Bayesiaans leren

Les 2: Markov Chain Monte Carlo

Joris Bierkens

Vakantiecursus

augustus 2019

Samenvatting en vooruitblik

- Veel statistische problemen kunnen we opvatten in een Bayesiaanse context

$$\pi(\theta) \propto \pi_0(\theta) \prod_{i=1}^{n} p(y_i \mid \theta), \quad \theta \in \mathcal{S}.$$

Samenvatting en vooruitblik

- Veel statistische problemen kunnen we opvatten in een Bayesiaanse context

$$\pi(\theta) \propto \pi_0(\theta) \prod_{i=1}^{n} p(y_i \mid \theta), \quad \theta \in \mathcal{S}.$$

- Bij predictie hebben we een model $p_{\text{model}}(y \mid \theta, x)$ en krijgen we de Bayesiaanse predictor

$$\widehat{p(y \mid x)} = \int_{\mathcal{S}} \pi(\theta) p_{\text{model}}(y \mid \theta, x) \, d\theta,$$

Samenvatting en vooruitblik

- Veel statistische problemen kunnen we opvatten in een Bayesiaanse context

$$\pi(\theta) \propto \pi_0(\theta) \prod_{i=1}^{n} p(y_i \mid \theta), \quad \theta \in \mathcal{S}.$$

- Bij predictie hebben we een model $p_{\text{model}}(y \mid \theta, x)$ en krijgen we de Bayesiaanse predictor

$$\widehat{p(y \mid x)} = \int_{\mathcal{S}} \pi(\theta) p_{\text{model}}(y \mid \theta, x) \, d\theta,$$

waarbij

$$\pi(\theta) \propto \pi_0(\theta) \prod_{i=1}^{n} p_{\text{model}}(y_i \mid \theta, x_i).$$

Samenvatting en vooruitblik

- Veel statistische problemen kunnen we opvatten in een Bayesiaanse context
$$\pi(\theta) \propto \pi_0(\theta) \prod_{i=1}^{n} p(y_i \mid \theta), \quad \theta \in \mathcal{S}.$$

- Bij predictie hebben we een model $p_{\text{model}}(y \mid \theta, x)$ en krijgen we de Bayesiaanse predictor
$$\widehat{p(y \mid x)} = \int_{\mathcal{S}} \pi(\theta) p_{\text{model}}(y \mid \theta, x)\, d\theta,$$
waarbij
$$\pi(\theta) \propto \pi_0(\theta) \prod_{i=1}^{n} p_{\text{model}}(y_i \mid \theta, x_i).$$

- Uitdaging: Hoe kunnen we de integraal over $\pi(\theta)$ berekenen?

Samenvatting en vooruitblik

- Veel statistische problemen kunnen we opvatten in een Bayesiaanse context

$$\pi(\theta) \propto \pi_0(\theta) \prod_{i=1}^{n} p(y_i \mid \theta), \quad \theta \in \mathcal{S}.$$

- Bij predictie hebben we een model $p_{\text{model}}(y \mid \theta, x)$ en krijgen we de Bayesiaanse predictor

$$\widehat{p(y \mid x)} = \int_{\mathcal{S}} \pi(\theta) p_{\text{model}}(y \mid \theta, x)\, d\theta,$$

waarbij

$$\pi(\theta) \propto \pi_0(\theta) \prod_{i=1}^{n} p_{\text{model}}(y_i \mid \theta, x_i).$$

- Uitdaging: Hoe kunnen we de integraal over $\pi(\theta)$ berekenen?
- Merk op: we kennen ook de normaliseringsconstante $\pi(\theta)$ niet zonder een integraal uit te rekenen!

Integralen uitrekenen

- We willen een integraal van de vorm $\pi(g) := \int_{\mathcal{S}} g(x)\pi(x)\,dx$ uitrekenen, waarbij we de normalisatieconstante van π niet kennen.

Integralen uitrekenen

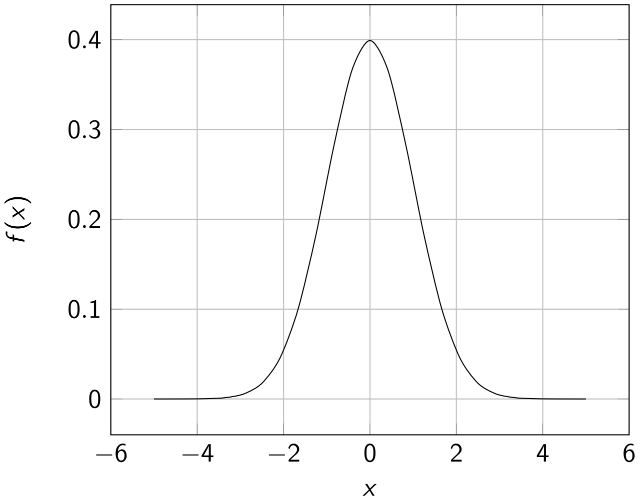
- We willen een integraal van de vorm $\pi(g) := \int_{\mathcal{S}} g(x)\pi(x)\, dx$ uitrekenen, waarbij we de normalisatieconstante van π niet kennen.
- Mogelijke aanpakken:
  - Numerieke integratie: aantal gridpunten schaalt helaas exponentieel in dimensie.

Integralen uitrekenen

- We willen een integraal van de vorm $\pi(g) := \int_{\mathcal{S}} g(x)\pi(x)\,dx$ uitrekenen, waarbij we de normalisatieconstante van π niet kennen.
- Mogelijke aanpakken:
  - Numerieke integratie: aantal gridpunten schaalt helaas exponentieel in dimensie.
  - Monte Carlo: als we $X_1, X_2, \ldots, X_d$ onafhankelijk kunnen trekken uit de verdeling $\pi(x)$ kunnen we de wet van de grote aantallen gebruiken:

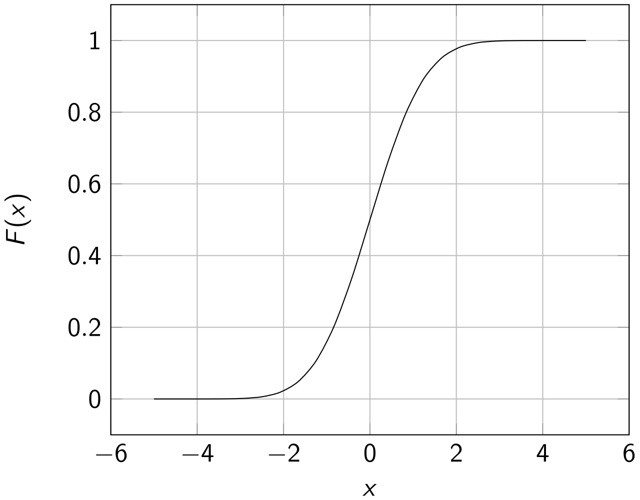
$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} g(X_i) = \pi(g).$$

Intermezzo: onafhankelijk trekken uit univariate verdeling



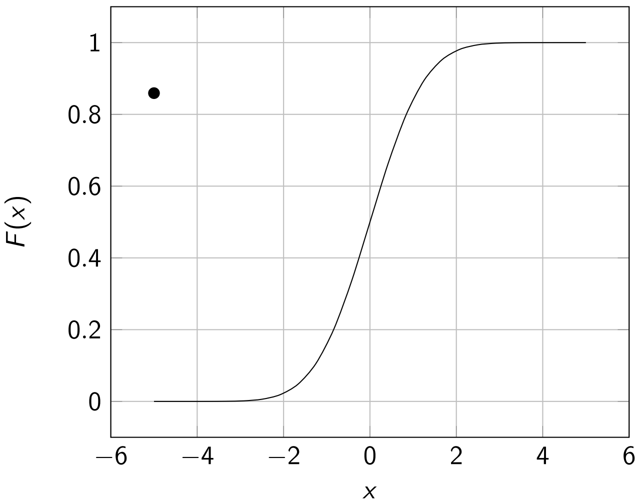
Kansdichtheidsfunctie van de standardnormale verdeling

Intermezzo: onafhankelijk trekken uit univariate verdeling

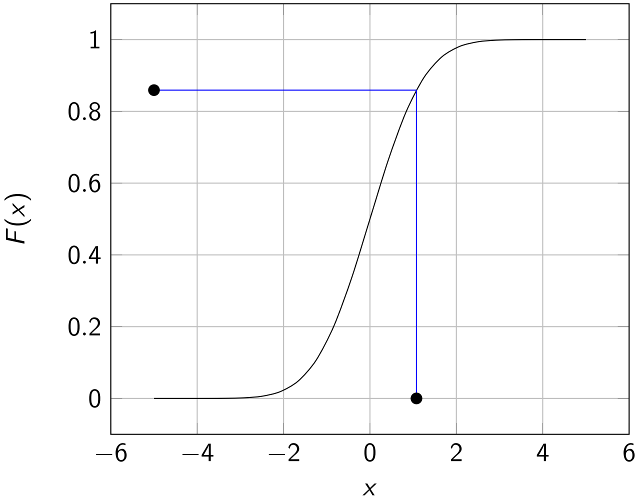


Cumulatieve verdelingsfunctie van de standardnormale verdeling

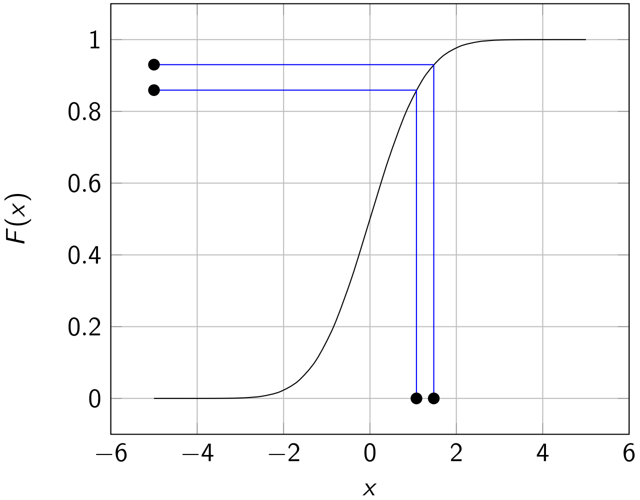
Intermezzo: onafhankelijk trekken uit univariate verdeling



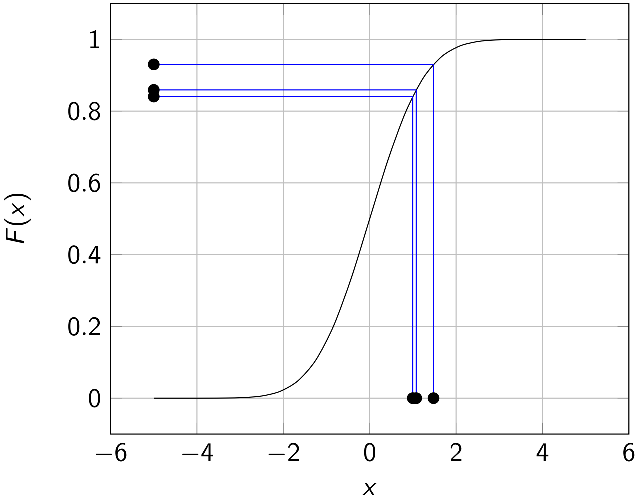
Intermezzo: onafhankelijk trekken uit univariate verdeling



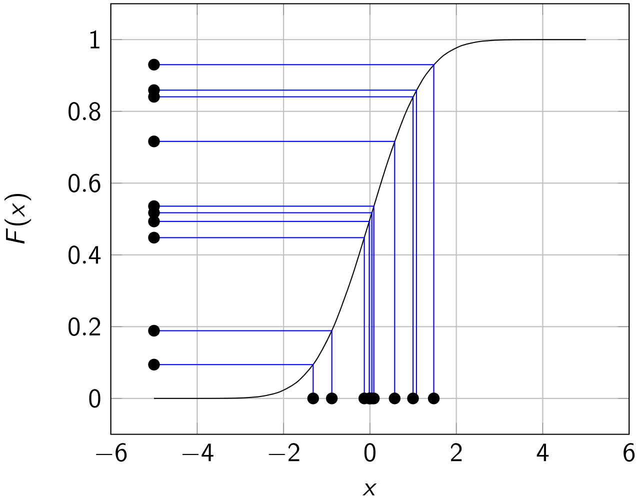
Intermezzo: onafhankelijk trekken uit univariate verdeling



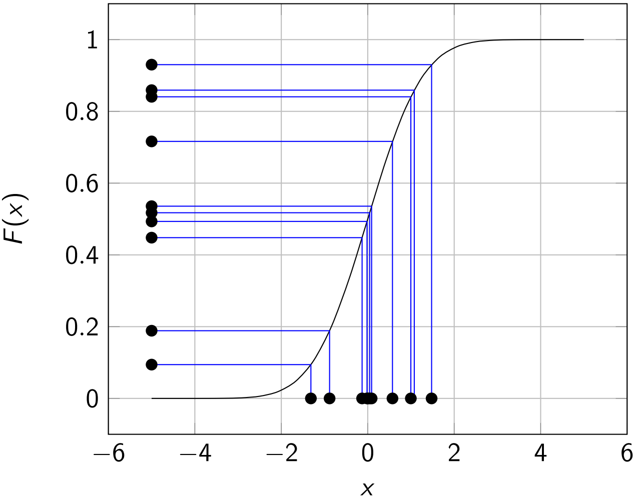
Intermezzo: onafhankelijk trekken uit univariate verdeling



Intermezzo: onafhankelijk trekken uit univariate verdeling



Intermezzo: onafhankelijk trekken uit univariate verdeling



Deze methode heet de inverse CDF techniek.

Intermezzo: onafhankelijk trekken met ‘rejection sampling’

Gooien van een dobbelsteen is verdeeld volgens uniforme verdeling over $\{1, 2, 3, 4, 5, 6\}$.
Hoe kunnen we nu trekken uit een uniforme verdeling over $\{1, 2, 3, 4, 5\}$?

Intermezzo: onafhankelijk trekken met 'rejection sampling'

Gooien van een dobbelsteen is verdeeld volgens uniforme verdeling over $\{1, 2, 3, 4, 5, 6\}$.
Hoe kunnen we nu trekken uit een uniforme verdeling over $\{1, 2, 3, 4, 5\}$?

1. Gooi de dobbelsteen met uitkomst X
2. Als $X = 6$, herhaal vanaf stap 1; anders retourneer X.

Intermezzo: onafhankelijk trekken met ‘rejection sampling’

Gooien van een dobbelsteen is verdeeld volgens uniforme verdeling over $\{1,2,3,4,5,6\}$.
Hoe kunnen we nu trekken uit een uniforme verdeling over $\{1,2,3,4,5\}$?

1. Gooi de dobbelsteen met uitkomst X
2. Als $X = 6$, herhaal vanaf stap 1; anders retourneer X.

Rejection sampling

- Doel: trekken van π (niet genormaliseerd).
- Middel: kansverdeling $\widetilde{\pi}$ zodat $\pi(x) \leq c\widetilde{\pi}(x)$ voor alle x.

Intermezzo: onafhankelijk trekken met 'rejection sampling'

Gooien van een dobbelsteen is verdeeld volgens uniforme verdeling over $\{1, 2, 3, 4, 5, 6\}$.
Hoe kunnen we nu trekken uit een uniforme verdeling over $\{1, 2, 3, 4, 5\}$?

1. Gooi de dobbelsteen met uitkomst X
2. Als $X = 6$, herhaal vanaf stap 1; anders retourneer X.

Rejection sampling

- Doel: trekken van π (niet genormaliseerd).
- Middel: kansverdeling $\widetilde{\pi}$ zodat $\pi(x) \leq c\widetilde{\pi}(x)$ voor alle x.

1. Trek X uit $\widetilde{\pi}$.
2. Met kans $1 - \pi(X)/c\widetilde{\pi}(X)$: herhaal vanaf 1, anders retourneer X.

Intermezzo: onafhankelijk trekken met 'rejection sampling'

Gooien van een dobbelsteen is verdeeld volgens uniforme verdeling over $\{1, 2, 3, 4, 5, 6\}$.
Hoe kunnen we nu trekken uit een uniforme verdeling over $\{1, 2, 3, 4, 5\}$?

1. Gooi de dobbelsteen met uitkomst X
2. Als $X = 6$, herhaal vanaf stap 1; anders retourneer X.

Rejection sampling

- Doel: trekken van π (niet genormaliseerd).
- Middel: kansverdeling $\widetilde{\pi}$ zodat $\pi(x) \leq c\widetilde{\pi}(x)$ voor alle x.

1. Trek X uit $\widetilde{\pi}$.
2. Met kans $1 - \pi(X)/c\widetilde{\pi}(X)$: herhaal vanaf 1, anders retourneer X.

Nadeel: in typische situaties wordt de acceptatiekans exponentieel klein.

Integralen uitrekenen

- We willen een integraal van de vorm $\pi(g) := \int_{\mathcal{S}} g(x)\pi(x)\,dx$ uitrekenen, waarbij we de normalisatieconstante van π niet kennen.
- Mogelijke aanpakken:
  - Numerieke integratie: aantal gridpunten schaalt helaas exponentieel in dimensie.
  - Monte Carlo: als we $X_1, X_2, \ldots, X_d$ onafhankelijk kunnen trekken uit de verdeling $\pi(x)$ kunnen we de wet van de grote aantallen gebruiken:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} g(X_i) = \pi(g).$$

Integralen uitrekenen

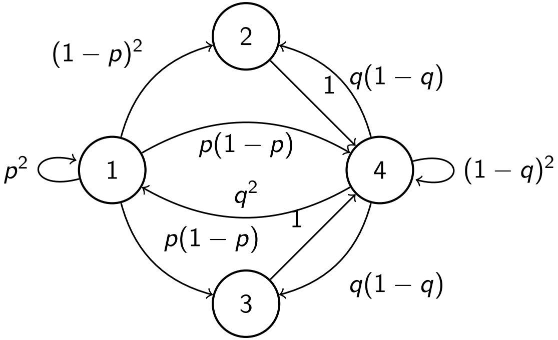
- We willen een integraal van de vorm $\pi(g) := \int_{\mathcal{S}} g(x)\pi(x)\,dx$ uitrekenen, waarbij we de normalisatieconstante van π niet kennen.
- Mogelijke aanpakken:
  - Numerieke integratie: aantal gridpunten schaalt helaas exponentieel in dimensie.
  - Monte Carlo: als we $X_1, X_2, \ldots, X_d$ onafhankelijk kunnen trekken uit de verdeling $\pi(x)$ kunnen we de wet van de grote aantallen gebruiken:

    $$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} g(X_i) = \pi(g).$$

    Onafhankelijk trekken uit een verdeling kan niet zomaar efficient gedaan worden.
  - Markov Chain Monte Carlo: het woord 'onafhankelijk' hierboven blijkt niet essentieel te zijn. We gaan een Markov keten bouwen met de correcte verdeling.

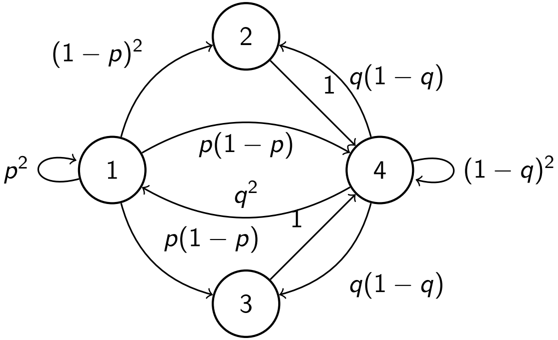
Markov-ketens

Markov-keten



Markov-ketens

Markov-keten



Transitiematrix

$$P = \begin{pmatrix} p^2 & (1-p)^2 & p(1-p) & p(1-p) \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ q^2 & q(1-q) & q(1-q) & (1-q)^2 \end{pmatrix}$$

$P(i,j)$ is de kans om van i naar j te springen.

Stationaire verdeling

$$P = \begin{pmatrix} p^2 & (1-p)^2 & p(1-p) & p(1-p) \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ q^2 & q(1-q) & q(1-q) & (1-q)^2 \end{pmatrix}$$

Stationaire verdeling

$$P = \begin{pmatrix} p^2 & (1-p)^2 & p(1-p) & p(1-p) \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ q^2 & q(1-q) & q(1-q) & (1-q)^2 \end{pmatrix}$$

Stel we beginnen vanuit een kansverdeling π en maken een willekeurige stap volgens P.

Stationaire verdeling

$$P = \begin{pmatrix} p^2 & (1-p)^2 & p(1-p) & p(1-p) \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ q^2 & q(1-q) & q(1-q) & (1-q)^2 \end{pmatrix}$$

Stel we beginnen vanuit een kansverdeling π en maken een willekeurige stap volgens P. De kans dat we in toestand j zijn na één stap is $\sum_{i=1}^{4} \pi(i)P(i,j)$.

Stationaire verdeling

π is een stationaire verdeling voor P als de verdeling na een stap vanuit π weer gelijk is aan π.

Stationaire verdeling

$$P = \begin{pmatrix} p^2 & (1-p)^2 & p(1-p) & p(1-p) \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ q^2 & q(1-q) & q(1-q) & (1-q)^2 \end{pmatrix}$$

Stel we beginnen vanuit een kansverdeling π en maken een willekeurige stap volgens P. De kans dat we in toestand j zijn na één stap is $\sum_{i=1}^4 \pi(i)P(i,j)$.

Stationaire verdeling

π is een stationaire verdeling voor P als de verdeling na een stap vanuit π weer gelijk is aan π.

$$\sum_i \pi(i)P(i,j) = \pi(j) \quad \text{voor alle } j.$$

Stationaire verdeling

$$P = \begin{pmatrix} p^2 & (1-p)^2 & p(1-p) & p(1-p) \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ q^2 & q(1-q) & q(1-q) & (1-q)^2 \end{pmatrix}$$

Stel we beginnen vanuit een kansverdeling π en maken een willekeurige stap volgens P. De kans dat we in toestand j zijn na één stap is $\sum_{i=1}^{4} \pi(i)P(i,j)$.

Stationaire verdeling

π is een stationaire verdeling voor P als de verdeling na een stap vanuit π weer gelijk is aan π.

$$\sum_i \pi(i)P(i,j) = \pi(j) \quad \text{voor alle } j.$$

Zij X_n de toestand van de Markov keten na n transities.

Convergentie

Onder milde condities (irreducibiliteit en aperiodiciteit), $X_n \to \pi$ in verdeling.

Stationaire verdeling

$$P = \begin{pmatrix} p^2 & (1-p)^2 & p(1-p) & p(1-p) \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ q^2 & q(1-q) & q(1-q) & (1-q)^2 \end{pmatrix}$$

Stel we beginnen vanuit een kansverdeling π en maken een willekeurige stap volgens P. De kans dat we in toestand j zijn na één stap is $\sum_{i=1}^{4} \pi(i)P(i,j)$.

Stationaire verdeling

π is een stationaire verdeling voor P als de verdeling na een stap vanuit π weer gelijk is aan π.

$$\sum_i \pi(i)P(i,j) = \pi(j) \quad \text{voor alle } j.$$

Zij X_n de toestand van de Markov keten na n transities.

Convergentie

Onder milde condities (irreducibiliteit en aperiodiciteit), $X_n \to \pi$ in verdeling. Dus we kunnen (bij benadering) trekken van π!

De uitdaging

- Voor een gegeven Markov transitiematrix P zijn we in staat de stationaire verdeling te berekenen

$$\sum_{i=1} \pi(i)P(i,j) = \pi(j) \quad \text{voor alle } j.$$

- Kunnen we dit ook omdraaien: gegeven een verdeling, wat is de bijbehorende Markov keten?
- Het blijkt dat er heel veel manieren zijn om Markov ketens te bouwen.

Markov Chain Monte Carlo algoritmen

Er zijn twee belangrijke generieke algoritmen:

- Metropolis-Hastings
- Gibbs sampling

Metropolis-Hastings

(Metropolis et al, 1953, Hastings et al, 1970)

Metropolis-Hastings algoritme

Vanuit een toestand X_i:

1. Stel een nieuwe toestand Y voor volgens 'zomaar' een Markov transitiematrix Q
2. Bereken de acceptatiekans
$$\alpha(X_i, Y) = \min\left(1, \frac{\pi(Y)Q(Y, X_0)}{\pi(X_i)Q(X_i, Y)}\right).$$
3. Definieer
$$X_{i+1} := \begin{cases} Y \quad (\text{accepteer}) & \text{met kans } \alpha(X_i, Y), \\ X_i \quad (\text{verwerp}) & \text{met kans } 1 - \alpha(X_i, Y). \end{cases}$$

- Q heet een proposal transitiematrix.
- Demo: `https://chi-feng.github.io/mcmc-demo/`

Gibbs sampling

(Glauber, 1963, Geman & Geman, 1984)

- Voor Gibbs sampling is een productstructuur vereist: $\mathcal{S} = \mathcal{S}_1 \times \cdots \times \mathcal{S}_n$.

Gibbs sampling

(Glauber, 1963, Geman & Geman, 1984)

- Voor Gibbs sampling is een productstructuur vereist: $\mathcal{S} = \mathcal{S}_1 \times \cdots \times \mathcal{S}_n$.
- Bovendien eisen we dat we kunnen trekken uit de conditionele verdelingen
$$\pi(x^j \mid x^1, \ldots, x^{j-1}, x^{j+1}, \ldots, x^n).$$

Gibbs sampling

(Glauber, 1963, Geman & Geman, 1984)

- Voor Gibbs sampling is een productstructuur vereist: $\mathcal{S} = \mathcal{S}_1 \times \cdots \times \mathcal{S}_n$.
- Bovendien eisen we dat we kunnen trekken uit de conditionele verdelingen
$$\pi(x^j \mid x^1, \ldots, x^{j-1}, x^{j+1}, \ldots, x^n).$$
- Een iteratie van het Gibbs sampling algoritme vanuit $\boldsymbol{X}_i = (X_i^1, \ldots, X_i^n) \in \mathcal{S}$ is als volgt:
  1. Kies willekeurig een component $j \in \{1, \ldots, n\}$
  2. Definieer
  $$X_{i+1}^k = X_i^k \quad \text{voor } k \neq j,$$
  en trek
  $$X_{i+1}^j \sim \pi(\cdot \mid X_i^1, \ldots, X_i^{j-1}, X_i^j, \ldots, X_i^n).$$

Samenvatting

We hebben kennisgemaakt met:

- Bayesiaanse statistiek als een heel algemeen conceptueel raamwerk voor statistiek
- Een manier om voorspellingsmethoden zoals neurale netwerken op een Bayesiaanse manier te interpreteren
- Numerieke methoden (MCMC) om berekeningen met Bayesiaanse statistiek uit te kunnen voeren.