Monte Carlo methods

Geir Storvik

Geilo Winter school 2023

UiO : Universitetet i Oslo

Norsk Regnesentral
NORWEGIAN COMPUTING CENTER

Outline

- Topics
  - What, how and why Monte Carlo
  - Markov chain Monte Carlo
  - Sequential Monte Carlo
  - Advanced/recent methods
- Goal
  - Introduce a range of Monte Carlo methods
  - Some mathematical background
    - Mainly to understand why and how methods work
    - Somewhat informal
- Mainly theory/illustration through examples
- Mainly statistical examples
  - Practial use (MCMC): Turing/Julia by Jose and Tor Erlend

Outline today

1. Monte Carlo methods
   - Monte Carlo for calculating integrals
2. Examples of integration problems
   - Bayesian inference
   - Other examples
3. Properties of Monte Carlo
4. Simulation techniques
5. Auxiliary variables
6. Variance reduction methods
7. Approximate Bayesian compututation

Monte Carlo methods

What is the Monte Carlo method?

- Essentially a numerical method for calculating integrals $I = \int_{\mathcal{X}} h(\boldsymbol{x}) d\boldsymbol{x}$
- Reformulate integral:

$$I = \int_{\mathcal{X}} h(\boldsymbol{x}) d\boldsymbol{x} = \int_{\mathcal{X}} \frac{h(\boldsymbol{x})}{q(\boldsymbol{x})} q(\boldsymbol{x}) d\boldsymbol{x} = E^{q(\boldsymbol{x})} \left[\frac{h(\boldsymbol{x})}{q(\boldsymbol{x})} q(\boldsymbol{x}) \right]$$

Last equality if $q(\boldsymbol{x})$ is a density over $\mathcal{X}$.
$E^{q(\boldsymbol{x})}$ means the expectation with respect to the distribution $q(\boldsymbol{x})$

What is the Monte Carlo method?

- Essentially a numerical method for calculating integrals $I = \int_{\mathcal{X}} h(\boldsymbol{x})d\boldsymbol{x}$
- Reformulate integral:

$$I = \int_{\mathcal{X}} h(\boldsymbol{x})d\boldsymbol{x} = \int_{\mathcal{X}} \frac{h(\boldsymbol{x})}{q(\boldsymbol{x})} q(\boldsymbol{x})d\boldsymbol{x} = E^{q(\boldsymbol{x})}\left[\frac{h(\boldsymbol{x})}{q(\boldsymbol{x})} q(\boldsymbol{x})\right]$$

  Last equality if $q(\boldsymbol{x})$ is a density over $\mathcal{X}$.
  $E^{q(\boldsymbol{x})}$ means the expectation with respect to the distribution $q(\boldsymbol{x})$
- Expectations can be approximated by averages:

$$\widehat{I}_N = \frac{1}{N}\sum_{i=1}^{N} \frac{h(\boldsymbol{x}^i)}{q(\boldsymbol{x}^i)}, \quad \boldsymbol{x}_i \overset{iid}{\sim} q(\boldsymbol{x}) \quad \text{Monte Carlo integration}$$

- $\widehat{I}_N$ is a Monte Carlo estimate of I.
- References:
  - Givens and Hoeting (2012): *Computational statistics*
  - Robert and Casella (1999): *Monte Carlo statistical methods*

Examples of integration problems

Why interest in integrals?

- Can in practice solve a huge range of problems
    - Bayesian inference
        - Missing data
        - Hierarchical models
    - Tool for efficient learning of neural networks
    - Solving PDEs
    - Monte Carlo testing
    - ...

Bayesian inference

- Data model (likelihood) $p(\boldsymbol{y}|\boldsymbol{\theta})$.
- Bayesian approach: Include prior information through a density $p(\boldsymbol{\theta})$.
- Prior: Describe our knowledge before data are collected
- Bayesians: Treat θ as a random variable

Bayesian inference

- Data model (likelihood) $p(\boldsymbol{y}|\boldsymbol{\theta})$.
- Bayesian approach: Include prior information through a density $p(\boldsymbol{\theta})$.
- Prior: Describe our knowledge before data are collected
- Bayesians: Treat θ as a random variable
- Bayes theorem:

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

$$p(\boldsymbol{y}) = \int_{\boldsymbol{\theta}} p(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\theta})d\boldsymbol{\theta}$$

Bayesian inference

- Data model (likelihood) $p(\boldsymbol{y}|\boldsymbol{\theta})$.
- Bayesian approach: Include prior information through a density $p(\boldsymbol{\theta})$.
- Prior: Describe our knowledge before data are collected
- Bayesians: Treat θ as a random variable
- Bayes theorem:

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

$$p(\boldsymbol{y}) = \int_{\boldsymbol{\theta}} p(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\theta})d\boldsymbol{\theta}$$

- Posterior: Updated knowledge based on both prior and data
- Bayesian paradigm: All relevant information about $\boldsymbol{\theta}$ is contained in the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{y})$

Bayesian inference

- Data model (likelihood) $p(\boldsymbol{y}|\boldsymbol{\theta})$.
- Bayesian approach: Include prior information through a density $p(\boldsymbol{\theta})$.
- Prior: Describe our knowledge before data are collected
- Bayesians: Treat θ as a random variable
- Bayes theorem:

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

$$p(\boldsymbol{y}) = \int_{\theta} p(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\theta})d\boldsymbol{\theta}$$

- Posterior: Updated knowledge based on both prior and data
- Bayesian paradigm: All relevant information about $\boldsymbol{\theta}$ is contained in the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{y})$
- Extract summaries: $\hat{\mu}_f = E[f(\boldsymbol{\theta})|\boldsymbol{y}] = \int_{\theta} f(\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}$

Statistical physics - Landau and Binder (2021)

- Define $w_{\mathbf{x}}(t)$ to be the probability of a system being in state $\boldsymbol{x}$ at time t.
- Define $P(\boldsymbol{x}^*|\boldsymbol{x})$ to be the transition rate from $\boldsymbol{x}$ to $\boldsymbol{x}^*$ (assumed time-independent)
- The master equation for evolution of $w_{\mathbf{x}}(t)$:

$$\frac{dw_{\mathbf{x}}}{dt} = \sum_{\boldsymbol{x}^*} \left[w_{\mathbf{x}^*}(t) P(\boldsymbol{x}|\boldsymbol{x}^*) - w_{\mathbf{x}}(t) P(\boldsymbol{x}^*|\boldsymbol{x}) \right]$$

with $\sum_{\boldsymbol{x}} w_{\mathbf{x}}(t) = 1$

Statistical physics - Landau and Binder (2021)

- Define $w_{\boldsymbol{x}}(t)$ to be the probability of a system being in state $\boldsymbol{x}$ at time t.
- Define $P(\boldsymbol{x}^*|\boldsymbol{x})$ to be the transition rate from $\boldsymbol{x}$ to $\boldsymbol{x}^*$ (assumed time-independent)
- The master equation for evolution of $w_{\boldsymbol{x}}(t)$:

$$\frac{dw_{\boldsymbol{x}}}{dt} = \sum_{\boldsymbol{x}^*} [w_{\boldsymbol{x}^*}(t)P(\boldsymbol{x}|\boldsymbol{x}^*) - w_{\boldsymbol{x}}(t)P(\boldsymbol{x}^*|\boldsymbol{x})]$$

  with $\sum_{\boldsymbol{x}} w_{\boldsymbol{x}}(t) = 1$
- Equilibrium: $\frac{dw_{\boldsymbol{x}}}{dt} = 0$,

$$p_{\boldsymbol{x}} = \lim_{t\to\infty} w_{\boldsymbol{x}}(t) = \frac{1}{Z} e^{-E(\boldsymbol{x})/(kT)}$$

  where $E(\boldsymbol{x})$ is the energy of state $\boldsymbol{x}$, k while T is the temperature.
- With $\beta = (kT)^{-1}$, the partition function is $Z = \sum_{\boldsymbol{x}} e^{-\beta E(\boldsymbol{x})}$ Typically impossible to compute exactly, unkown
- Of interest:

$$E[Q] = \sum_{\boldsymbol{x}} Q(\boldsymbol{x}) \frac{1}{Z} e^{-\beta E(\boldsymbol{x})}$$

Bayesian inference and statistical physics

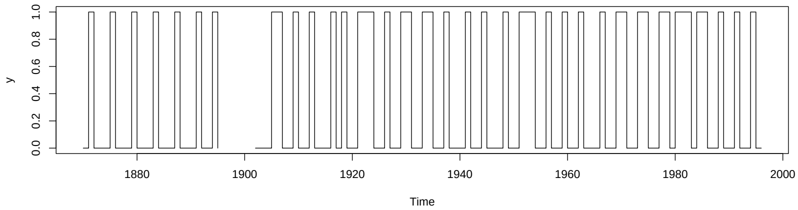
- Probability distributions:
  - Bayesian: $p(\boldsymbol{x}|\boldsymbol{y}) = \frac{p(\boldsymbol{x})p(\boldsymbol{y}|\boldsymbol{x})}{p(\boldsymbol{y})}$
  - Physics: $p_{\boldsymbol{x}} = \frac{1}{Z}e^{-\beta E(\boldsymbol{x})}$
- Nominator on (minus) log-scale
  - Bayesian: $-\log p(\boldsymbol{x}) - \log p(\boldsymbol{y}|\boldsymbol{x})$
  - Physics: Scaled energy $\beta E(\boldsymbol{x})$
- Denominator:
  - Bayesian: Marginal likelihood $p(\boldsymbol{y})$
  - Physics: Partition function Z
- In both cases: Expectations of interest
  - Possibly expectations of several different functions simultaneously
  - Physics: For different values of β
  - Statistics: Possible for different models

Hierarchical/state space models

- Interest in cyclic behaviour of lemmings populations
- Possible simple model: $\boldsymbol{x}_t = \log(N_t)$

$$x_t = ax_{t-1} + \sigma\varepsilon_t \qquad\qquad \varepsilon_t \sim N(0,1)$$

- Trap data: Typically very short time series
- Old church books: Written down if large or small lemmings populations within a year.



Lemmings data - cont

- Process model: $x_t = ax_{t-1} + \sigma\varepsilon_t$
- Possible observation model

$$
\begin{aligned}
Y_t \sim& \text{Binom}(1, p_t) \qquad\qquad\qquad t = 1, ..., T\\
p_t =& \exp(x_t)/(1 + \exp(x_t))
\end{aligned}
$$

- Parameters $\boldsymbol{\theta} = (a, \sigma^2)$

$$
\begin{tikzcd}
 & & \boldsymbol{\theta} \arrow[dl] \arrow[d] \arrow[dr] & & & \text{Parameters} \\
{} \arrow[r] & \boldsymbol{x}_{t-1} \arrow[r] \arrow[d] & \boldsymbol{x}_t \arrow[r] \arrow[d] & \boldsymbol{x}_{t+1} \arrow[r] \arrow[d] & {} & \text{Process} \\
 & \boldsymbol{y}_{t-1} & \boldsymbol{y}_t & \boldsymbol{y}_{t+1} & & \text{Observations}
\end{tikzcd}
$$

- Likelihood for data:

$$
L(\boldsymbol{\theta}) \equiv p(\boldsymbol{y}|\boldsymbol{\theta}) = \int_{x_{1:T}} p(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\theta})p(\boldsymbol{x}|\boldsymbol{\theta})d\boldsymbol{x}
$$

- Maximum likelihood: Need to optimize an integral

Bayesian extension

- Consider the previous example, but within a Bayesian setting.
- In that case, describe our prior knowledge about $\boldsymbol{\theta} = (a, \sigma^2)$ through a probability distribution $p(\boldsymbol{\theta})$.
- Update our knowledge by Bayes theorem:

$$
\begin{aligned}
p(\boldsymbol{\theta}|\boldsymbol{y}) &= \frac{p(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\theta})}{p(\boldsymbol{y})} \\
p(\boldsymbol{y}) &= \int_{\boldsymbol{\theta}} p(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\theta})d\boldsymbol{\theta} \\
p(\boldsymbol{y}|\boldsymbol{\theta}) &= \int_{x_{1:T}} p(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\theta})p(\boldsymbol{x}|\boldsymbol{\theta})d\boldsymbol{x}
\end{aligned}
$$

Bayesian extension

- Consider the previous example, but within a Bayesian setting.
- In that case, describe our prior knowledge about $\boldsymbol{\theta} = (a, \sigma^2)$ through a probability distribution $p(\boldsymbol{\theta})$.
- Update our knowledge by Bayes theorem:

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

$$p(\boldsymbol{y}) = \int_{\theta} p(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\theta})d\boldsymbol{\theta}$$

$$p(\boldsymbol{y}|\boldsymbol{\theta}) = \int_{x_{1:T}} p(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\theta})p(\boldsymbol{x}|\boldsymbol{\theta})d\boldsymbol{x}$$

- Summary statistics:

$$E[\boldsymbol{\theta}_i|\boldsymbol{y}] = \int_{\theta} \theta_i p(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}$$

$$\text{Var}[\boldsymbol{\theta}_i|\boldsymbol{y}] = \int_{\theta} \theta_i^2 p(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta} - (E[\boldsymbol{\theta}_i|\boldsymbol{y}])^2$$

Tracking automobiles using GPS measurements

$$
\begin{aligned}
(v_t^x, v_t^y, v_t^z) &= \text{Position of vehicle} \\
(s_{t,i}^x, s_{t,i}^y, s_{t,i}^z) &= \text{Position of satellite } i \\
y_{t,i} &= \text{time of signal from satellite } i \text{ to GPS}
\end{aligned}
$$

- Simplified model

$$y_{t,i} = \sqrt{(v_t^x - s_{t,i}^x)^2 + (v_t^y - s_{t,i}^y)^2 + (v_t^z - s_{t,i}^z)^2} + \varepsilon_{t,i}, \quad i = 1, 2, \ldots, n_t$$

with $\{\varepsilon_{t,i}\}$ independent noise terms.

Tracking automobiles using GPS measurements

$$
\begin{aligned}
(v_t^x, v_t^y, v_t^z) &= \text{Position of vehicle}\\
(s_{t,i}^x, s_{t,i}^y, s_{t,i}^z) &= \text{Position of satellite } i\\
y_{t,i} &= \text{time of signal from satellite } i \text{ to GPS}
\end{aligned}
$$

- Simplified model

$$y_{t,i} = \sqrt{(v_t^x - s_{t,i}^x)^2 + (v_t^y - s_{t,i}^y)^2 + (v_t^z - s_{t,i}^z)^2} + \varepsilon_{t,i}, \quad i = 1, 2, \dots, n_t$$

  with $\{\varepsilon_{t,i}\}$ independent noise terms.
- Assume available model for movement: $p(\boldsymbol{v}_t | \boldsymbol{v}_{t-1})$.
- Aim:

$$p(\boldsymbol{v}_t|\boldsymbol{y}_{1:t}) = \int_{\boldsymbol{v}_{1:t-1}} p(\boldsymbol{v}_{1:t}|\boldsymbol{y}_{1:t})d\boldsymbol{v}_{1:t-1} = \int_{\boldsymbol{v}_{1:t-1}} \frac{p(\boldsymbol{v}_{1:t})p(\boldsymbol{y}_{1:t}|\boldsymbol{v}_{1:t})}{p(\boldsymbol{y}_{1:t})} d\boldsymbol{v}_{1:t-1}$$

$$p(\boldsymbol{v}_{1:t}) = p(\boldsymbol{v}_1)\prod_{s=2}^{t} p(\boldsymbol{v}_s|\boldsymbol{v}_{s-1})$$

$$p(\boldsymbol{y}_{1:t}) = \int_{\boldsymbol{v}_{1:t}} p(\boldsymbol{v}_{1:t})p(\boldsymbol{y}_{1:t}|\boldsymbol{v}_{1:t})d\boldsymbol{v}_{1:t}$$

Model dynamics - simplified model

- Linear dynamics

$$\begin{aligned} \boldsymbol{v}_t &= (v_t^x, v_t^y, v_t^z, \dot{v}_t^x, \dot{v}_t^y, \dot{v}_t^z)^T \\ &= \boldsymbol{\Phi} \boldsymbol{v}_{t-1} + \boldsymbol{\eta}_t, \qquad \boldsymbol{\eta}_t \sim N(\boldsymbol{0}, \sigma_Q^2 \boldsymbol{Q}) \end{aligned}$$

where

$$\boldsymbol{\Phi} = \begin{pmatrix} \boldsymbol{I}_3 & \boldsymbol{I}_3 \\ \boldsymbol{0} & \boldsymbol{I}_3 \end{pmatrix}, \qquad \boldsymbol{Q} = \begin{pmatrix} \frac{q_v^2 D_t^3}{3} \boldsymbol{I}_3 & t\frac{q_{cd}^2 D_t}{2} \boldsymbol{I}_3 \\ \frac{q_{cd}^2 D_t}{2} \boldsymbol{I}_3 & q_{cb}^2 D_t \boldsymbol{I}_3 \end{pmatrix}$$

Model dynamics - simplified model

- Linear dynamics

$$\begin{aligned} \boldsymbol{v}_t =& (v_t^x, v_t^y, v_t^z, \dot{v}_t^x, \dot{v}_t^y, \dot{v}_t^z)^T \\ =& \boldsymbol{\Phi} \boldsymbol{v}_{t-1} + \boldsymbol{\eta}_t, \qquad \boldsymbol{\eta}_t \sim N(\mathbf{0}, \sigma_Q^2 \boldsymbol{Q}) \end{aligned}$$

where

$$\boldsymbol{\Phi} = \begin{pmatrix} \boldsymbol{I}_3 & \boldsymbol{I}_3 \\ \mathbf{0} & \boldsymbol{I}_3 \end{pmatrix}, \qquad \boldsymbol{Q} = \begin{pmatrix} \frac{q_v^2 D_t^3}{3} \boldsymbol{I}_3 & t \frac{q_{cd}^2 D_t}{2} \boldsymbol{I}_3 \\ \frac{q_{cd}^2 D_t}{2} \boldsymbol{I}_3 & q_{cb}^2 D_t \boldsymbol{I}_3 \end{pmatrix}$$

- Combined model

$$\begin{aligned} \boldsymbol{v}_t =& \boldsymbol{\Phi} \boldsymbol{v}_{t-1} + \boldsymbol{\eta}_t, \\ y_{t,i} =& \sqrt{(v_t^x - s_{t,i}^x)^2 + (v_t^y - s_{t,i}^y)^2 + (v_t^z - s_{t,i}^z)^2} + \varepsilon_{t,i}, \;\; i = 1, 2, \ldots, n_t \end{aligned}$$

example of a state space model

Model dynamics - simplified model

- Linear dynamics

$$\begin{aligned}\boldsymbol{v}_t =&(v_t^x, v_t^y, v_t^z, \dot{v}_t^x, \dot{v}_t^y, \dot{v}_t^z)^T\\ =&\boldsymbol{\Phi}\boldsymbol{v}_{t-1} + \boldsymbol{\eta}_t, \qquad\qquad \boldsymbol{\eta}_t \sim N(\boldsymbol{0}, \sigma_Q^2\boldsymbol{Q})\end{aligned}$$

where

$$\boldsymbol{\Phi} = \begin{pmatrix} \boldsymbol{I}_3 & \boldsymbol{I}_3 \\ \boldsymbol{0} & \boldsymbol{I}_3 \end{pmatrix}, \qquad\qquad \boldsymbol{Q} = \begin{pmatrix} \frac{q_v^2 D_t^3}{3}\boldsymbol{I}_3 & t\frac{q_{cd}^2 D_t}{2}\boldsymbol{I}_3 \\ \frac{q_{cd}^2 D_t}{2}\boldsymbol{I}_3 & q_{cb}^2 D_t \boldsymbol{I}_3 \end{pmatrix}$$

- Combined model

$$\begin{aligned}\boldsymbol{v}_t =&\boldsymbol{\Phi}\boldsymbol{v}_{t-1} + \boldsymbol{\eta}_t,\\ y_{t,i} =&\sqrt{(v_t^x - s_{t,i}^x)^2 + (v_t^y - s_{t,i}^y)^2 + (v_t^z - s_{t,i}^z)^2} + \varepsilon_{t,i}, \quad i = 1, 2, ..., n_t\end{aligned}$$

example of a state space model

- Challenge: Compute $p(\boldsymbol{v}_t|\boldsymbol{y}_{1:t})$ for each t in real time
- Need to utilize computation performed on previous time step

Model selection

- Genetic association studies: Which genes influence a certain phenotype (presence of cancer, size, etc)
- Linear model including all possible variables:

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i$$

Model selection

- Genetic association studies: Which genes influence a certain phenotype (presence of cancer, size, etc)
- Linear model including all possible variables:

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i$$

- Reasonable to assume that some x_{ij}'s do not influence the response, modification:

$$Y_i = \beta_0 + \sum_{j=1}^{p} \gamma_j \beta_j x_{ij} + \varepsilon_i \qquad \gamma_j \in \{0, 1\}.$$

Model selection

- Genetic association studies: Which genes influence a certain phenotype (presence of cancer, size, etc)
- Linear model including all possible variables:

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i$$

- Reasonable to assume that some x_{ij}'s do not influence the response, modification:

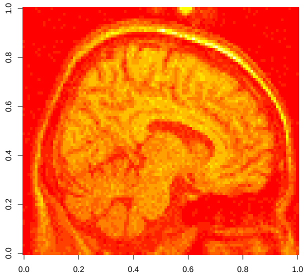
$$Y_i = \beta_0 + \sum_{j=1}^{p} \gamma_j \beta_j x_{ij} + \varepsilon_i \qquad \gamma_j \in \{0, 1\}.$$

- 2^p possible models, how to find the best ones?
  - $p = 20, 2^p = 1\,048\,576$, $p = 100, 2^p = 1.267651 * 10^{30}$
- Combinatorial problem

Image segmentation

- MRI tissue classification problem
- Three major tissue classes (cerebrospinal fluid (CSF), gray matter (GM), white matter (WM))
- Intensities assumed normally distributed with class-dependent means and variances:

$$y_{ij}|C_{ij} = k \sim N(\mu_k, \sigma_k^2)$$



- Bayes formula ($\pi_k = \Pr(C_{ij} = k)$):

$$\Pr(C_{ij} = k|y_{ij}) = \frac{\pi_k p(y_{ij}|C_{ij} = k)}{\sum_{l=1}^{3} \pi_l p(y_{ij}|C_{ij} = l)}$$

- Easy to calculate individually for each square (pixel)

Image segmentation - spatial structure

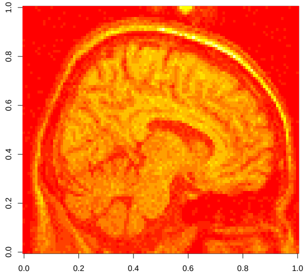
- Expect some smoothness in class-structure
- Markov Random field/Potts model:

$$\begin{aligned}\Pr(\boldsymbol{C}) =& \Pr(C_{11}, \ldots, C_{n_1 n_2})\\ =& \frac{1}{Z} e^{-\beta \sum_{||(i,j)-(i'j')||=1} I(C_{ij} \neq C_{i'j'})}\end{aligned}$$

- Now interested in

$$\Pr(\boldsymbol{C}|\boldsymbol{y}) = \frac{\Pr(\boldsymbol{C}) \prod_{ij} p(y_{ij}|C_{ij})}{\sum_{\boldsymbol{C}'} \Pr(\boldsymbol{C}') \prod_{ij} p(y_{ij}|C'_{ij})}$$

  - The sum in the denominator contains K^n terms,
    - K = number of class
    - n = number of pixels.
  - Discrete type of "integration"



Machine learning

- Search engines, recommendation platforms, speech and image recognition
- Large data sets, complex models
- Deep neural networks

Deep neural network

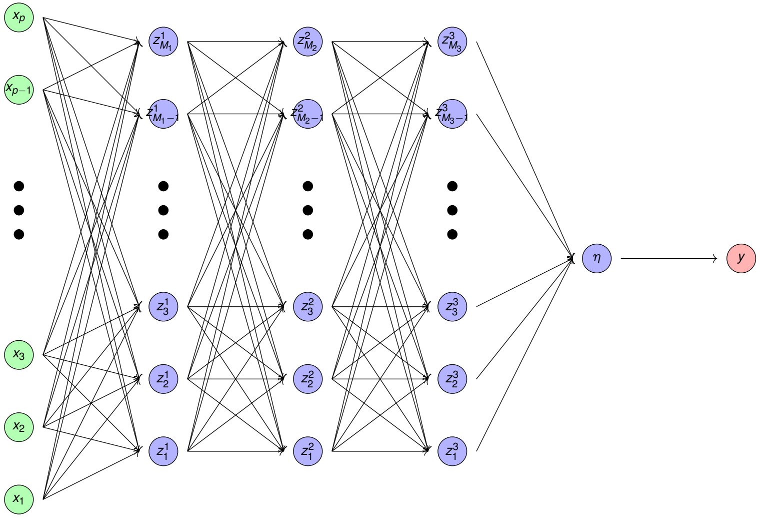


Figure: (Deep) Neural network with three hidden layer.

Learning neural networks

- Neural networks: $\boldsymbol{y} \approx f(\boldsymbol{x};\boldsymbol{\omega})$
- Typical criterion for continuous output:

$$g(\boldsymbol{\omega}) = \sum_{i=1}^{n}(y_i - f(\boldsymbol{x}_i;\boldsymbol{\omega}))^2$$

- Gradient decent:

$$\boldsymbol{\omega}^{(s+1)} = \boldsymbol{\omega}^{(s)} + \alpha\nabla g(\boldsymbol{\omega}^{(s)})$$

- If n is large, an unbiased estimate of $\nabla g(\boldsymbol{\omega}^{(s)})$ can be applied
- Simple Monte Carlo application: Use subsample
  - Need to use the reparametrization trick in order to obtain unbiasedness

Bayesian Neural networks

- Neural networks: $\boldsymbol{y} \approx f(\boldsymbol{x};\boldsymbol{\omega})$
- Bayesian approaches
    - Priors on $\boldsymbol{\omega}$.
    - Bayesian inference

$$p(y^*|x^*,\boldsymbol{x},\boldsymbol{y}) = \int_{\boldsymbol{\omega}} p(y^*|x^*,\boldsymbol{\omega})p(\boldsymbol{\omega}|\boldsymbol{x},\boldsymbol{y})d\boldsymbol{\omega}$$

    - Standard NN:

$$p(y^*|x^*,\boldsymbol{x},\boldsymbol{y}) \approx p(y^*|x^*,\widehat{\boldsymbol{\omega}})$$

    - Bayesian approach a huge computational challenge
- Discussion: Why do we want to do this?

Properties of Monte Carlo

Properties of Monte Carlo integration

- Reformulated integral:

$$I = \int_{\mathcal{X}} \frac{h(\boldsymbol{x})}{q(\boldsymbol{x})} q(\boldsymbol{x}) d\boldsymbol{x} = E^{q(\boldsymbol{x})}\left[\frac{h(\boldsymbol{x})}{q(\boldsymbol{x})} q(\boldsymbol{x})\right]$$

- Monte Carlo estimate:

$$\widehat{I}_N = \frac{1}{N}\sum_{i=1}^{N} \frac{h(\boldsymbol{x}_i)}{q(\boldsymbol{x}_i)} \qquad \boldsymbol{x}_i \sim q(\boldsymbol{x})$$

Properties of Monte Carlo integration

- Reformulated integral:

$$I = \int_{\mathcal{X}} \frac{h(\boldsymbol{x})}{q(\boldsymbol{x})} q(\boldsymbol{x}) d\boldsymbol{x} = E^{q(\boldsymbol{x})}\left[\frac{h(\boldsymbol{x})}{q(\boldsymbol{x})} q(\boldsymbol{x})\right]$$

- Monte Carlo estimate:

$$\widehat{I}_N = \frac{1}{N} \sum_{i=1}^{N} \frac{h(\boldsymbol{x}_i)}{q(\boldsymbol{x}_i)} \qquad \boldsymbol{x}_i \sim q(\boldsymbol{x})$$

- Properties:

$$\begin{aligned} E^{q(\boldsymbol{x})}[\widehat{I}_N] &= I && \text{Unbiased} \\ \mathrm{Var}^{q(\boldsymbol{x})}[\widehat{I}_N] &= \frac{1}{N} \mathrm{Var}^{p(\boldsymbol{x})}\left[\frac{h(\boldsymbol{x})}{q(\boldsymbol{x})}\right] && \text{If independent samples} \\ &= \frac{1}{N}\sigma_h^2 && \text{In general} \end{aligned}$$

- Discussion: Discuss this result

Simulation techniques

Simulation techniques

- Monte Carlo require $\boldsymbol{x}_i \sim q(\boldsymbol{x})$
- Exact methods
  - Inversion/transformation methods
  - Rejection sampling

Simulation techniques

- Monte Carlo require $\boldsymbol{x}_i \sim q(\boldsymbol{x})$
- Exact methods
  - Inversion/transformation methods
  - Rejection sampling
- Approximate methods
  - Sampling importance resampling
  - Approximate Bayesian computing
  - Sequential Monte Carlo
  - Markov chain Monte Carlo

Simulation techniques

- Monte Carlo require $\boldsymbol{x}_i \sim q(\boldsymbol{x})$
- Exact methods
  - Inversion/transformation methods
  - Rejection sampling
- Approximate methods
  - Sampling importance resampling
  - Approximate Bayesian computing
  - Sequential Monte Carlo
  - Markov chain Monte Carlo
- Variance reduction methods
  - Importance sampling
- Auxiliary variables

The inversion method and the transformation methods

- Assume continuous distribution, density $p(x)$, CDF

$$p(x) = \int_{-\infty}^{x} p(u)du$$

- Assume $U \sim \text{Unif}[0, 1]$
- Define $X = F^{-1}(U)$:

The inversion method and the transformation methods

- Assume continuous distribution, density $p(x)$, CDF

$$p(x) = \int_{-\infty}^{x} p(u)du$$

- Assume $U \sim \text{Unif}[0, 1]$
- Define $X = F^{-1}(U)$:

$$\begin{aligned} \Pr(X \leq x) =& \Pr(F^{-1}(U) \leq x) \\ =& \Pr(U \leq p(x)) = p(x) \end{aligned}$$

  showing that $X \sim p(x)$!
- Assumes possible to generate U (good routines available)
- Assumes $F^{-1}(U)$ available
- Only applicable for univariate distributions
- Special case of transformation methods: $X = g(U)$

Pseudo-random variables

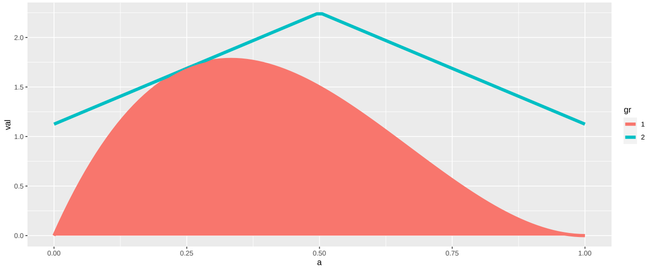
- All (?) random generators on computers rely on $U \sim \text{Unif}[0, 1]$
- Computers are deterministic
- Pseudo sequence:

$$u_{t+1} = (a * u_t + b) \text{ modulo } m$$

- Unix: $a = 1103515245$, $b = 12345$, $m = 2^{31}$
- Discuss this setting

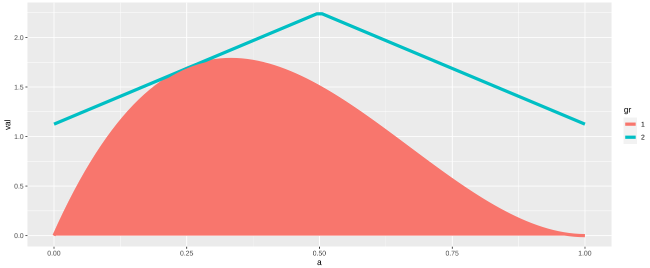
Rejection sampling

- Difficult to simulate from $p(x)$ directly
- Easy to simulate from $g(x) \approx p(x)$.
- Assume $\exists\alpha \leq 1$ such that for all x: $p(x) \leq g(x)/\alpha \equiv e(x)$ (the envelope)



Rejection sampling

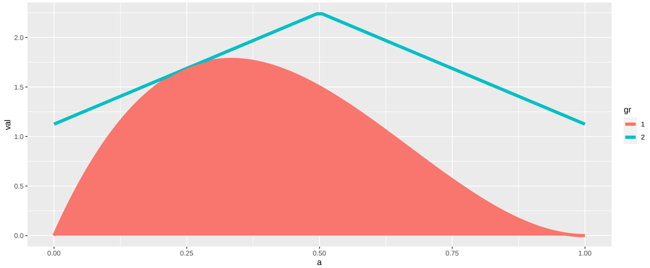
- Difficult to simulate from $p(x)$ directly
- Easy to simulate from $g(x) \approx p(x)$.
- Assume $\exists\alpha \leq 1$ such that for all x: $p(x) \leq g(x)/\alpha \equiv e(x)$ (the envelope)



- Algorithm:
  1. Sample $Y \sim g(\cdot)$.
  2. Sample $U \sim \text{Unif}(0, 1)$.
  3. If $U \leq p(Y)/e(Y)$, put $X = Y$, otherwise return to step 1

Rejection sampling

- Difficult to simulate from $p(x)$ directly
- Easy to simulate from $g(x) \approx p(x)$.
- Assume $\exists\alpha \leq 1$ such that for all x: $p(x) \leq g(x)/\alpha \equiv e(x)$ (the envelope)



- Algorithm:
  1. Sample $Y \sim g(\cdot)$.
  2. Sample $U \sim \text{Unif}(0, 1)$.
  3. If $U \leq p(Y)/e(Y)$, put $X = Y$, otherwise return to step 1
- $\alpha = \Pr(U \leq \frac{p(Y)}{e(Y)})$ is the probability for acceptance
- α^{-1} is the expected number of iterations.

Proof rejection sampling

- Distribution of X:

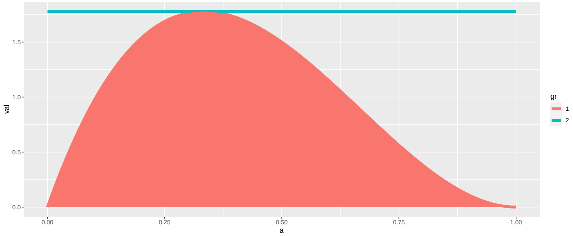
$$\Pr(X \leq x) = \Pr(Y \leq x | U \leq \tfrac{p(Y)}{e(Y)}) = \frac{\Pr(Y \leq x, U \leq \frac{p(Y)}{e(Y)})}{\Pr(U \leq \frac{p(Y)}{e(Y)})}$$

$$= \frac{\int_{-\infty}^{x} \int_{0}^{p(y)/e(y)} du g(y) dy}{\int_{-\infty}^{\infty} \int_{0}^{p(y)/e(y)} du g(y) dy} = \frac{\int_{-\infty}^{x} \frac{p(y)}{e(y)} g(y) dy}{\int_{-\infty}^{\infty} \frac{p(y)}{e(y)} g(y) dy}$$

$$= \int_{-\infty}^{x} p(y) dy$$

Example - rejection sampling

1. Aim: Simulate from Beta distribution:
$$p(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$
2. $\arg\max_x p(x) = \frac{\alpha-1}{\alpha+\beta-2} = x^*$
3. Define $g(x) = 1; 0 < x < 1$. Then $g(x) \geq p(x)/p(x^*)$
4. Accept if $U \leq p(x)/p(x^*)$



5. `beta_rej.R`

Auxiliary variables

Auxiliary variables

- Assume interest is in

$$\pi(\boldsymbol{x}) = \int_{\boldsymbol{z}} p(\boldsymbol{x}, \boldsymbol{z}) d\boldsymbol{z}$$

- Simulation directly from $p(\boldsymbol{x})$ is difficult
  - but simulation from $p(\boldsymbol{x}, \boldsymbol{z})$ is easy!
- Assuming $(\boldsymbol{x}, \boldsymbol{z})$ is a sample from $p(\boldsymbol{x}, \boldsymbol{z})$
- Then $\boldsymbol{x}$ is a sample from $p(\boldsymbol{x})$

Example

- Model

$$\begin{aligned} \sigma \sim& \text{Unif}[0,2] \\ X|\sigma \sim& N(0,\sigma) \\ E[X] =& E[E[X|\sigma]] = E[0] = 0 \end{aligned}$$

- Simulation of X directly?

$$p(x) = \int_0^2 \frac{1}{2} \frac{1}{\sqrt{2\pi\sigma}} \exp(-0.5\sigma^{-2}x^2)dx$$

- Possible through numerical integration and rejection sampling
- Easier to simulate directly from model!

Variance reduction methods

Monte Carlo method

- Aim :

$$\mu = E^{f(\mathbf{X})}[h(\mathbf{X})] = \begin{cases} \int_{\mathbf{x}} h(\mathbf{x})f(\mathbf{x}))d\mathbf{x} & \mathbf{x} \text{ continuous} \\ \sum_{\mathbf{x}} h(\mathbf{x})f(\mathbf{x}) & \mathbf{x} \text{ discrete} \end{cases}$$

Monte Carlo method

- Aim :

$$\mu = E^{f(\mathbf{X})}[h(\mathbf{X})] = \begin{cases} \int_{\mathbf{x}} h(\mathbf{x})f(\mathbf{x}))d\mathbf{x} & \mathbf{x} \text{ continuous} \\ \sum_{\mathbf{x}} h(\mathbf{x})f(\mathbf{x}) & \mathbf{x} \text{ discrete} \end{cases}$$

- Monte Carlo:
  1. Simulate $\mathbf{X}_i \sim f(\mathbf{x})$, $i = 1, \ldots, n$
  2. Approximate μ by $\hat{\mu}_{MC} = \frac{1}{n}\sum_{i=1}^{n} h(\mathbf{x}_i)$.
- Properties:
  - Unbiased $E[\hat{\mu}_{MC}] = \mu$
  - If $X_1, \ldots, X_n$ are independent
    - Variance: $\text{var}[\hat{\mu}_{MC}] = \frac{1}{n}\text{var}[h(\mathbf{X})]$
    - Consistent: $\hat{\mu}_{MC} \rightarrow \mu$ as $n \rightarrow \infty$ if $\text{var}[h(\mathbf{X})] < \infty$

Monte Carlo method

- Aim :

$$\mu = E^{f(\mathbf{X})}[h(\mathbf{X})] = \begin{cases} \int_{\mathbf{x}} h(\mathbf{x})f(\mathbf{x})d\mathbf{x} & \mathbf{x} \text{ continuous} \\ \sum_{\mathbf{x}} h(\mathbf{x})f(\mathbf{x}) & \mathbf{x} \text{ discrete} \end{cases}$$

- Monte Carlo:
  1. Simulate $\mathbf{X}_i \sim f(\mathbf{x})$, $i = 1, \ldots, n$
  2. Approximate μ by $\hat{\mu}_{MC} = \frac{1}{n}\sum_{i=1}^{n} h(\mathbf{x}_i)$.
- Properties:
  - Unbiased $E[\hat{\mu}_{MC}] = \mu$
  - If $X_1, \ldots, X_n$ are independent
    - Variance: $\text{var}[\hat{\mu}_{MC}] = \frac{1}{n}\text{var}[h(\mathbf{X})]$
    - Consistent: $\hat{\mu}_{MC} \to \mu$ as $n \to \infty$ if $\text{var}[h(\mathbf{X})] < \infty$
  - Estimate of variance:

$$\widehat{\text{var}}[\hat{\mu}_{MC}] = \frac{1}{n-1}\sum_{i=1}^{n}(h(\mathbf{x}_i) - \hat{\mu}_{MC})^2$$

- Can we do better than this?

Importance sampling

- Rewriting

$$\mu = \int h(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} = \int \frac{h(\boldsymbol{x})p(\boldsymbol{x})}{g(\boldsymbol{x})}g(\boldsymbol{x})d\boldsymbol{x} = \frac{\int \frac{h(\boldsymbol{x})p(\boldsymbol{x})}{g(\boldsymbol{x})}g(\boldsymbol{x})d\boldsymbol{x}}{\int \frac{p(\boldsymbol{x})}{g(\boldsymbol{x})}g(\boldsymbol{x})d\boldsymbol{x}}$$

- Assume $X_1, \ldots, X_n$ iid from $g(\boldsymbol{x})$.
- Two alternative estimates

$$\hat{\mu}^*_{IS} = \frac{1}{n}\sum_{i=1}^{n} h(\boldsymbol{X}_i)w^*(\boldsymbol{X}_i), \quad w^*(X_i) = \frac{p(\boldsymbol{X}_i)}{g(\boldsymbol{X}_i)}$$

$$\hat{\mu}_{IS} = \sum_{i=1}^{n} h(\boldsymbol{X}_i)w(\boldsymbol{X}_i), \quad w(\boldsymbol{X}_i) = \frac{w^*(\boldsymbol{X}_i)}{\sum_{j=1}^{n} w^*(\boldsymbol{X}_i)}$$

- $w^*(\boldsymbol{X}_i)$ called importance weights
- $w(\boldsymbol{X}_i)$ called the normalized importance weights
- Discussion: Which one to use (in which situations)?

Importance sampling

- Rewriting

$$\mu = \int h(\mathbf{x})f(\mathbf{x})d\mathbf{x} = \int \frac{h(\mathbf{x})f(\mathbf{x})}{g(\mathbf{x})}g(\mathbf{x})d\mathbf{x} = \frac{\int \frac{h(\mathbf{x})f(\mathbf{x})}{g(\mathbf{x})}g(\mathbf{x})d\mathbf{x}}{\int \frac{f(\mathbf{x})}{g(\mathbf{x})}g(\mathbf{x})d\mathbf{x}}$$

- Assume $X_1, \ldots, X_n$ iid from $g(\mathbf{x})$.
- Two alternative estimates

$$\hat{\mu}^*_{IS} = \frac{1}{n}\sum_{i=1}^{n} h(\mathbf{X}_i)w^*(\mathbf{X}_i), \quad w^*(X_i) = \frac{f(\mathbf{X}_i)}{g(\mathbf{X}_i)}$$

$$\hat{\mu}_{IS} = \sum_{i=1}^{n} h(\mathbf{X}_i)w(\mathbf{X}_i), \quad w(\mathbf{X}_i) = \frac{w^*(\mathbf{X}_i)}{\sum_{j=1}^{n} w^*(\mathbf{X}_j)}$$

- $w^*(\mathbf{X}_i)$ called importance weights
- $w(\mathbf{X}_i)$ called the normalized importance weights
- Choise of g:
  - Simple to simulate from
  - Result in low variance

Other variance reduction methods

- Rao-Blackwellization
- Antitetic variables
- Common rando numbers
- Control variates

Approximate Bayesian compututation

Approximate Bayesian computation

- Assume of interest $p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\theta})}{p(\boldsymbol{y})}$
- Possible approach:
  1. Simulate $(\boldsymbol{\theta}^*, \boldsymbol{y}^*) \sim p(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\theta})$
  2. Accept if $\boldsymbol{y}^* = \boldsymbol{y}$
- Can show: Accepted $\boldsymbol{\theta}^* \sim p(\boldsymbol{\theta}|\boldsymbol{y})$

Approximate Bayesian computation

- Assume of interest $p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\theta})}{p(\boldsymbol{y})}$
- Possible approach:
  1. Simulate $(\boldsymbol{\theta}^*, \boldsymbol{y}^*) \sim p(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\theta})$
  2. Accept if $\boldsymbol{y}^* = \boldsymbol{y}$
- Can show: Accepted $\boldsymbol{\theta}^* \sim p(\boldsymbol{\theta}|\boldsymbol{y})$
- Problem: Very unlikely that $\boldsymbol{y}^* = \boldsymbol{y}$
- The ABC method: Accept if $\text{Dist}(\boldsymbol{y}^*, \boldsymbol{y}) < \varepsilon$
- Typically: $\text{Dist}(\boldsymbol{y}^*, \boldsymbol{y}) = \text{Dist}(S(\boldsymbol{y}^*), S(\boldsymbol{y}))$ where $S(\boldsymbol{y})$ is some summary statistic
- Gives an approximate sample
  - Robust with respect to model assumptions

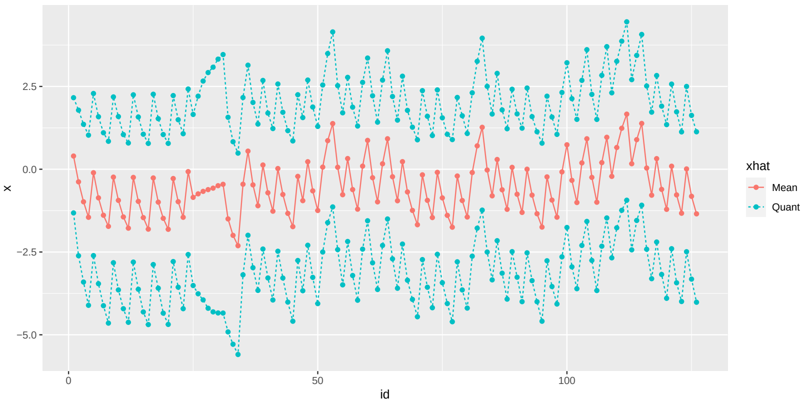
Lemmings data

- Model (simplified, $a_2 = 0$, σ known)

$$
\begin{aligned}
\boldsymbol{y}_t &\sim \text{Binom}\left(1, \frac{\exp(\boldsymbol{x}_t)}{1+\exp(\boldsymbol{x}_t)}\right) \\
\boldsymbol{x}_t &= a\boldsymbol{x}_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma^2) \\
a &\sim \text{Uniform}[0, 1]
\end{aligned}
$$

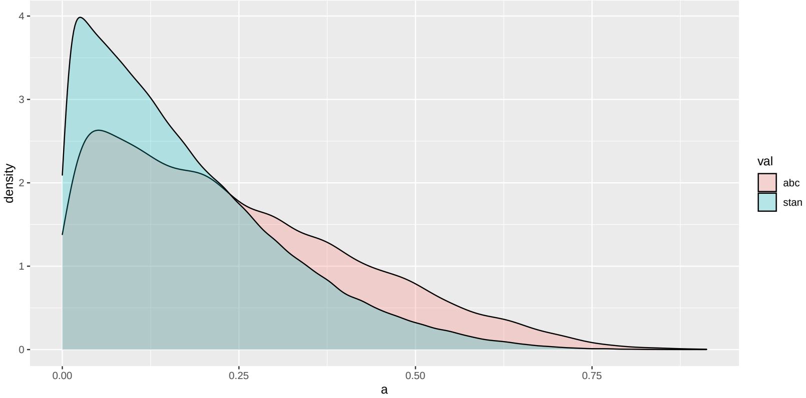
- Of interest: $p(\boldsymbol{x}_t | \boldsymbol{y}_{1:t})$, $p(a | \boldsymbol{y}_{1:t})$

Lemmings - latent process



Results - lemmings data

- $S(\boldsymbol{y}) = (\frac{1}{n}\sum_i I(y_i - y_{i-1} = 1), \frac{1}{n}\sum_i I(y_i - y_{i-1} = -1)$
- $N = 100\,000$, accepted=15 294
- R-script: `ABC_lemmings_parest.R`



References

G. H. Givens and J. A. Hoeting. *Computational statistics*, volume 710. John Wiley & Sons, 2012.

D. Landau and K. Binder. *A guide to Monte Carlo simulations in statistical physics*. Cambridge university press, 2021.

C. P. Robert and G. Casella. *Monte Carlo statistical methods*, volume 2. Springer, 1999.