p(\theta_1 \mid \theta_2, \theta_3, \mathbf{y})$
  - Sample $p(\theta_2 \mid \theta_1, \theta_3, \mathbf{y})$
  - Sample $p(\theta_3 \mid \theta_1, \theta_2, \mathbf{y})$
- When a **full conditional is not easily sampled** we can simulate from it using **MH**.
- Example: at ith iteration, propose θ_2 from $q(\theta_2 \mid \theta_1, \theta_3, \theta_2^{(i-1)}, \mathbf{y})$. Accept/reject.
- **Gibbs sampling is a special case of MH** when $q(\theta_2 \mid \theta_1, \theta_3, \theta_2^{(i-1)}, \mathbf{y}) = p(\theta_2 \mid \theta_1, \theta_3, \mathbf{y})$, which gives $\alpha = 1$. Always accept.

The efficiency of MCMC

- **How efficient** is MCMC compared to iid sampling?
- If $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(N)}$ are **iid** with variance σ^2, then:

$$\text{Var}(\bar{\theta}) = \frac{\sigma^2}{N}$$

- Autocorrelated $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(N)}$ generated by **MCMC**

$$\text{Var}(\bar{\theta}) = \frac{\sigma^2}{N}\left(1 + 2\sum_{k=1}^{\infty}\rho_k\right)$$

  where $\rho_k = \text{Corr}(\theta^{(i)}, \theta^{(i+k)})$ is the autocorrelation at lag k.
- **Inefficiency factor**

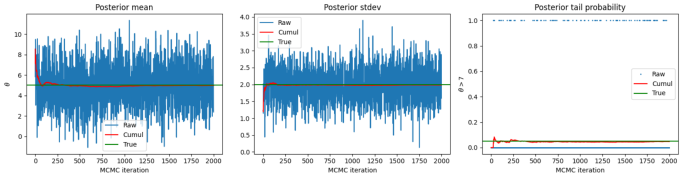
$$\text{IF} = 1 + 2\sum_{k=1}^{\infty}\rho_k$$

- **Effective sample size** from MCMC:

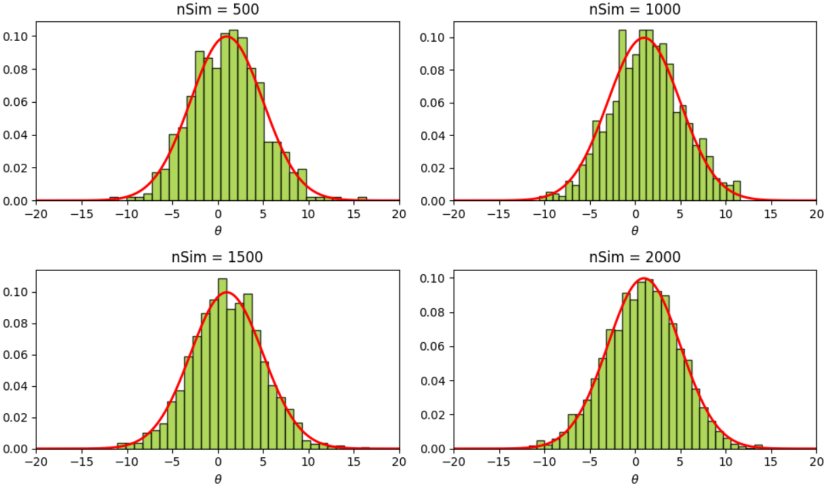
$$\text{ESS} = \frac{N}{\text{IF}}$$

Burn-in and convergence

- How long **burn-in**?
- **How long to sample** after burn-in?
- **Thinning**? Keeping every h draw reduces autocorrelation.
- **Convergence diagnostics**
  - Raw plots of simulated sequences (trajectories)
  - CUSUM plots
  - **Variance reduction**: the error in a direct Monte Carlo simulation goes as $\sigma/\sqrt{n}$. Two ways we can reduce the error, Run the simulation for a longer time, i.e., increase n or find a different formulation of the Monte Carlo that has a smaller σ.
  - Paper:
    Convergence diagnostics for Markov chain Monte Carlo, Roy 2019



Burn-in and convergence



Wrapping up

Today we discussed:

- Monte Carlo simulation
- Gibbs sampling
- Data augmentation
  - Mixture models
  - Probit regression
- Regularized regression

Some References

- Pratt J., Raiffa H., Schlaifer R., Introduction to Statistical Decision Theory, The MIT Press, 2008
- Hoffmann M. D., Gelman A., The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo, arXiv:1111.4246, 2011
- A. Gelman, J. B. Carlin, H. S. Stern, Bayesian Data Analysis, CRC Press, 2013
- Walsh B., Markov Chain Monte Carlo and Gibbs Sampling, Lecture Notes for EEB 596z, 2002
- R. A. Howard, Dynamic Programming and Markov Process, The MIT Press, 1960
- Rabiner L. R., A tutorial on hidden Markov models and selected applications in speech recognition, Proceedings of the IEEE 77.2, 1989
- W. K. Hastings, Monte Carlo sampling methods using Markov chains and their applications, Biometrik, 57/1, 04/1970
- Kevin B. Korb, Ann E. Nicholson, Bayesian Artificial Intelligence, CRC Press, 2010 Pearl J., Causality, Cambridge University Press, 2009
- L. E. Baum, T. Petrie, Statistical Inference for Probabilistic Functions of Finite State Markov Chains, The Annals of Mathematical Statistics, 37, 1966

Thank you very much!

ANY QUESTIONS OR COMMENTS?