Metody MCMC (STP139)

1. Úvod

Charakteristika: MCMC je třída algoritmů umožňující simulovat složité stochastické systémy.
Idea: Když chceme generovat z nějakého pravděpodobnostního rozdělení, tak zkonstruujeme markovský řetězec, jehož stacionární rozdělení je požadované rozdělení. Simulujeme markovský řetězec a po dostatečně velkém počtu kroků dostaneme přibližně výběr z daného rozdělení, pokud jsou splněny jisté předpoklady na řetězec, které zaručí, že limitní rozdělení existuje a splývá se stacionárním.
Otázky: Kolik je dostatečně velký počet kroků? Jak zkonstruovat takový markovský řetězec?
Odpovědi: Konstrukce markovského řetězce s daným stacionárním rozdělením není těžká, existuje řada postupů. Těžší je určit, po kolika krocích řetězec zkonverguje k limitnímu rozdělení s rozumnou chybou. Existují MCMC algoritmy, které dají přesný výběr z limitního rozdělení (tzv. perfektní simulace), a to v konečném čase, který je ovšem náhodný. Navíc je to za cenu dodatečných výpočtů.
Použití:

- možnost generovat výběry z komplikovaného modelu, který nás zajímá,
- kromě toho lze MCMC metody využít k výpočtu složitých (typicky vícerozměrných) integrálů. Dejme tomu, že chceme numericky spočítat $\int_{\mathcal{X}} h(x)f(x)\,\mathrm{d}x$, kde h je nějaká funkce a f je hustota nějakého rozdělení na prostoru $\mathcal{X}$. Vytvoříme Markovův řetězec se stacionárním rozdělením f a simulujeme jeden běh $X_1, X_2, \ldots$ tohoto řetězce. Po jistém čase T máme hodnoty z rozdělení přibližného f. Daný integrál pak aproximujeme pomocí $\frac{1}{N}\sum_{t=T+1}^{T+N} h(X_t)$. Využíváme silný zákon velkých čísel pro markovské řetězce (X_t nejsou nezávislé, jisté předpoklady jsou nutné – ergodicita). Výpočty takovýchto integrálů se objevují při statistické analýze modelu (maximální věrohodnost, bayesovská statistika).
- Metoda simulovaného žíhání se používá pro optimalizaci (hledání argumentu maxima nějaké funkce).

Teoretický základ: K pochopení simulace metodami MCMC je třeba rozumět vlastnostem markovských řetězců s diskrétním časem a obecnou množinou stavů (prostor $\mathcal{X}$ je většinou nespočetný). K analýze generovaných dat jsou potřebné postupy matematické statistiky.
Historie:

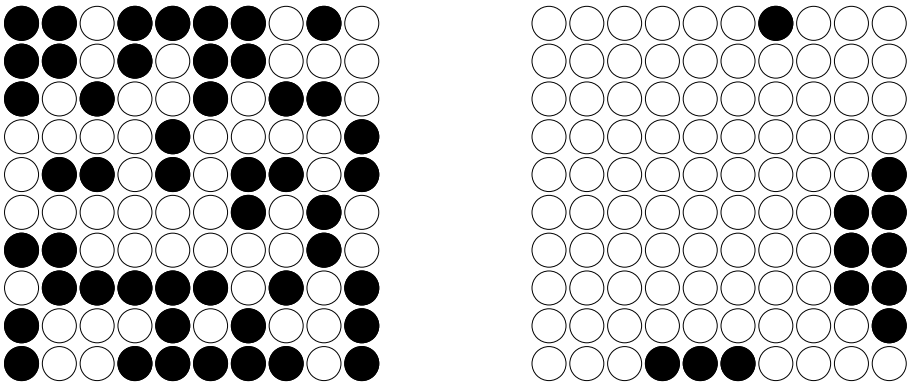
- První MCMC algoritmus byl vyvinut pro aplikace ve statistické fyzice – Metropolis a kol. (1953) [16], zobecnění algoritmu – Hastings (1970) [10], tzv. Metropolis-Hastingsův algoritmus.
- Řešení optimalizačních problémů metodou simulovaného žíhání – Kirkpatrick, Gelatt a Vecchi (1983) [14], Černý (1985) [3].
- Aplikace na statistické problémy poprvé až v 80. letech 20. století (Gibbsův výběrový plán): Geman, Geman (1984) [7] – restaurování digitálních obrázků; Gelfand, Smith (1990) [6] – bayesovská statistika.
- V 90. letech rozvoj teorie, sofistikovanější metody: Propp, Wilson (1996) [20] – perfektní simulace.
- V posledních 15 letech největší rozmach díky lepší výkonnosti počítačů a širokému množství aplikací v různých oborech.

Aplikace: všude, kde se vyskytují pravděpodobnostní modely, které vedou k složitým rozdělením (většinou na prostorech velké dimenze) přinášejícím výpočetní problémy.

(a) statistická fyzika: modely různých fyzikálních systémů, studium fázových přechodů.

Ilustrační příklad: Isingův model (1925) [11] – matematický model používaný ve statistické mechanice. Užívá se jako zjednodušený model feromagnetismu nebo k modelování chování kapalin a plynů. Uvažujme čtvercovou konečnou mříž. Každému vrcholu x mříže je přiřazena hodnota $\xi(x) \in \{-1, +1\}$ (orientace rotace atomu). Definujme hamiltonián $H(\xi) = -\sum_{x\sim y} \xi(x)\xi(y)$, kde $x \sim y$ značí, že x a y jsou sousedé na mříži (uvažujeme periodické okrajové podmínky). Pravděpodobnost konfigurace ξ je $\pi_\beta(\xi) = \frac{1}{Z_\beta}\mathrm{e}^{-\beta H(\xi)}$, kde parametr $\beta \geq 0$ se nazývá inverzní teplota a Z_β je normující konstanta. Pro $\beta = 0$ má každá konfigurace stejnou pravděpodobnost, jedná se o náhodné přiřazení -1 a $+1$ vrcholům mříže. Pro $\beta > 0$ má větší pravděpodobnost konfigurace, kde se sousedi přitahují. Pro $\beta \to \infty$ převládá jeden stav. Na levém obrázku je simulace modelu na mříži 10×10 pro $\beta = 0$, zatímco na pravém pro $\beta = 0.5$, černé kolečka představují vrcholy s hodnotami $+1$, bílé s hodnotami -1. Pro Isingův model v $\mathbb{Z}^2$ je kritická hodnota, kdy dochází k tzv. fázovému přechodu, rovna $\beta = \beta_c = \frac{1}{2}\log(1+\sqrt{2}) \doteq 0{,}441$ (analyticky spočtena Onsagerem [19], 1944). Pro $\beta > \beta_c$ je kov magnetizovaný, pro $\beta \leq \beta_c$ je neuspořádaný (více různých rovnovážných stavů, oba spiny zastoupeny

stejně). Zobecnění: obecnější graf než mřížka; energie odpovídající dvojici spinu (hraně grafu) může být obecnější než $\xi(x)\xi(y)$; větší dimenze; vnější magnetické pole (v definici H).



(b) informatika: přibližné určení počtů prvků velké množiny (čítací problémy), umělá inteligence, optimalizační problémy (např. problém obchodního cestujícího), studium znáhodněných algoritmů (chování pro rostoucí velikost problému).

Ilustrační příklad: hard-core model – mějme graf $G = (V, E)$, každému vrcholu grafu je přiřazena hodnota 0 nebo 1. Zajímají nás takové konfigurace, kde dva vrcholy spojené hranou nemají hodnotu 1 zároveň. Kolik je přípustných konfigurací? Jaký je střední počet jedniček v náhodné přípustné konfiguraci? Pro n vrcholů je všech možných přiřazení 0 a 1 celkem 2^n. Pro velké n nemožné počítat přímo. Např. pro $n = 64$ (mříž 8×8) je $2^{64} \doteq 1{,}8 \cdot 10^{19}$, což přesahuje možnosti počítačů. Proto se simuluje náhodná přípustná konfigurace metodami MCMC.

Druhý ilustrační příklad: náhodná q-obarvení – každý vrchol grafu má jednu z q barev tak, aby sousedé neměli stejnou. Pro rovinný graf stačí $q = 4$, aby množina všech q-obarvení byla neprázdná. Kolik je všech q-obarvení?

(c) prostorová statistika: vzorky z prostorových stochastických modelů – bodové procesy, náhodná pole.

(d) aplikovaná statistika: především bayesovský kontext – lze formulovat statistické modely, které by jinak nebylo možné efektivně analyzovat (použití v obrazové analýze, grafických modelech nebo při detekci změny). Uplatnění pro statistickou inferenci pro problémy v biostatistice, genetice, epidemiologii nebo finanční matematice (GARCH modely).

Literatura: Z celé řady knížek a monografií věnované problematice MCMC jmenujme [2], [5], [8], [9], [13], [18] a [21].

2. Simulace

Zde zmíníme pouze základy, zájemce o podrobnosti mohou navštěvovat přednášku prof. Antocha *Simulační metody a statistika*.

Při náhodných simulacích se využívá generátor pseudonáhodných čísel (jsou vytvořená deterministickým algoritmem, ale mají vlastnosti jako náhodná čísla). Budeme předpokládat, že máme dobrý generátor z rovnoměrného rozdělení na $[0, 1]$, měl by mít tyto vlastnosti: náhodnost (rovnoměrnost a nekorelovanost), dlouhá perioda, výpočetní efektivita, opakovatelnost (nastavení seed), přenositelnost (na různé počítače), homogenita.

2.1 Přímé metody

1. simulace náhodných veličin s diskrétním rozdělením: (x_k, p_k), $k = 1, 2, \ldots$
   (a) obecná metoda: interval $[0, 1]$ rozdělíme na disjunktní podintervaly

$$I_1 = [0, p_1], \quad I_n = (\sum_{k=1}^{n-1} p_k, \sum_{k=1}^{n} p_k] \quad \text{pro } n > 1.$$

Tedy každé hodnotě x_k přísluší interval I_k délky odpovídající pravděpodobnosti p_k. Nechť $U \sim R(0, 1)$, pokud $U \in I_n$, pak x_n je výběr z daného rozdělení. V praxi je podstatné, kolik

porovnání provedeme pro nalezení takového n. Nejpřirozenější je použití while cyklu (syntaxe jako v R):

```
k <- 1; u <- runif(1); s <- p[1];
while (s<u) { k <- k+1; s <- s+p[k]; }; print(x[k]);
```

Předpokládáme, že ve vektorech `p` a `x` jsou uloženy pravděpodobnosti p_k a hodnoty x_k. Pro tuto situaci je nejvýhodnější, když x_k jsou seřazeny od největší pravděpodobnosti k nejmenší. Pokud veličina může nabývat spočetně mnoha hodnot, nelze mít `p` a `x` uloženo jako vektor. Je možné pravděpodobnosti p_k počítat v každém kroku cyklu (často se s výhodou použije znalost předchozí hodnoty pravděpodobnosti).

Příklad: simulace z Poissonova rozdělení.

(b) využití interpretace nebo vlastností daného rozdělení.
Příklady: binomické (součet alternativních), geometrické (čekání na první úspěch), Poissonovo (definice Poissonova procesu přes exponenciální přírůstky).

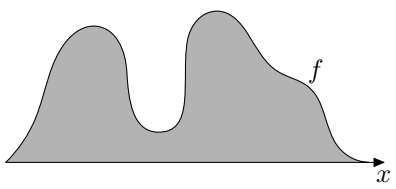
2. simulace náhodných veličin se spojitým rozdělením: hustota $f(x)$, distribuční funkce $F(x)$, kvantilová funkce $F^{-1}(u)$
   (a) inverzní metoda: pokud $U \sim R(0,1)$, pak $F^{-1}(U) \sim F$.
   Důkaz: $\mathbb{P}(F^{-1}(U) \le x) = \mathbb{P}(U \le F(x)) = F(x)$.
   Pozn.: metoda (a) u diskrétního rozdělení odpovídá této metodě.
   Příklad: $-\frac{1}{\lambda}\log U \sim Exp(\lambda)$.
   (b) využití interpretace daného rozdělení.
   Příklady: $\Gamma(n,\lambda)$ pro $n \in \mathbb{N}$ lze simulovat jako součet exponenciálních, χ^2_n jako součet druhých mocnin nezávislých normálních (jedná se o $\Gamma(n/2, 1/2)$).
   (c) transformační metoda: vhodná transformace ze známých.
   Příklad: normální $N(\mu,\sigma^2)$ – Box, Muller: $\sqrt{-2\log U_1}\cos 2\pi U_2$, $\sqrt{-2\log U_1}\sin 2\pi U_2 \sim N(0,1)$ nezávislé, když U_1, $U_2 \sim R(0,1)$ jsou nezávislé. Jde vlastně o to, že se dvojrozměrná hustota normálního rozdělení přepíše do polárních souřadnic. Když $X \sim N(0,1)$, tak $\mu+\sigma X \sim N(\mu,\sigma^2)$. Změna polohy a měřítka je jednoduchá transformace, která se dá využít v mnoha jiných rozděleních.
3. simulace náhodných vektorů
   (a) transformace: vhodná transformace z náhodného vektoru, který umíme simulovat (nejčastěji s nezávislými složkami).
   Příklad: normální $N_d(\mu,\Sigma)$ – nalezneme-li matici A (tzv. odmocninová matice) takovou, že $\Sigma = AA^T$, pak můžeme využít toho, že pokud $X \sim N_d(0, I_d)$, tak $Y = \mu + AX \sim N_d(\mu,\Sigma)$. K nalezení odmocninové matice lze užít Choleského rozklad (A bude dolní trojúhelníková).
   (b) využití interpretace daného rozdělení.
   Příklad: Wishartovo rozdělení (zobecnění Γ-rozdělení) – když $X_1,\dots,X_n$ je výběr z $N_d(\mu,\Sigma)$, tak $\sum_{i=1}^n (X_i-\mu)(X_i-\mu)^T \sim W_d(n/2, \Sigma^{-1}/2)$. Pro $d=1$ se jedná o $\sigma^2\chi^2_n$, obecně je $W_1(\alpha,\beta)$ přesně $\Gamma(\alpha,\beta)$.

2.2 Zamítací metoda

Uvažujme měřitelný prostor $\mathcal{X}$ se σ-konečnou mírou μ. Chceme simulovat náhodný element z rozdělení dané hustotou f vzhledem k μ.

Lemma 1. *Simulace $X \sim f$ je ekvivalentní simulaci (X,U) z rovnoměrného rozdělení na množině $\{(x,u) : 0 < u < f(x)\}$.*

Důkaz: Když $(X,U) \sim R(\{(x,u) : 0 < u < f(x)\})$, pak marginální rozdělení X je $f(x)$. Naopak když máme $X \sim f$ a vygenerujeme $U \sim R(0, f(X))$, tak $(X,U) \sim R(\{(x,u) : 0 < u < f(x)\})$, protože $f(x,u) = f(u \mid x)f(x) = 1$.

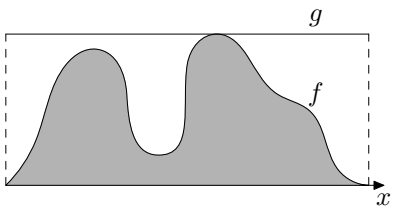


□

Pokud neumíme generovat rovnoměrně z oblasti $\{(x,u) : 0 < u < f(x)\}$, tak můžeme generovat z větší a omezit se jen na body, které padnou dovnitř.

Předpokládejme, že známe hustotu f až na normující konstantu, tj. známe $f^* = cf$ a c neznáme. Mějme pomocnou hustotu g splňující $f^*(x) \leq Mg(x)$ pro všechna $x \in \mathcal{X}$. Předpokládejme, že známe konstantu M a že z hustoty g umíme jednoduše simulovat. Potom můžeme definovat následující algoritmus simulace z rozdělení s hustotou f.

Algoritmus 1. *Zamítací metoda (rejection method)*:
1. generuj $X \sim g$,
2. generuj $U \sim R(0,1)$ nezávisle na X,
3. když $U \leq \frac{f^*(X)}{Mg(X)}$, tak polož $Z = X$, jinak se vrať na 1.



Pozn.: Pro omezené hustoty na omezené množině lze volit g konstantní (jako na obrázku). Potom stačí v prvním kroku generovat z rovnoměrného rozdělení. Tato volba ovšem může být velmi neefektivní.

Věta 2. *Hodnota Z z algoritmu 1 představuje výběr z f. Počet iterací předcházejících jejímu vygenerování má geometrické rozdělení s parametrem $\frac{c}{M}$. To znamená, že očekávaný počet iterací pro vygenerování Z je $\frac{M}{c}$.*

Důkaz: Ukážeme, že podmíněné rozdělení $[X \mid U \leq \frac{f^*(X)}{Mg(X)}]$ má hustotu f:

$$\begin{aligned} f\left(x \mid U \leq \frac{f^*(X)}{Mg(X)}\right) &= \frac{\mathbb{P}\left(U \leq \frac{f^*(X)}{Mg(X)} \mid X = x\right) g(x)}{\int \mathbb{P}\left(U \leq \frac{f^*(X)}{Mg(X)} \mid X = x\right) g(x)\,\mu(\mathrm{d}x)} \\ &= \frac{\frac{f^*(x)}{Mg(x)} g(x)}{\int \frac{f^*(x)}{Mg(x)} g(x)\,\mu(\mathrm{d}x)} \\ &= \frac{f^*(x)}{\int f^*(x)\,\mu(\mathrm{d}x)} = f(x). \end{aligned}$$

Využili jsme Bayesovu větu. Z věty o úplné pravděpodobnosti pak dostaneme pravděpodobnost přijetí

$$\begin{aligned} \mathbb{P}\left(U \leq \frac{f^*(X)}{Mg(X)}\right) &= \int \mathbb{P}\left(U \leq \frac{f^*(X)}{Mg(X)} \mid X = x\right) g(x)\,\mu(\mathrm{d}x) \\ &= \int \frac{f^*(x)}{Mg(x)} g(x)\,\mu(\mathrm{d}x) = \frac{c}{M}. \end{aligned}$$

□

Pozn.: U g rovněž není nutné znát normující konstantu. Důležitá je znalost konstanty M, kterou není vždy lehké určit! Určuje efektivitu algoritmu (čím blíž c, tím lépe). Aby bylo možné zamítací metodu použít, tak f/g musí být omezené, to znamená, že g musí mít těžší chvosty než f, např. simulace $N(0,1)$ pomocí Cauchyho rozdělení (ne naopak).

Příklad: simulace náhodného vektoru s FGM (Farlie, Gumbel, Morgenstern) rozdělením, které je dané hustotou $f(x,y) = g_1(x)g_2(y)(1+\lambda(2G_1(x)-1)(2G_2(y)-1))$, kde $-1 \leq \lambda \leq 1$ a g_i resp. G_i jsou hustota resp. distribuční funkce nějakých náhodných veličin ($i = 1,2$). Platí $f(x,y) \leq 2g_1(x)g_2(y)$, tedy $M = 2$, $c = 1$ a pravděpodobnost přijetí je $1/2$.

2.3 Směšovací metody

Na rozdíl od zamítací metody teď cílovou hustotu f „aproximujeme zespodu“. V podstatě využíváme rozkladu $f(x) = \sum p_i f_i(x)$, kde $\sum p_i = 1$ a f_i jsou hustoty, ze kterých umíme simulovat.
Příklad: dvojně exponenciální rozdělení – $f(x) = (f_1(x) + f_1(-x))/2$, kde $f_1(x)$ je hustota exponenciálního.

Obecně jsou směšovací metody založené na vztahu $f(x, y) = f(x \mid y)f(y)$. Pokud je směs Y diskrétní dostáváme předchozí vyjádření. Všimněme si, že nepotřebujeme znát marginální rozdělení $f(x)$.
Příklad: pro t-rozdělení s n stupni volnosti je $X \mid Y = y \sim N(0, n/y)$, $Y \sim \chi^2_n$.

Pro d-rozměrné t-rozdělení s n stupni volnosti a maticí Σ je $X \mid Y = y \sim N_d(\mu, n\Sigma/y)$, $Y \sim \chi^2_n$.

2.4 Monte Carlo integrace

Cílem je vyčíslit integrál $\mathbb{E}_f h(X) = \int_{\mathcal{X}} h(x)f(x)\,\mu(\mathrm{d}x)$. Monte Carlo integrace je založena na simulaci $X_1, \ldots, X_m$ i.i.d. s hustotou f a aproximaci daného integrálu průměrem $\bar{h}_m = \frac{1}{m}\sum_{j=1}^m h(X_j)$, který podle silného zákona velkých čísel konverguje k $\mathbb{E}_f h(X)$.

Víme, že $\operatorname{var} \bar{h}_m = \frac{1}{m} \operatorname{var} h(X_1)$ lze nestranně odhadnout pomocí

$$v_m = \frac{1}{m(m-1)} \sum_{j=1}^m \left(h(x_j) - \bar{h}_m\right)^2.$$

Z centrální limitní věty plyne, že pro velká m má $\frac{\bar{h}_m - \mathbb{E}_f h(X)}{\sqrt{v_m}}$ přibližně $N(0,1)$ rozdělení. Můžeme tak sestrojit přibližné intervaly spolehlivosti pro aproximaci integrálu $\mathbb{E}_f h(X)$. Nasimulovaný výběr $X_1, \ldots, X_m$ se dá použít opakovaně pro různá h.

Ovšem ne vždy je optimální simulovat přímo z f (někdy to ani neumíme). Alternativní přístup je tzv. *importance sampling* založený na vztahu

$$\mathbb{E}_f h(X) = \int_{\mathcal{X}} \left(h(x)\frac{f(x)}{g(x)}\right) g(x)\,\mu(\mathrm{d}x).$$

Pro vyčíslení $\mathbb{E}_f h(X)$ se nyní použije aproximace

$$\frac{1}{m}\sum_{j=1}^m h(X_j)\frac{f(X_j)}{g(X_j)}, \tag{1}$$

kde $X_1, \ldots, X_m$ je náhodný výběr z rozdělení s hustotou g (tzv. *importance hustota*). Pokud $\operatorname{supp} g \supseteq \operatorname{supp} f$ ($f(x) > 0 \Rightarrow g(x) > 0$), tak (1) konverguje k $\mathbb{E}_f h(X)$ podle silného zákona velkých čísel.

Hustota g může být teoreticky libovolná, ale je vhodné, aby měla následující vlastnosti:

1. jednoduše se z ní simuluje (nebo máme k dispozici výběr z g),
2. jednoduše se dá spočítat $g(x)$ pro libovolné x,
3. $\int h(x)^2 \frac{f(x)^2}{g(x)}\,\mu(\mathrm{d}x) < \infty$, což zajistí konečný rozptyl (1). Dá se ukázat (z Cauchyovy-Schwarzovy nerovnosti), že rozptyl (1) je minimální (dokonce roven 0) pro $g(x) \propto h(x)f(x)$, kde $\propto$ značí rovnost až na multiplikativní konstantu. Proto je dobré, když g je blízká $ch(x)f(x)$.
4. Když $\sup f/g = \infty$, tak velký význam je dáván několika málo hodnotám X_j, pro které je podíl f/g velký, proto není dobré, když g má lehké chvosty (např. normální rozdělení). Když $\sup f/g = M < \infty$, tak lze užít zamítací metodu pro simulaci přímo z f.

Pokud neznáme normující konstanty u f a g, tak lze použít *samonormující (selfnormalised) importance sampling*:

$$\frac{\sum_{j=1}^m h(X_j)\frac{f^*(X_j)}{g^*(X_j)}}{\sum_{j=1}^m \frac{f^*(X_j)}{g^*(X_j)}}. \tag{2}$$

Podle silného zákona velkých čísel konverguje (2) k

$$\frac{\int h(x)f^*(x)g(x)/g^*(x)\,\mu(\mathrm{d}x)}{\int f^*(x)g(x)/g^*(x)\,\mu(\mathrm{d}x)} = \frac{(c_f/c_g)\mathbb{E}_f h(X)}{c_f/c_g} = \mathbb{E}_f h(X),$$

kde $c_f = f^*/f$ a $c_g = g^*/g$. Místo nestrannosti teď máme pouze asymptotickou nestrannost (2).

Výhoda MCMC je, že místo generování nezávislých vzorků (což může být těžké) generujeme markovský řetězec. Závislosti v markovském řetězci způsobí větší rozptyl, ovšem ten může být snížen větším počtem vygenerovaných vzorků.

3. Bayesovská statistika

Existuje speciální přednáška *Bayesovské metody* prof. Huškové.

3.1 Bayesova věta

Ve statistice obvykle pracujeme s pozorováním x, které se považuje za realizaci náhodného elementu X v nějakém měřitelném prostoru $(\mathcal{X}, \mathfrak{X})$. Předpokládá se, že X má rozdělení s hustotou $f(x \mid \theta)$ vzhledem k σ-konečné míře μ. O funkci $f(x \mid \theta)$ se mluví také jako o *věrohodnosti (likelihood)*. V klasickém parametrickém přístupu je $\theta \in \Theta \in \mathcal{B}(\mathbb{R}^d)$ vektor neznámých hodnot. Oproti tomu bayesovský přístup považuje θ za d-rozměrný náhodný vektor s hustotou vůči nějaké σ-konečné míře ν. Bayesovský přístup je založen na kombinaci historické informace o parametru θ a pozorovaných dat. Informace o možných hodnotách θ před experimentem určuje *apriorní (prior) hustota* $\pi(\theta)$. *Aposteriorní (posterior) rozdělení* θ za podmínky $X = x$ je pak dáno Bayesovou větou

$$\pi(\theta \mid x) = \frac{f(x \mid \theta)\pi(\theta)}{\int_\Theta f(x \mid \theta)\pi(\theta)\,\nu(\mathrm{d}\theta)} \propto f(x \mid \theta)\pi(\theta),$$

pokud je jmenovatel nenulový. Pro užití zamítací metody nebo importance samplingu znalost $\pi(\theta \mid x)$ až na normující konstantu stačí, tedy pro tyto metody není nutné znát jmenovatel přesně.

Pokud bychom nyní uvažovali nové nezávislé pozorování y spojené s θ, použijeme $\pi(\theta \mid x)$ jako apriorní hustotu pro θ a opětnou aplikací Bayesovy věty dostaneme nové aposteriorní rozdělení. Není těžké ukázat, že výsledek nezávisí na pořadí, v jakém pozorování zpracováváme (viz cvičení).

Můžeme si všimnout, že u $\pi(\theta)$ se nemusí specifikovat normující konstanta (ve výpočtu $\pi(\theta \mid x)$ se zkrátí). Pokud je jmenovatel konečný pro μ-s.v. $x \in \mathcal{X}$ lze užít i nevlastní hustotu $\pi(\theta)$, tj. $\int \pi(\theta)\,\nu(\mathrm{d}\theta) = \infty$. Například neurčité apriorní rozdělení ($\pi(\theta)$ je konstantní) je obvyklou volbou, pokud nemáme představu o apriorním rozdělení. Nevlastní apriorní hustoty mohou vést k divným závěrům: kromě toho, že aposteriorní rozdělení nemusí být vlastní (např. $X \mid \theta \sim R(0, \theta)$, $\pi(\theta) = 1, \theta > 0$, pak $\int f(x \mid \theta)\pi(\theta)\,\mathrm{d}\theta = \int \frac{1}{\theta}\,\mathrm{d}\theta$), tak může záviset na parametrizaci, proto je třeba je používat opatrně. Neurčité apriorní rozdělení, které nezávisí na parametrizaci θ, je dáno tzv. *Jeffreysovou hustotou (Jeffreys' density)* $\pi(\theta) = \sqrt{\det I(\theta)}$, kde

$$I(\theta)_{i,j} = \mathbb{E}_\theta \left(\frac{\partial \log f(x \mid \theta)}{\partial \theta_i} \frac{\partial \log f(x \mid \theta)}{\partial \theta_j} \right)$$

je Fisherova informační matice o θ.

Statistická inference je založena na aposteriorním rozdělení θ. Jakmile ho máme, tak nás většinou zajímají jeho charakteristiky, které shrnují informaci o rozdělení θ, např. *aposteriorní střední hodnota (posterior mean)* je $\mathbb{E}[\theta \mid X = x]$ nebo *aposteriorní rozptyl (posterior variance)* je $\mathrm{var}[\theta \mid X = x]$. K výpočtu aposteriorní střední hodnoty nějaké funkce θ je třeba se vypořádat s (většinou vícerozměrným) integrálem typu $\int h(\theta)\pi(\theta \mid x)\,\nu(\mathrm{d}\theta)$. Ten lze analyticky spočítat jen v některých speciálních případech, proto se musí využít numerické integrace, Monte Carlo integrace, asymptotické aproximace nebo nejčastěji MCMC metod.

3.2 Konjugovaná rozdělení

Definice 1. Řekneme, že P je systém hustot *konjugovaných (conjugate)* s $f(x \mid \theta)$, pokud pro každé $\pi(\theta) \in P$ je $\pi(\theta \mid x) \in P$ pro s.v. x.

Pozn.: Zřejmě systém všech hustot je konjugovaný s libovolným modelem pro pozorování. Definice konjugovaných rozdělení je užitečná pouze pro rozumně velké třídy hustot.

Příklad: Třída normálních rozdělení je konjugovaná pro normální pozorování se známým rozptylem (viz cvičení).

Výhoda konjugovaných rozdělení spočívá v tom, že přechod od apriorního k aposteriornímu je pouze změna parametrů bez nutnosti dalších výpočtů. Neboli aktualizace rozdělení θ je jednoduchá. Na druhou stranu nevýhodou je jisté omezení na volbu apriorního rozdělení. Konjugované rozdělení nemusí být vhodnou volbou pro daný problém. Je třeba volit kompromis mezi realitou a výpočetní zvládnutelností.

Popíšeme možnou konstrukci systému konjugovaných hustot pro exponenciální rodinu (exponential family) rozdělení, tj. rozdělení s hustotou tvaru

$$f(x \mid \theta) = \exp\{\sum_{j=1}^{d} c_j(\theta)T_j(x) + A(\theta) + B(x)\}. \tag{3}$$

Tvrzení 3. *Systém hustot*

$$\pi(\theta) = C(\alpha_1, \ldots, \alpha_d, \beta) \exp\{\sum_{j=1}^{d} \alpha_j c_j(\theta) + \beta A(\theta)\}$$

je konjugovaný s $f(x \mid \theta)$ daným (3).

Důkaz:

$$\pi(\theta \mid x) \propto f(x \mid \theta)\pi(\theta) \propto \exp\{\sum_{j=1}^{d} c_j(\theta)T_j(x) + A(\theta)\} \exp\{\sum_{j=1}^{d} \alpha_j c_j(\theta) + \beta A(\theta)\}$$

$$= \exp\{\sum_{j=1}^{d} (\alpha_j + T_j(x))c_j(\theta) + (\beta + 1)A(\theta)\}.$$

□

Mnoho používaných rozdělení je exponenciálního typu (např. normální, Γ, binomické, Poissonovo). Do exponenciální rodiny nepatří např. rovnoměrné nebo t-rozdělení.

S rostoucí dimenzí a komplikovaností modelu je těžší obdržet konjugovaná rozdělení. Pro víceparametrické modely hraje v MCMC metodách důležitou roli tzv. *podmíněná konjugovanost (conditional conjugacy)*. Pro $\theta = (\theta_1, \ldots, \theta_d)$ položme $\theta_{-i} = (\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_d)$, $i = 1, \ldots, d$. O podmíněné konjugovanosti (conditional conjugacy) parametru θ_i mluvíme, pokud $\pi(\theta_i \mid \theta_{-i})$ a $\pi(\theta_i \mid x, \theta_{-i})$ jsou stejného typu pro $i \in \{1, \ldots, d\}$.

Příklad: Když apriorní rozdělení má nezávislé složky ($\pi(\theta) = \pi_1(\theta_1) \cdots \pi_d(\theta_d)$) a marginální složky $\pi_i(\theta_i)$ jsou konjugované s $f(x \mid \theta_i)$, pak dostáváme podmíněnou konjugovanost.
Pozn.: Existují příklady podmíněné konjugovanosti, kdy složky apriorního rozdělení nejsou nezávislé.

3.3 Hierarchické modely

V hierarchických modelech (hierarchial models) je apriorní rozdělení specifikováno ve více stupních.

Příklad: Klasický model lineární regrese, kde nezávislá pozorování $Y_1, \ldots, Y_n$ mají normální rozdělení, má tvar

$$Y_i = x_{i1}\beta_1 + \cdots + x_{id}\beta_d + \sigma^2\varepsilon_i, \quad i = 1, \ldots, n,$$

kde $x_{i1}, \ldots, x_{id}$ jsou vysvětlující proměnné pro i-té pozorování, $\beta_1, \ldots, \beta_d$ jsou regresní koeficienty, chyby ε_i jsou nezávislé s normovaným normálním rozdělením $N(0, 1)$ a $\sigma^2 > 0$ je rozptyl. Maticový zápis je

$$Y \sim N(X\beta, \tau^{-1}I_n),$$

kde $Y = (Y_1, \ldots, Y_n)^T$, $\beta = (\beta_1, \ldots, \beta_d)^T$, $X = (x_{ij})$ je matice typu $n \times d$, $\tau = \sigma^{-2} > 0$ a I_n je jednotková matice řádu n.

V bayesovském přístupu musí být model doplněn o apriorní rozdělení pro parametry (β, τ). Budeme uvažovat, že β a τ jsou apriorně nezávislé, $\beta \sim N_d(b_0, B_0)$ a $\tau \sim \Gamma(n_0/2, n_0\sigma_0^2/2)$, což je ekvivalentní

tomu, že $n_0\sigma_0^2/\sigma^2 \sim \chi^2_{n_0}$. Aposteriorní rozdělení, které můžeme vyjádřit z Bayesovy věty, je komplikovaného tvaru. Nelze očekávat, že v tomto případě budeme mít konjugovanost. Ovšem dostat podmíněná aposteriorní rozdělení není tak složité (viz cvičení):

$$\beta \mid y, \tau \sim N(b_1, B_1) \quad \text{a} \quad \tau \mid y, \beta \sim \Gamma(n_1/2, n_1\sigma_1^2/2),$$

kde $b_1 = B_1(B_0^{-1}b_0 + \tau X^T y)$, $B_1^{-1} = B_0^{-1} + \tau X^T X$, $n_1 = n_0 + n$ a $n_1\sigma_1^2 = (y - X\beta)^T(y - X\beta) + n_0\sigma_0^2$.

Apriorní rozdělení bylo určeno pomocí hyperparametrů $b_0 \in \mathbb{R}^d$, $B_0 \in \mathbb{R}^{d\times d}$, $n_0 \in \mathbb{N}$, $\sigma_0^2 > 0$, které se pokládají za známé konstanty. Někdy může být vhodné tyto parametry považovat rovněž za náhodné. Apriorní rozdělení je pak dáno ve více krocích. Jako příklad budeme nyní uvažovat dvoustupňový model normální regrese (two-stage normal regression model). Předpokládejme, že vektor β je specifikován pomocí regresního modelu s vysvětlujícími proměnnými $\tilde{x}_{ij}$, $i = 1, \ldots, d$, $j = 1, \ldots, \tilde{d}$ a regresními koeficienty $\tilde{\beta}_1, \ldots, \tilde{\beta}_{\tilde{d}}$. Celý model (viz [15]) má potom maticový tvar

$$\begin{aligned} Y \mid \beta, \tau &\sim N_d(X\beta, \tau^{-1}I_n), \\ \beta \mid \tilde{\beta} &\sim N_{\tilde{d}}(\tilde{X}\tilde{\beta}, C_0), \\ \tau &\sim \Gamma\left(\frac{n_0}{2}, \frac{n_0\sigma_0^2}{2}\right), \\ \tilde{\beta} &\sim N(b_0, B_0), \end{aligned}$$

kde $C_0 \in \mathbb{R}^{d\times d}$, $B_0 \in \mathbb{R}^{\tilde{d}\times\tilde{d}}$, $n_0 \in \mathcal{B}$, $\sigma_0^2 > 0$ jsou známé hyperparametry. Sdružená hustota $(Y, \beta, \tilde{\beta}, \tau)$ proto splňuje

$$f(y, \beta, \tilde{\beta}, \tau) = f(y \mid \beta, \tau)\pi(\beta \mid \tilde{\beta})\pi(\tilde{\beta})\pi(\tau),$$

což bohužel neumožňuje analyticky zvládnout analýzu modelu. Marginální aposteriorní rozdělení parametrů β, $\tilde{\beta}$ a τ není možné analyticky vyjádřit, ale s plnými podmíněnými aposteriorními rozděleními lze pracovat:

$$\begin{aligned} \beta \mid y, \tilde{\beta}, \tau &\sim N_d(b, C_1), \\ \tau \mid y, \beta, \tilde{\beta} &\sim \Gamma\left(\frac{n_1}{2}, \frac{n_1\sigma_1^2}{2}\right), \\ \tilde{\beta} \mid y, \beta, \tau &\sim N_{\tilde{d}}(b_1, B_1), \end{aligned}$$

kde $b = C_1(C_0^{-1}\tilde{X}\tilde{\beta} + \tau X^T y)$, $C_1^{-1} = C_0^{-1} + \tau X^T X$, $n_1 = n + n_0$, $n_1\sigma_1^2 = n_0\sigma_0^2 + (y - X^T\beta)^T(y - X^T\beta)$, $b_1 = B_1(B_0^{-1}b_0 + \tilde{X}^T C_0^{-1}\beta)$ a $B_1 = (B_0^{-1} + \tilde{X}^T C_0^{-1}\tilde{X})^{-1}$. Vidíme tedy, že všechny parametry jsou podmíněně konjugované. První dvě podmíněná rozdělení dostáváme z toho, že podmíněně při $\tilde{\beta}$ jsme v situaci předchozího modelu. Poslední se vypočte ze vztahu $\pi(\tilde{\beta} \mid y, \beta, \tau) \propto \pi(\beta \mid \tilde{\beta})\pi(\tilde{\beta})$. Všimněme si, že $\pi(\tilde{\beta} \mid y, \beta, \tau)$ nezávisí na pozorování y. To je důsledkem hierarchické struktury modelu. Veškerá informace daná pozorováním y se přenáší na $\tilde{\beta}$ prostřednictvím β. Neboli Y a $\tilde{\beta}$ jsou podmíněně nezávislé při daném β.

Pozn.: Hierarchické modely mají zřídka více než tři stupně a obvykle je apriorní rozdělení v nejvyšším stupni neurčité.

4. Příklady MCMC algoritmů

Ukážeme si některé MCMC algoritmy pro simulaci z cílového rozdělení (target distribution) π, o kterém předpokládáme, že má hustotou f vzhledem k nějaké σ-konečné referenční míře μ na měřitelném prostoru $(\mathcal{X}, \mathfrak{X})$. Označme $\mathcal{X}^+ = \{x : f(x) > 0\}$.

Naším cílem je zkonstruovat markovský řetězec, jehož limitní rozdělení bude π. Množina stavů tohoto řetězce bude obecně nespočetná. Roli pravděpodobností přechodu z markovských řetězců se spočetnými stavy bude nyní hrát tzv. *přechodové jádro (transition probability kernel)* $P(x, A)$, které určuje pravděpodobnost přechodu ze stavu x do stavu v množině A. Předpokládáme, že P je markovské jádro na $\mathcal{X}$.

Definice 2. Měřitelné zobrazení $P : \mathcal{X} \times \mathfrak{X} \to [0, 1]$ nazveme *markovským jádrem (Markov kernel)* na $(\mathcal{X}, \mathfrak{X})$, pokud

(i) pro každé $A \in \mathfrak{X}$ je $P(\cdot, A)$ nezáporná měřitelná funkce na $\mathcal{X}$,
(ii) pro každé $x \in \mathcal{X}$ je $P(x, \cdot)$ pravděpodobnostní míra na $\mathfrak{X}$.

Existuje-li limitní rozdělení řetězce, je stacionární (invariantní), tedy splňuje

$$\pi(A) = \int_{\mathcal{X}} P(x, A)\, \pi(\mathrm{d}x) \quad \forall A \in \mathfrak{X}. \tag{4}$$

Postačující podmínka (nikoli však nutná) pro invarianci π je *reverzibilita (reversibility)* Markovova řetězce vzhledem k π:

$$\pi(\mathrm{d}x)P(x, \mathrm{d}y) = \pi(\mathrm{d}y)P(y, \mathrm{d}x),$$

což lze přepsat pomocí hustot jako

$$f(x)p(x, y) = f(y)p(y, x) \quad \forall x, y \in \mathcal{X}, \tag{5}$$

kde $p(x, y)$ je tzv. přechodová hustota. Je to hustota přechodového jádra P, tj. $P(x, A) = \int_A p(x, y)\, \mu(\mathrm{d}y)$. Podmínka (5) je známá jako *detailní podmínka rovnováhy (detailed balance condition)*. Není těžké se přesvědčit, že (4) je splněna, když (5) platí. Ke konstrukci markovského řetězce s daným stacionárním rozdělením proto stačí nalézt přechodové jádro splňující detailní podmínku rovnováhy. To je vždy možné (např. Metropolisův-Hastingsův algoritmus).

4.1 Gibbsův výběrový plán

Budeme předpokládat, že prostor $\mathcal{X}$ má součinový tvar $\prod_{i=1}^{d} \mathcal{X}_i$, nejčastější případ je $\mathcal{X} = \mathbb{R}^d$. Cílové rozdělení přísluší nějakému náhodnému vektoru $(X_1, \ldots, X_d)$.

Algoritmus 2. *Gibbsův výběrový plán (Gibbs sampler)*:

1. zvol počáteční stav $x^{(0)} = (x_1^{(0)}, \ldots, x_d^{(0)}) \in \mathcal{X}^+$, polož $t = 0$,
2. simuluj $x_1^{(t+1)}$ z podmíněného rozdělení $X_1 \mid x_2^{(t)}, \ldots, x_d^{(t)}$,
   simuluj $x_2^{(t+1)}$ z podmíněného rozdělení $X_2 \mid x_1^{(t+1)}, x_3^{(t)}, \ldots, x_d^{(t)}$,
   $\vdots$
   simuluj $x_d^{(t+1)}$ z podmíněného rozdělení $X_d \mid x_1^{(t+1)}, \ldots, x_{d-1}^{(t+1)}$,
3. pokud $t + 1 < T$, tak t zvětši o jedničku a jdi na 2., jinak ukonči algoritmus.

Pozn.: Gibbsův výběrový plán předpokládá, že umíme simulovat ze všech plně podmíněných rozdělení $f(x_i \mid x_{-i})$. I pro vícerozměrné problémy tak můžou být všechny simulace jednorozměrné. Jak jsme již viděli, pokud v bayesovských metodách dochází k podmíněné konjugovanosti, tak z plně podmíněných rozdělení není těžké simulovat, zatímco sdružené rozdělení může být poměrně komplikované a je obtížné z něho simulovat. Jedná se tedy o situaci vhodnou pro Gibbsův výběrový plán.

Vždy d kroků algoritmu dá novou iteraci vektoru. Výstupem je realizace $x^{(0)}, \ldots, x^{(T)}$ Markovova řetězce $X^{(t)}$.

Můžeme si položit otázku, zda $X^{(t)}$ konverguje pro $t \to \infty$ slabě k náhodnému vektoru X bez ohledu na volbu počátečního stavu $x^{(0)}$. Jednoduchý příklad ukazuje, že tomu tak nemusí být.
Příklad: Nechť (X_1, X_2) má rovnoměrné rozdělení na množině $A \cup B$, kde $A = A_1 \times A_2$ a $B = B_1 \times B_2$ jsou obdélníky v $\mathbb{R}^2$. Potom plně podmíněná rozdělení jsou rovnoměrná:

$$X_1 \mid x_2 \sim \begin{cases} R(A_1) & \text{pokud } x_2 \in A_2, \\ R(B_1) & \text{pokud } x_2 \in B_2, \end{cases}$$
$$X_2 \mid x_1 \sim \begin{cases} R(A_2) & \text{pokud } x_1 \in A_1, \\ R(B_2) & \text{pokud } x_1 \in B_1. \end{cases}$$

V závislosti na volbě počátečního rozdělení zůstává řetězec buď v A nebo B, nikdy se nedostane z A do B ani z B do A. Je tedy rozložitelný a limitní rozdělení je rovnoměrné na A nebo B (podle volby $x^{(0)}$).

Za jistých předpokladů se však dá dokázat, že cílové rozdělení je stacionární pro markovský řetězec $X^{(t)}$.

Přechodová hustota (pokud existuje) je rovna součinu plně podmíněných rozdělení f:

$$p(x,y) = \prod_{i=1}^{d} f(y_i \mid y_1, \ldots, y_{i-1}, x_{i+1}, \ldots, x_d).$$

Příslušné přechodové jádro P se nazývá *Gibbsovo jádro (Gibbs kernel)*.

Lemma 4. *(Besag [1]) Nechť $f_i(x_i) > 0$ pro každé $i \in \{1, \ldots, d\}$ implikuje $f(x) > 0$, kde $x = (x_1, \ldots, x_d)$ a f_i jsou marginály f. Potom*

$$\frac{f(y)}{f(x)} = \prod_{i=1}^{d} \frac{f(y_i \mid y_1, \ldots, y_{i-1}, x_{i+1}, \ldots, x_d)}{f(x_i \mid y_1, \ldots, y_{i-1}, x_{i+1}, \ldots, x_d)}, \quad x \in \mathcal{X}^+.$$

Důkaz: Z definice podmíněného rozdělení máme:

$$f(y) = f(y_d \mid y_1, \ldots, y_{d-1}) f(y_1, \ldots, y_{d-1}),$$
$$f(y_1, \ldots, y_{d-1}, x_d) = f(x_d \mid y_1, \ldots, y_{d-1}) f(y_1, \ldots, y_{d-1}),$$

a odtud

$$f(y) = \frac{f(y_d \mid y_1, \ldots, y_{d-1})}{f(x_d \mid y_1, \ldots, y_{d-1})} f(y_1, \ldots, y_{d-1}, x_d).$$

Dále

$$f(y_1, \ldots, y_{d-1}, x_d) = f(y_{d-1} \mid y_1, \ldots, y_{d-2}, x_d) f(y_1, \ldots, y_{d-2}, x_d),$$
$$f(y_1, \ldots, y_{d-2}, x_{d-1}, x_d) = f(x_{d-1} \mid y_1, \ldots, y_{d-2}, x_d) f(y_1, \ldots, y_{d-2}, x_d),$$

což znamená, že

$$f(y) = \frac{f(y_d \mid y_1, \ldots, y_{d-1})}{f(x_d \mid y_1, \ldots, y_{d-1})} \frac{f(y_{d-1} \mid y_1, \ldots, y_{d-2}, x_d)}{f(x_{d-1} \mid y_1, \ldots, y_{d-2}, x_d)} f(y_1, \ldots, y_{d-2}, x_{d-1}, x_d).$$

Takto postupně dostaneme požadované tvrzení.

□

Z lemmatu plyne, že $f(x)p(x,y) \neq f(y)p(y,x)$, tedy Gibbsův výběrový plán nedává reverzibilní řetězec. Můžeme však uvažovat přechodové jádro s hustotou $p^*(y,x) = \prod_{i=1}^{d} f(x_i \mid y_1, \ldots, y_{i-1}, x_{i+1}, \ldots, x_d)$ odpovídající simulaci „odzadu“ – od d-té souřadnici k první. Potom se dá ukázat invariance f vzhledem k P. Proto existuje-li limitní rozdělení řetězce, je to nutně π.

Věta 5. *Za výše uvedených předpokladů je π invariantní míra vzhledem ke Gibbsovu jádru, tedy splňuje (4).*

Důkaz: Využijeme toho, že $f(y)p^*(y,x) = f(x)p(x,y)$, viz lemma 4. Pro $A \in \mathfrak{X}$ je

$$\begin{aligned}\int P(x,A)\,\pi(\mathrm{d}x) &= \int \int_A p(x,y)\,\mu(\mathrm{d}y)\,\pi(\mathrm{d}x) = \int \int_A p(x,y) f(x)\,\mu(\mathrm{d}y)\,\mu(\mathrm{d}x) \\ &= \int_A \int f(x)p(x,y)\,\mu(\mathrm{d}x)\,\mu(\mathrm{d}y) = \int_A \int f(y)p^*(y,x)\,\mu(\mathrm{d}x)\,\mu(\mathrm{d}y) = \int_A f(y)\,\mu(\mathrm{d}y) = \pi(A).\end{aligned}$$

□

Algoritmu 2 se také říká *systematický (systematic)* Gibbsův výběrový plán, protože složky vektoru se procházejí systematicky od první k poslední. Modifikací tohoto algoritmu je tzv. *náhodné procházení (random scan)*, kdy složky vektoru, ze kterých simulujeme, vybíráme náhodně (každou s pravděpodobností $1/d$).

Algoritmus 3. *Náhodné procházení v Gibbsově výběrovém plánu (random scan Gibbs sampler)*:

1. zvol počáteční stav $x^{(0)} = (x_1^{(0)}, \ldots, x_d^{(0)})$, polož $t = 0$,
2. vygeneruj k z rovnoměrného rozdělení na množině $\{1, \ldots, d\}$ a simuluj $x_k^{(t+1)}$ z podmíněného rozdělení $X_k \mid x_1^{(t)}, \ldots, x_{k-1}^{(t)}, x_{k+1}^{(t)}, \ldots, x_d^{(t)}$, polož $x_j^{(t+1)} = x_j^{(t)}$ pro $j \neq k$,
3. pokud $t + 1 < T$, tak t zvětši o jedničku a jdi na 2., jinak ukonči algoritmus.

Tento algoritmus již vede na reverzibilní markovský řetězec.

4.2 Metropolisův-Hastingsův algoritmus

Buď Q markovské jádro na $\mathcal{X}$. Nechť $Q(x, \mathrm{d}y) = q(x, y)\mu(\mathrm{d}y)$ pro nějaké q a $Q(x, \mathcal{X}^+) = 1$ pro $x \notin \mathcal{X}^+$. Funkce q se nazývá *návrhová hustota (proposal density)*.

Definujme *pravděpodobnost přijetí návrhu (proposal acceptance probability)* jako

$$\alpha(x, y) = \begin{cases} \min\left\{\frac{f(y)q(y,x)}{f(x)q(x,y)}, 1\right\} & \text{pro } f(x)q(x, y) > 0, \\ 1 & \text{jinak.} \end{cases}$$

Algoritmus 4. *Metropolisův-Hastingsův algoritmus (Metropolis-Hastings algorithm)*:

1. Zvol $x^{(0)} \in \mathcal{X}^+$ libovolně, polož $t = 0$.
2. Generuj y z rozdělení $Q(x^{(t)}, \cdot)$. S pravděpodobností $\alpha(x^{(t)}, y)$ je kandidát y přijat ($x^{(t+1)} = y$), s pravděpodobností $1 - \alpha(x^{(t)}, y)$ je zamítnut ($x^{(t+1)} = x^{(t)}$).
3. Pokud $t + 1 < T$, tak zvětši t o jedničku a jdi na 2., jinak ukonči algoritmus.

Pozn.: Vygenerovaný řetězec skoro jistě neopustí $\mathcal{X}^+$, protože když $f(y) = 0$, tak $\alpha(x, y) = 0$.

Algoritmus závisí na f jen přes podíl $f(y)/f(x)$, proto není nutné znát normující konstantu u hustoty f. Podobně není nutné znát normující konstantu u q. Další výhoda Metropolisova-Hastingsova algoritmu spočívá v tom, že simulujeme z rozdělení q, které si volíme libovolně. Na rozdíl od Gibbsova výběrového plánu tedy nemusíme znát podmíněné hustoty cílového rozdělení (a umět z nich generovat). Nevýhodou je, že pokud q je nevhodně zvoleno může být pravděpodobnost přijetí návrhu často malá (tudíž počet zamítnutí je velký a řetězec dlouho zůstává v jednom stavu), což snižuje efektivitu algoritmu.

Definujme

$$p_0(x, y) = \begin{cases} q(x, y)\alpha(x, y) & \text{pro } x \neq y, \\ 0 & \text{pro } x = y. \end{cases}$$

Potom $p_0(x, y)f(x) = p_0(y, x)f(y)$, protože $\alpha(x, y) < 1$ znamená $\alpha(y, x) = 1$ a naopak. Položme $r(x) = 1 - \int p_0(x, y)\,\mu(\mathrm{d}y)$ pravděpodobnost, že $X^{(t)}$ neopustí x v jednom kroku. Potom přechodové (Metropolisovo-Hastingsovo) jádro je $P(x, \mathrm{d}y) = p_0(x, y)\mu(\mathrm{d}y) + r(x)\delta_x(\mathrm{d}y)$, kde δ_x je Diracova míra v bodě x, tj.

$$\delta_x(A) = \begin{cases} 1, & x \in A, \\ 0, & x \notin A. \end{cases}$$

Věta 6. *Cílové rozdělení π s hustotou f je invariantní pro Metropolisovo-Hastingsovo jádro.*

Důkaz: Pro $A \in \mathfrak{X}$ je

$$\begin{aligned} \int P(x, A)\,\pi(\mathrm{d}x) &= \int P(x, A)f(x)\,\mu(\mathrm{d}x) = \int \left(\int_A p_0(x, y)\,\mu(\mathrm{d}y)\right) f(x)\,\mu(\mathrm{d}x) + \int r(x)\delta_x(A)f(x)\,\mu(\mathrm{d}x) \\ &= \int_A \left(\int p_0(x, y)f(x)\,\mu(\mathrm{d}x)\right)\,\mu(\mathrm{d}y) + \int_A r(x)f(x)\,\mu(\mathrm{d}x) \\ &= \int_A \left(p_0(y, x)f(y)\,\mu(\mathrm{d}x)\right)\,\mu(\mathrm{d}y) + \int_A r(x)f(x)\,\mu(\mathrm{d}x) \\ &= \int_A (1 - r(y))f(y)\,\mu(\mathrm{d}y) + \int_A r(x)f(x)\,\mu(\mathrm{d}x) = \int_A f(y)\,\mu(\mathrm{d}y) = \pi(A). \end{aligned}$$

□

Příklady:

(a) *náhodná procházka (random walk)*: $\mathcal{X} = \mathbb{R}^d$, $q(x, y) = q_0(y - x)$. Tedy pro dané x je návrh $Y = x + Z$, kde Z má hustotu q_0. Typická volba pro q_0 je hustota d-rozměrného normálního rozdělení nebo rovnoměrné rozdělení na d-rozměrné kouli. Je-li $q_0(x) = q_0(-x)$, mluvíme o symetrické náhodné procházce (symmetric random walk) a $\alpha(x, y) = \min\{f(y)/f(x), 1\}$, tedy kandidáta s větší hodnotou cílové hustoty přijmeme vždy. Není proto nutné vyčíslovat q. Algoritmus se symetrickou návrhovou funkci ($q(x, y) = q(y, x)$) se někdy nazývá krátce Metropolisův. Byl poprvé uvažován v článku Metropolise a kol. (1953), kde lze rovněž nalézt heuristický důkaz konvergence. Přínos Hastingse (1970) spočívá v zobecnění na nesymetrické návrhy, rigorózním důkazu konvergence a zaměření na statistické problémy.

(b) *multiplikativní náhodná procházka (multiplicative random walk)*: $\mathcal{X} = \mathbb{R}^d$, $q(x,y) = \frac{1}{y} q_0\left(\log \frac{y}{x}\right)$. Odpovídá situaci, kdy návrh je $Y = x\mathrm{e}^Z$, kde Z má hustotu q_0.

(c) *nezávislý výběr (independent sampler)*: $q(x,y) = q_0(y)$ pro všechna $x \in \mathcal{X}$ (návrhová hustota nezávisí na současném stavu). Definujme $w(x) = f(x)/q_0(x)$, potom je $\alpha(x,y) = \min\{w(y)/w(x), 1\}$. Je-li $q_0 = f$, je $w = 1$ a algoritmus dává náhodný výběr z rozdělení s hustotou f. Situace připomíná importance sampling, každému stavu je přiřazena váha (podíl cílové a pomocné hustoty). Návrhy s větší váhovou funkcí jsou častěji přijímány. Opět je vhodné, aby váhová funkce byla omezená (jinak řetězec může po dlouhou dobu zůstávat ve stavech s velkou váhou) a co nejbližší konstantní jedničce. Aby se zajistila omezenost w je dobré volit q_0 s těžkými chvosty (např. mnohorozměrné t-rozdělení pro $\mathcal{X} = \mathbb{R}^d$).

(d) *autoregresní řetězec (autoregressive chain)*: $\mathcal{X} = \mathbb{R}^d$, $q(x,y) = q_0(y - a - b(x-a))$, kde $a \in \mathbb{R}^d$ a $b \in \mathbb{R}$ jsou pevné. Návrh je $Y = a + b(x-a) + Z$, kde Z má hustotu q_0. Jedná se o prostředníka mezi náhodnou procházkou ($b = 1$) a nezávislým výběrem ($a = b = 0$). Pro $0 < b < 1$ tato strategie stahuje současný stav směrem k a. Volba $b < 0$ vede na záporné korelace v řetězci, což snižuje rozptyl odhadů středních hodnot funkcí stavů.

(e) *zamítací výběr (rejection sampler)*: Připomeňme, že při simulování zamítací metodou (algoritmu 1) potřebujeme, aby platilo $f(x) \le Mg(x)$ pro všechna x a nějakou konstantu M. Často je konstanta M tak velká, že pravděpodobnost přijetí je velmi malá. Pokud neplatí $f(x) \le Mg(x)$ a použijeme zamítací metodu, dostáváme výběr z rozdělení s hustotou $q_0(x) \propto \min(f(x), Mg(x))$. Nyní můžeme užít Metropolisův-Hastingsův nezávislý výběr s q_0. Označme $C = \{x : f(x) \le Mg(x)\}$, pravděpodobnost přijetí je potom

$$\alpha(x,y) = \min\{\frac{f(y)q_0(x)}{f(x)q_0(y)}, 1\} = \begin{cases} 1 & \text{pro } x \in C, \\ \frac{Mg(x)}{f(x)} & \text{pro } x \notin C,\ y \in C, \\ \min\{\frac{f(y)g(x)}{f(x)g(y)}, 1\} & \text{pro } x \notin C,\ y \notin C. \end{cases}$$

Hlavní část celého algoritmu se tedy skládá ze dvou kroků. V prvním se generuje návrh z rozdělení s hustotou úměrnou $\min(f(y), Mg(y))$ pomocí zamítacího výběru a v druhém se tento návrh přijme s pravděpodobností $\alpha(x,y)$.

(f) *Langevinův algoritmus (Langevin algorithm)*: $\mathcal{X} = \mathbb{R}^d$,

$$q(x,y) = \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left\{-\frac{\|y - x - \sigma^2 \nabla \log f(x)/2\|^2}{2\sigma^2}\right\},$$

kde σ je vhodný parametr a ∇ označuje gradient. Algoritmus užívá informace o gradientu cílové hustoty f, návrh není centrován v současném stavu, ale je nasměrován tam, kde bude pravděpodobně cílová hustota nabývat vyšší hodnoty.

(g) *hybridní algoritmus (hybrid algorithm)*: kombinace Metropolisova-Hastingsova algoritmu a Gibbsova výběrového plánu. Dejme tomu, že chceme simulovat náhodný vektor (X_1, X_2), přitom simulace z $X_1 \mid X_2$ je jednoduchá, ale z $X_2 \mid X_1$ nelze přímo simulovat. Místo toho použijeme pro aktualizaci druhé složky Metropolisovo-Hastingsovo jádro se stacionárním rozdělením $X_2 \mid X_1$. Tento hybridní algoritmus se někdy také označuje jako *Metropolis-within-Gibbs algorithm*.

5. Markovovy řetězce

V této kapitole zopakujeme základní vlastnosti markovských řetězců s diskrétní množinou stavů a poté přejdeme k situaci s obecnou množinou stavů. Stavový prostor budeme opět značit $(\mathcal{X}, \mathfrak{X})$.

5.1 Diskrétní množina stavů

Předpokládejme, že množina stavů $\mathcal{X} = S$ je nejvýše spočetná.

Definice 3. Mějme posloupnost náhodných veličin $\{X_n, n \in \mathbb{N}_0\}$ definovaných na pravděpodobnostním prostoru $(\Omega, \mathcal{A}, \mathbb{P})$ a nabývajících hodnot v prostoru S. Řekneme, že $\{X_n, n \in \mathbb{N}_0\}$ je *Markovův řetězec (Markov chain)* s množinou stavů S, jestliže platí

$$\mathbb{P}(X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1} \ldots, X_0 = i_0) = \mathbb{P}(X_{n+1} = j \mid X_n = i)$$

pro všechna $n \in \mathbb{N}_0$, $i, j, i_{n-1}, \ldots, i_0 \in S$, pro která $P(X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) > 0$. Když navíc $\mathbb{P}(X_{n+m+1} = j \mid X_{n+m} = i) = \mathbb{P}(X_{n+1} = j \mid X_n = i)$ pro libovolné $m \in \mathbb{N}$, $n \in \mathbb{N}_0$ a $i, j \in S$, tak mluvíme o *homogenním (homogenous)* řetězci.

Uvažujme homogenní Markovův řetězec $\{X_n, n \in \mathbb{N}_0\}$ a připomeňme základní definice a vlastnosti z přednášky *Náhodné procesy*, které rozšíříme o některé další poznatky.

Pro konečně rozměrná rozdělení $\{X_n, n \in \mathbb{N}_0\}$ platí

$$\mathbb{P}(X_0 = i_0, \ldots, X_n = i_n) = p_{i_0} p_{i_0 i_1} \cdots p_{i_{n-1} i_n}, \tag{6}$$

kde $p_j = \mathbb{P}(X_0 = j)$ jsou *počáteční pravděpodobnosti (initial probabilities)* a $p_{ij} = \mathbb{P}(X_{n+1} = j \mid X_n = i)$ jsou *pravděpodobnosti přechodu (transition probabilities)*.

Pravděpodobnosti přechodu řádu n jsou

$$p_{ij}^{(n)} = \mathbb{P}(X_n = j \mid X_0 = i), \quad i, j \in S.$$

Chapmanova-Kolmogorova rovnost má tvar

$$p_{ij}^{(n)} = \sum_{k \in S} p_{ik}^{(m)} p_{kj}^{(n-m)}, \quad i, j \in S, n, m \in \mathbb{N}_0, m \leq n.$$

Stav j řetězce $\{X_n\}$ je:

- *trvalý (recurrent)*, pokud $\mathbb{P}(\tau_{jj} < \infty) = 1$, kde $\tau_{jj} = \min\{n > 0 : X_n = j \mid X_0 = j\}$ je doba prvního návratu do j. Ekvivalentně, když platí $\sum_{n=0}^{\infty} p_{jj}^{(n)} = \infty$.
- *trvalý nenulový (positive recurrent)*, pokud $\mathbb{E}\tau_{jj} < \infty$.
- *neperiodický (aperiodic)*, pokud největší společný dělitel prvků množiny $\{n > 0 : p_{jj}^{(n)} > 0\}$ je roven 1.

Řetězec $\{X_n\}$ je *nerozložitelný (irreducible)*, když pro všechna $i, j \in S$ existuje $n \in \mathbb{N}$ tak, že $p_{ij}^{(n)} > 0$. Říkáme, že nerozložitelný řetězec je *ergodický (ergodic)*, pokud nějaký stav $j \in S$ (a potom všechny stavy) je trvalý nenulový a neperiodický.

V nerozložitelném řetězci je existence trvalého nenulového stavu ekvivalentní existenci stacionárního rozdělení. Pravděpodobnostní rozdělení π se nazývá *stacionární rozdělení (stationary distribution)*, jestliže $\pi_j = \sum_{i \in S} \pi_i p_{ij}^{(n)}$ pro všechna $j \in S$ a $n \in \mathbb{N}$. Pokud existuje $\{\eta_j \geq 0, j \in S\}$ splňující $\eta_j = \sum_{i \in S} \eta_i p_{ij}^{(n)}$ pro všechna $j \in S$ a $n \in \mathbb{N}$, pak se nazývá *invariantní míra (invariant measure)*. Pokud je počáteční rozdělení stacionární, tak $\{X_n\}$ je striktně stacionární proces (speciálně X_n má rozdělení π pro každé n).

Jsou-li v nerozložitelném všechny stavy trvalé nenulové a neperiodické (ergodický řetězec), tak stacionární rozdělení existuje, je jediné a splňuje

$$\lim_{n \to \infty} p_{ij}^{(n)} = \pi_j \quad \text{pro všechna } i, j \in S,$$

neboli je *limitní (limitting)*. Dokonce platí

$$\lim_{n \to \infty} \sum_{j \in S} |p_{ij}^{(n)} - \pi_j| = 0.$$

Navíc pro ergodický řetězec platí tzv. ergodická věta. Je-li $\mathbb{E}_\pi |h(X)| = \sum_{i \in S} |h(i)| \pi_i < \infty$, potom

$$\lim_{n \to \infty} \bar{h}_n = \mathbb{E}_\pi h(X) \quad \mathbb{P}\text{-s.j.},$$

kde h je reálná měřitelná funkce na S, $\bar{h}_n = \frac{1}{n} \sum_{i=1}^{n} h(X_i)$ a $\mathbb{E}_\pi h(X) = \sum_{j \in S} h(j) \pi_j$.

Řekneme, že markovský řetězec se stacionárním rozdělením π je *reverzibilní (reversible)*, splňuje-li

$$\pi_i p_{ij} = \pi_j p_{ji}, \quad \forall i, j \in S.$$

Ergodický řetězec je *geometricky ergodický (geometrically ergodic)*, existuje-li $0 \leq \lambda < 1$ a funkce V tak, že

$$\sum_{j \in S} |p_{ij}^{(n)} - \pi_j| \leq V(i)\lambda^n \quad \forall i \in S, n \in \mathbb{N}. \tag{7}$$

Nejmenší λ, pro něž existuje funkce V splňující (7), se značí λ^* a nazývá se *geometrický řád konvergence (geometric rate of convergence)*.

Nechť existují vlastní čísla $\lambda_0, \lambda_1, \ldots$ a vlastní levé vektory $e_0, e_1, \ldots$ matice $P = (p_{ij})$, tj.

$$\sum_{i \in S} e_k(i) p_{ij} = \lambda_k e_k(j).$$

Zřejmě $\lambda_0 = 1$ a $e_0 = \pi$. Pro geometricky ergodický řetězec jsou $\{\lambda_i, i \in \mathbb{N}\}$ stejnoměrně odražené od ± 1 a platí $\lambda^* = \sup_{i \in \mathbb{N}} |\lambda_i| < 1$, tj. λ^* je druhé největší vlastní číslo – SLEM (second largest eigenvalue modulus). Toto tvrzení je důsledkem Perronovy-Frobeniovy věty, která udává tvar matice P^n.

Za předpokladu geometrické ergodicity platí centrální limitní věta pro ergodické průměry $\bar{h}_n = \frac{1}{n}\sum_{i=1}^n h(X_i)$:

$$\sqrt{n}\left(\bar{h}_n - \mathbb{E}_\pi h(X)\right) \xrightarrow[n\to\infty]{\mathcal{D}} N(0, \sigma^2),$$

kde pro limitní rozptyl platí $\sigma^2 \leq \frac{1+\lambda^*}{1-\lambda^*} \operatorname{var}_\pi h(X)$.

5.2 Obecná množina stavů

Nechť $\mathcal{X}$ je obecná množina a σ-algebra $\mathfrak{X}$ je spočetně generovaná. Podrobnosti k teorii markovských řetězců s obecným prostorem stavů lze nalézt v [17].

Mnoho výsledků pro diskrétní prostor stavů se dá zobecnit na situaci s obecným prostorem stavů. Místo pravděpodobností přechodu je třeba používat markovská přechodová jádra. Následující věta je zobecnění vztahu (6), ve kterém jsou konečně rozměrná rozdělení vyjádřena pomocí pravděpodobností přechodu.

Věta 7. *Je dáno markovské jádro P na $(\mathcal{X}, \mathfrak{X})$ a pravděpodobnostní rozdělení ϱ na $\mathfrak{X}$. Existuje náhodný proces $\{X_n, n \in \mathbb{N}_0\}$ takový, že*

$$\mathbb{P}(X_0 \in A_0, \ldots, X_n \in A_n) = \int_{A_0} \cdots \int_{A_{n-1}} P(y_{n-1}, A_n) P(y_{n-2}, \mathrm{d}y_{n-1}) \cdots P(y_0, \mathrm{d}y_1) \varrho(\mathrm{d}y_0) \tag{8}$$

pro všechna $n \in \mathbb{N}_0$, $A_0, \ldots, A_n \in \mathfrak{X}$.

Důkaz: (náznak) Projektivnost se ověří položením $A_n = \mathcal{X}$. Existence plyne z Danielovy-Kolmogorovovy věty.

□

Definice 4. Řekneme, že náhodný proces $\{X_n\}$ s obecnou množinou stavů $\mathcal{X}$ je *homogenní Markovův řetězec (homogenous Markov chain)* s přechodovým jádrem P a počátečním rozdělením ϱ, pokud jeho konečně rozměrná rozdělení splňují (8) pro každé $n \in \mathbb{N}_0$ a pro všechna $A_0, \ldots, A_n \in \mathfrak{X}$.

Pro libovolnou měřitelnou funkci f na $\mathcal{X}$ a σ-konečnou míru μ na $\mathfrak{X}$ budeme psát

$$Pf(x) = \int_{\mathcal{X}} f(y)\, P(x, \mathrm{d}y), \quad \mu P(A) = \int_{\mathcal{X}} P(x, A)\, \mu(\mathrm{d}x),$$

neboli Pf je funkce na $\mathcal{X}$ a μP je míra na $\mathfrak{X}$.

Markovský řetězec lze ekvivalentně zavést pomocí markovské vlastnosti.

Tvrzení 8. *Nechť $\{X_n\}$ je homogenní markovský řetězec generovaný přechodovým jádrem P a h je omezená měřitelná funkce na $\mathcal{X}$. Potom pro každé $n \in \mathbb{N}_0$ platí*

$$\mathbb{E}[h(X_{n+1}) \mid X_n, \ldots, X_0] = Ph(X_n).$$

Pozn.: Pravá strana je vlastně $\mathbb{E}[h(X_{n+1}) \mid X_n]$.

Definice 5. Položme $P^0(x, A) = \delta_x(A)$. *Přechodové jádro n-tého řádu (n-step transition probability kernel)* je dáno induktivně vztahem

$$P^n(x, A) = \int_{\mathcal{X}} P(y, A)\, P^{n-1}(x, \mathrm{d}y), \quad n \in \mathbb{N}.$$

Tvrzení 9. *(Chapmanova-Kolmogorovova rovnost) Pro $n, m \in \mathbb{N}_0$ a $m \leq n$ platí*

$$P^n(x, A) = \int_{\mathcal{X}} P^{n-m}(y, A)\, P^m(x, \mathrm{d}y).$$

Důkaz: V (8) stačí položit $\varrho = \delta_x$, $A_i = \mathcal{X}$, $i = 0, \ldots, n-1$ a $A_n = A$. Definice P^m a P^{n-m} se použije pro prvních m a posledních $n - m$ integrandů.

□

Definice 6. Pravděpodobnostní rozdělení π na $\mathfrak{X}$ nazveme *limitní rozdělení (limitting distribution)* Markovova řetězce $\{X_n\}$ generovaného P, jestliže

$$\lim_{n\to\infty} P^n(x, A) = \pi(A) \quad \text{pro } \pi\text{-s.v. } x \in \mathcal{X}, \text{ pro všechna } A \in \mathfrak{X}.$$

Pro dané počáteční rozdělení ϱ je $\mathbb{P}(X_n \in A) = \int_{\mathcal{X}} P^n(x, A)\, \varrho(\mathrm{d}x)$, tedy $\mathbb{P}(X_n \in A) \underset{n\to\infty}{\longrightarrow} \pi(A)$.

Na definici stacionárního rozdělení jsme již narazili, viz (4).

Definice 7. Řekneme, že σ-konečná míra π na $\mathfrak{X}$ se nazývá *invariantní (invariant)*, jestliže $\pi = \pi P$, tj.

$$\pi(A) = \int_{\mathcal{X}} P(x, A)\, \pi(\mathrm{d}x) \quad \forall A \in \mathfrak{X}.$$

Pokud je π pravděpodobnostní rozdělení, nazývá se *stacionární (stationary)* rozdělení Markovova řetězce s přechodovým jádrem P.

Pokud zvolíme stacionární rozdělení π jako počáteční, pak X_n je striktně stacionární proces.

Tvrzení 10. *Je-li π limitní rozdělení, potom je stacionární.*
Důkaz: Pro $A \in \mathfrak{X}$ je

$$\pi(A) = \lim_{n\to\infty} P^n(x, A) = \lim_{n\to\infty} \int_{\mathcal{X}} P(y, A)\, P^{n-1}(x, \mathrm{d}y) = \int_{\mathcal{X}} P(y, A)\, \pi(\mathrm{d}y) = \pi P(A).$$

□

Definice 8. Markovský řetězec generovaný přechodovým jádrem P je *reverzibilní (reversible)* vzhledem k π, jestliže pro každé $A, B \in \mathfrak{X}$ platí

$$\int_A P(x, B)\, \pi(\mathrm{d}x) = \int_B P(x, A)\, \pi(\mathrm{d}x). \tag{9}$$

Tvrzení 11. *Je-li Markovův řetězec reverzibilní vzhledem k π, potom π je stacionární rozdělení.*
Důkaz: Stačí položit $A = \mathcal{X}$ v (9).

□

Definice 9. Markovský čas $\tau_A = \min\{n \in \mathbb{N} : X_n \in A\}$ se nazývá *doba prvního návratu do A (first return time on A)*. Označme $L(x, A) = \mathbb{P}(\tau_A < \infty \mid X_0 = x)$ pravděpodobnost návratu do A.

Definice 10. Nechť φ je pravděpodobnostní míra na $\mathfrak{X}$. Řekneme, že markovský řetězec $\{X_n\}$ je *φ-nerozložitelný (φ-irreducible)*, jestliže pro každé $x \in \mathcal{X}$ a $A \in \mathfrak{X}$ s $\varphi(A) > 0$ je $P^n(x, A) > 0$ pro nějaké $n \in \mathbb{N}$, neboli $L(x, A) > 0$.

Příklad: Řetězec se spočetně mnoha stavy, který není nerozložitelný v diskrétní definici, může být φ-nerozložitelný. Uvažujme náhodnou procházku s absorpčním stavem 0, tedy $p_{00} = 1$, $p_{i,i+1} = p \in (0, 1)$

a $p_{i,i-1} = 1 - p$ pro $i \in \mathbb{N}$, $p_{ij} = 0$ jinak. Potom v diskrétní definici jsou stavy $1, 2, \ldots$ přechodné a stav 0 je nenulový trvalý (absorpční). Ve spojité definici je řetězec φ-nerozložitelný pro $\varphi = \delta_0$.

Definice 11. Pro $0 < \varepsilon < 1$ a Markovův řetězec s přechodovým jádrem P definujeme *rezolventu (resolvent)* jako $K_\varepsilon(x, A) = (1 - \varepsilon) \sum_{n=0}^{\infty} \varepsilon^n P^n(x, A)$.

Věta 12. *Nechť Markovův řetězec $\{X_n\}$ je φ-nerozložitelný, potom existuje pravděpodobnostní míra ψ na $\mathfrak{X}$ tak, že*

(i) $\{X_n\}$ je ψ-nerozložitelný,

(ii) pro libovolnou pravděpodobnostní míru φ' na $\mathfrak{X}$ platí: $\{X_n\}$ je φ'-nerozložitelný právě tehdy, když φ' je absolutně spojitá k ψ.

Důkaz: Buď $A \in \mathfrak{X}$ a $\psi(A) = \int_{\mathcal{X}} K_{\frac{1}{2}}(y, A)\, \varphi(\mathrm{d}y)$. Označme $\bar{A}(k) = \{y : \sum_{n=1}^{k} P^n(y, A) > \frac{1}{k}\}$. Pro $y \in \mathcal{X}$ takové, že $y \notin \bar{A}(k)$ pro žádné k, je $\sum_{n=1}^{k} P^n(y, A) \leq \frac{1}{k}$ pro každé $k \in \mathbb{N}$, tedy $P^n(y, A) = 0$ pro každé $n \in \mathbb{N}$. Proto

$$\psi(A) = \int_{\mathcal{X}} \sum_{n=0}^{\infty} P^n(x, A) 2^{-(n+1)}\, \varphi(\mathrm{d}x) = \int_{\cup_k \bar{A}(k)} \sum_{n=0}^{\infty} P^n(x, A) 2^{-(n+1)}\, \varphi(\mathrm{d}x).$$

Tedy $\psi(A) > 0$ implikuje existenci k takového, že $\varphi(\bar{A}(k)) > 0$. Potom (z φ-nerozložitelnosti) je

$$\sum_{n=1}^{k} P^{m+n}(x, A) = \int_{\mathcal{X}} \sum_{n=1}^{k} P^n(y, A)\, P^m(x, \mathrm{d}y) \geq \frac{1}{k} P^m(x, \bar{A}(k)) > 0$$

pro nějaké m, a tudíž je řetězec ψ-nerozložitelný.

Nechť $\{X_n\}$ je φ'-nerozložitelný. Je-li $\varphi'(A) > 0$, je $\sum_{n=0}^{\infty} P^n(y, A) > 0$ pro každé $y \in \mathcal{X}$, tedy $\varphi' \ll \psi$.

Nechť $\{X_n\}$ je ψ-nerozložitelný a $\varphi' \ll \psi$. Je-li $\varphi'(A) > 0$, je $\psi(A) > 0$ a z ψ-nerozložitelnosti plyne $K_{\frac{1}{2}}(x, A) > 0$ pro každé $x \in \mathcal{X}$, tudíž $\{X_n\}$ je φ'-nerozložitelný.

□

Pozn.: Když budeme mluvit o ψ-nerozložitelném řetězci, máme na mysli, že řetězec je φ-nerozložitelný pro nějaké φ a míra ψ je maximální ve smyslu předchozí věty.

Pokud v nerozložitelném řetězci existuje stacionární rozdělení, tak je jediné.

Věta 13. *Nechť π je stacionární rozdělení a existuje míra φ tak, že $\{X_n\}$ je φ-nerozložitelný. Potom $\{X_n\}$ je π-nerozložitelný a π je jediné stacionární rozdělení.*

Důkaz: [24]

□

Definice 12. Markovův řetězec, který je ψ-nerozložitelný a ve kterém existuje stacionární rozdělení, se nazývá *nenulový (positive)*.

Definice 13. Markovský řetězec $\{X_n\}$ je *periodický (periodic)*, jestliže existuje $q \in \mathbb{N}$, $q > 1$ a neprázdné disjunktní množiny $A_0, \ldots, A_{q-1}, A_q = A_0 \in \mathfrak{X}$ tak, že $P(x, A_{i+1}) = 1$ pro každé $x \in A_i$, $i \in \{0, \ldots, q-1\}$. V opačném případě je $\{X_n\}$ *neperiodický (aperiodic)*.

Definice 14. Nechť ν_1 a ν_2 jsou dvě pravděpodobnostní míry na $\mathfrak{X}$. Definujeme jejich *vzdálenost v totální variaci (total variation distance)* jako

$$\|\nu_1 - \nu_2\|_{TV} = \sup_{A \in \mathfrak{X}} |\nu_1(A) - \nu_2(A)|.$$

Pozn.: Pokud existují hustoty f_1, f_2 měr ν_1, ν_2 vzhledem k nějaké σ-konečné míře μ, tak

$$\|\nu_1 - \nu_2\|_{TV} = \frac{1}{2} \int_{\mathcal{X}} |f_1(x) - f_2(x)|\, \mu(\mathrm{d}x).$$

Věta 14. *Mějme markovský řetězec se stacionárním rozdělením π, který je φ-nerozložitelný a neperiodický, potom*

$$\|P^n(x, \cdot) - \pi(\cdot)\|_{TV} \underset{n \to \infty}{\longrightarrow} 0 \quad \textit{pro } \pi\textit{-s.v. } x \in \mathcal{X}.$$

Důkaz: [24]

□

Abychom neměli výjimečné body, pro které konvergence neplatí, potřebujeme podmínku na trvalost.

Definice 15. Pro $A \in \mathfrak{X}$ položme $\eta_A = \sum_{n=1}^{\infty} \mathbf{1}_{[X_n \in A]}$ počet navštívení množiny A, tzv. *čas okupace (occupation time)*. Řekneme, že množina A je *trvalá (recurrent)*, jestliže $U(x, A) = \mathbb{E}[\eta_A \mid X_0 = x] = \infty$ pro každé $x \in A$. Markovský řetězec $\{X_n\}$ nazveme *trvalý (recurrent)*, je-li ψ-nerozložitelný a každá A s $\psi(A) > 0$ je trvalá.
Pozn.: Nenulový Markovův řetězec je trvalý.

Definice 16. Řekneme, že množina A je *harrisovsky trvalá (Harris recurrent)*, když $L(x, A) = \mathbb{P}(\exists n : X_n \in A \mid X_0 = x) = 1$ pro každé $x \in A$. Markovský řetězec $\{X_n\}$ nazveme *harrisovsky trvalý (Harris reccurent)*, jestliže je ψ-nerozložitelný a každé $A \in \mathfrak{X}$ splňující $\psi(A) > 0$ je harrisovsky trvalá množina.
Pozn.: Ekvivalentně lze harrisovsky trvalou množinu definovat vlastností $Q(x, A) = \mathbb{P}(\eta_A = \infty \mid X_0 = x) = 1$ pro každé $x \in A$. Odtud je vidět, že $U(x, A) = \mathbb{E}[\eta_A \mid X_0 = x] = \infty$, a tudíž je A i trvalá.
Pozn.: Pro harrisovsky trvalý řetězec je $Q(x, A) = 1$ pro každé $x \in \mathcal{X}$ a A s $\psi(A) > 0$.

Definice 17. Je-li řetězec $\{X_n\}$ harrisovsky trvalý, neperiodický a nenulový, nazývá se *ergodický (ergodic)*.

Podle tvrzení 10 je stacionární rozdělení přirozený kandidát pro limitní rozdělení. Pro ergodický řetězec je stacionární rozdělení limitní. Na rozdíl od věty 14 máme konvergenci pro všechna x.

Věta 15. *Nechť markovský řetězec $\{X_n\}$ se stacionárním rozdělením π je ergodický, pak*

$$\|P^n(x, \cdot) - \pi(\cdot)\|_{TV} \underset{n\to\infty}{\longrightarrow} 0 \quad \textit{pro všechna } x \in \mathcal{X}.$$

Důkaz: [17], Theorem 13.3.3.

□

Uvažujme měřitelnou funkci $h : \mathcal{X} \to \mathbb{R}$. V praxi jsou často výstupem MCMC průměry $\bar{h}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, proto nás zajímá vyšetřování asymptotických vlastností $\bar{h}_n$. Uvedeme si limitní věty pro konvergenci průměrů ke střední hodnotě $\mathbb{E}_\pi h(X) = \int h(x)\,\pi(\mathrm{d}x)$ vzhledem ke stacionárnímu rozdělení.

Věta 16. *(silný zákon velkých čísel, ergodická věta) Nechť $\{X_n\}$ je ergodický Markovův řetězec, potom pro libovolnou funkci h splňující $\mathbb{E}_\pi|h(X)| < \infty$ platí*

$$\mathbb{P}(\lim_{n\to\infty} \bar{h}_n = \mathbb{E}_\pi h(X)) = 1 \quad \textit{pro libovolné počáteční rozdělení } \varrho.$$

Důkaz: [17]

□

Definice 18. Ergodický řetězec $\{X_n\}$ nazveme *stejnoměrně ergodický (uniformly ergodic)*, pokud $\|P^n(x, \cdot) - \pi(\cdot)\|_{TV}$ konverguje k nule stejnoměrně v x pro $n \to \infty$.

Definice 19. Řekneme, že ergodický Markovův řetězec $\{X_n\}$ je *geometricky ergodický (geometrically ergodic)*, existuje-li $\lambda \in [0, 1)$ a reálná integrovatelná funkce V na $\mathcal{X}$ tak, že

$$\|P^n(x, \cdot) - \pi(\cdot)\|_{TV} \le V(x)\lambda^n \quad \text{pro každé } x \in \mathcal{X},\ n \in \mathbb{N}.$$

Infimum takových λ se nazývá *geometrický řád konvergence (geometric rate of convergence)* Markovova řetězce.
Pozn.: Stejnoměrná ergodicita je ekvivalentní tomu, že v definici geometrické ergodicity je V konstantní. Proto stejnoměrná ergodicita implikuje geometrickou ergodicitu a ta pro změnu implikuje ergodicitu.

Pro geometricky ergodické řetězce platí centrální limitní věta, která nám umožňuje statistickou inferenci výstupů z MCMC.

Věta 17. *Nechť $\{X_n\}$ je geometricky ergodický markovský řetězec. Buď $\mathbb{E}_\pi|h(X)|^{2+\varepsilon} < \infty$ pro nějaké $\varepsilon > 0$ nebo $\{X_n\}$ je reverzibilní a $\mathbb{E}_\pi h(X)^2 < \infty$, potom*

$$\sqrt{n}\left(\bar{h}_n - \mathbb{E}_\pi h(X)\right) \underset{n\to\infty}{\xrightarrow{\mathcal{D}}} N(0, \sigma_h^2),$$

kde

$$\sigma_h^2 = \operatorname{var}_\pi h(X_0) + 2\sum_{i=1}^{\infty} \operatorname{cov}_\pi(h(X_0), h(X_i)).$$

Důkaz: [17]

□

Pozn.: Konečnost sumy v σ_h^2 je zajištěna geometrickou ergodicitou.

Definice 20. Množina $B \in \mathfrak{X}$ je *atom (atom)* řetězce $\{X_n\}$, pokud existuje míra ν na $\mathfrak{X}$ tak, že $P(x, A) = \nu(A)$ pro všechna $x \in B$. Je-li $\{X_n\}$ ψ-nerozložitelný a $\psi(B) > 0$, je B *dosažitelný atom (accessible atom)*.
Pozn.: Jediný bod je vždy atom. Pro diskrétní množinu stavů a nerozložitelný řetězec je každý bod dosažitelný atom.
Pozn.: Protože pro atom B je pravděpodobnost $P(x, A)$ stejná pro všechny $x \in B$, můžeme ji zkráceně psát symbolem $P(B, A)$. Totéž platí pro P^n, protože

$$P^n(x, A) = \int P^{n-1}(y, A)\, P(x, \mathrm{d}y) = \int P^{n-1}(y, A)\, \nu(\mathrm{d}y) = \nu_n(A) \quad \text{pro } x \in B.$$

Podobně píšeme $U(B, A) = \sum_{n=1}^{\infty} P^n(B, A)$.

Tvrzení 18. *Je-li B atom s $\sum_n P^n(x, B) > 0$ pro všechna $x \in \mathcal{X}$, je B dosažitelný atom a $\{X_n\}$ je ν-nerozložitelný, kde $\nu(\cdot) = P(B, \cdot)$.*

Důkaz: Z Chapman-Kolmogorovovy rovnosti je pro každé $n \in \mathbb{N}$

$$P^{n+1}(x, A) \geq \int_B P(y, A)\, P^n(x, \mathrm{d}y) = P^n(x, B)\nu(A).$$

Sečtením přes n plyne ν-nerozložitelnost.

□

Věta 19. *Nechť $\{X_n\}$ je ψ-nerozložitelný řetězec a B je dosažitelný atom, který je trvalý (tj. $U(B, B) = \infty$). Potom každá A s $\psi(A) > 0$ je trvalá množina.*

Důkaz: Z ψ-nerozložitelnosti plyne, že pro každé $x \in \mathcal{X}$ existuje $k \in \mathbb{N}$ takové, že $P^k(x, B) > 0$ a $l \in \mathbb{N}$ takové, že $P^l(B, A) > 0$. Z tvrzení 9 potom máme

$$\sum_n P^{k+l+n}(x, A) \geq \sum_n \int_B \int_B P^l(y, A)\, P^n(z, \mathrm{d}y)\, P^k(x, \mathrm{d}z) = P^l(B, A)P^k(x, B) \sum_n P^n(B, B).$$

Nyní už si stačí jen uvědomit, že trvalost B znamená, že pravá strana diverguje.

□

Abychom zformulovali postačující podmínky pro geometrickou a stejnoměrnou ergodicitu budeme potřebovat zavést minorizační podmínku a malou množinu.
Definice 21. Řekneme, že φ-nerozložitelný Markovův řetězec splňuje *minorizační podmínku (minorization condition)* $M(m, \varepsilon, C, \nu)$, jestliže pro $m \in \mathbb{N}$, $\varepsilon > 0$, množinu $C \in \mathfrak{X}$ a pravděpodobnostní míru ν platí $P^m(x, A) \geq \varepsilon\nu(A)$ pro všechna $x \in C$ a $A \in \mathfrak{X}$.

Definice 22. Řekneme, že $C \in \mathfrak{X}$ je *malá množina (small set)*, když řetězec splňuje $M(m, \varepsilon, C, \nu)$ pro nějaké $m \in \mathbb{N}$, pravděpodobnostní míru ν a $\varepsilon > 0$.
Pozn.: V diskrétní situaci jsou atomy malé množiny.

Věta 20. *Nechť $\{X_n\}$ je ψ-nerozložitelný, potom pro každé $A \in \mathfrak{X}$ s $\psi(A) > 0$ existuje malá množina $C \subseteq A$ tak, že $\psi(C) > 0$ a $\nu(C) > 0$.*

Důkaz: [17]

□

Příklad: Uvažujme náhodnou procházku na polopřímce $[0, \infty)$: $X_{k+1} = \max(X_k + Z_{k+1}, 0)$, $k \in \mathbb{N}_0$, kde Z_k jsou nezávislé reálné náhodné veličiny s distribuční funkcí F a X_0 je nezávislá na $\{Z_k\}$. Předpokládejme, že $F(z) = \varepsilon > 0$ pro nějaké $z < 0$. Pro $A \subseteq (0, \infty)$ je $P(x, A) = \mathbb{P}(X_0 + Z_1 \in A \mid X_0 = x) = \mathbb{P}(Z_1 \in$

$A-x$), kde $A-x=\{y-x : y\in A\}$. Dále $P(x,\{0\})=\mathbb{P}(X_0+Z_1\le 0 \mid X_0=x)=\mathbb{P}(Z_1\le -x)=F(-x)$. Potom pro každé x je $P^n(x,\{0\})\ge \varepsilon^n>0$, kde $n=\lfloor \frac{x}{|z|}\rfloor+1$. Tedy $\{X_n\}$ je δ_0-nerozložitelný Markovův řetězec. Podle věty 12 je ψ-nerozložitelný s

$$\psi(A)=\int K_{\frac12}(y,A)\delta_0(\mathrm{d}y)=K_{\frac12}(0,A)=\frac12\sum_{n=0}^{\infty}\frac{1}{2^n}P^n(0,A).$$

Protože $P(0,\{0\})=F(0)>0$, je $\psi(\{0\})>0$ a $\{0\}$ je dosažitelný atom. Každý kompakt $[0,c]$, $c\ge 0$, je malá množina. Stačí zvolit $m=\lfloor \frac{c}{|z|}\rfloor+1$, pak $P^m(x,B)\ge \varepsilon^m\delta_0(B)$ pro každé $x\in[0,c]$ a každou borelovskou množinou B (pro $0\notin B$ platí triviálně, pro $0\in B$ platí díky $F(z)=\varepsilon>0$).

Příklad: Nyní uvažujme náhodnou procházku na přímce: $X_{k+1}=X_k+Z_{k+1}$, $k\in\mathbb{N}_0$, kde Z_k jsou nezávislé stejně rozdělené reálné náhodné veličiny s distribuční funkcí F a nezávislé na X_0. Předpokládejme, že F má absolutně spojitou složku vzhledem k Lebesgueově míře λ^1 s hustotou splňující $f(x)\ge\delta$ pro $|x|<\beta$ pro nějaké $\delta,\beta>0$. Potom $\mathbb{P}(Z_k\in A)\ge\int_A f(x)\,\mathrm{d}x$. Položme $C=\{x : |x|\le\frac{\beta}{2}\}$. Je-li $B\subseteq C$ a $x\in C$, potom

$$P(x,B)=\mathbb{P}(Z_k\in B-x)\ge\int_{B-x} f(y)\,\mathrm{d}y\ge\delta\lambda^1(B), \tag{10}$$

kde λ^1 značí Lebesgueovu míru. Z libovolného x můžeme dosáhnout C v nejvýš $n=\lfloor\frac{2|x|}{\beta}\rfloor$ krocích s kladnou pravděpodobností, tedy $\lambda^1|_C$ je míra nerozložitelnosti. Navíc C je malá množina díky (10), splňuje $M(1,\delta,C,\lambda^1|_C)$. Pokud má F hustotu, tak neexistuje dosažitelný atom.

Definice 23. Nechť V je měřitelná nezáporná funkce na $\mathcal{X}$. *Operátor driftu (drift operator)* Δ pro V a Markovův řetězec s přechodovým jádrem P je definován jako $\Delta V(x)=PV(x)-V(x)$. Hodnota $\Delta V(x)$ se nazývá *drift*.

Definice 24. Nechť C je malá množina. Řekneme, že řetězec splňuje *podmínku geometrického driftu (geometric drift condition)*, pokud existuje funkce $V:\mathcal{X}\to[1,\infty)$ taková, že

$$\Delta V(x)\le(\lambda-1)V(x)+b\mathbf{1}_C(x), \tag{11}$$

pro nějaké konstanty $b>0$ a $0<\lambda<1$.

Věta 21. *Markovův řetězec je geometricky ergodický, právě když splňuje podmínku geometrického driftu pro nějakou malou množinu C. Je-li V omezená, pak je řetězec stejnoměrně ergodický.*

Důkaz: [17], Chapter 15.

□

Věta 22. *Markovský řetězec je stejnoměrně ergodický právě tehdy, když $\mathcal{X}$ je malá množina. Řád konvergence je menší nebo roven než $(1-\varepsilon)^{1/m}$.*

Důkaz: [24]

□

5.3 Ergodicita MCMC algoritmů

Víme, že pro ergodický Markovův řetězec máme konvergenci ke stacionárnímu rozdělení. V kapitole jsme uvedli příklady konstrukce řetězců s předepsaným stacionárním rozdělením. Abychom měli zajištěnu konvergenci tohoto řetězce, potřebujeme ověřit nerozložitelnost a neperiodicitu. Rovněž můžeme zkoumat rychlost konvergence (geometrická ergodicita, stejnoměrná ergodicita).

Pro diskrétní množinu stavů je zjištění nerozložitelnosti a neperiodicity dobře známou úlohou z přednášky *Náhodné procesy*. Výpočtem vlastních čísel matice pravděpodobností přechodu lze zjistit geometrický řád konvergence (viz cvičení).

Uvedeme postačující podmínky, za kterých dostaneme konvergenci k cílovému rozdělení v případě Gibbsova výběrového plánu a Metropolisova-Hastingsova algoritmu.

Chceme simulovat z pravděpodobnostního rozdělení π na prostoru $\mathcal{X}$ s hustotou f vzhledem k σ-konečné míře μ. Připomeňme, že značíme $\mathcal{X}^+=\{x\in\mathcal{X} : f(x)>0\}$.

Věta 23. *Předpokládejme, že $\mathcal{X}^+=\prod_{i=1}^d\mathcal{X}_i$ a $\mu=\prod_{i=1}^d\mu_i$, kde $\mu_i(\mathcal{X}_i)>0$. Potom markovský řetězec generovaný Gibbsovým výběrovým plánem (algoritmus 2) je μ-nerozložitelný a neperiodický.*

Důkaz: Podmíněná hustota $f(x_i \mid x_{-i})$ je dobře definováno pro každé $x \in \mathcal{X}^+$ a každé $i \in \{1, \ldots, d\}$. Proto přechodová hustota Gibbsova jádra P je dobře definována a splňuje

$$p(x, y) = \prod_{i=1}^{d} f(y_i \mid y_1, \ldots, y_{i-1}, x_{i+1}, \ldots, x_d) > 0$$

pro každé $x, y \in \mathcal{X}^+$. Tudíž příslušný markovský řetězec je μ-nerozložitelný a neperiodický, neboť $P(x, A) = \int_A p(x, y)\, \mu(\mathrm{d}y) > 0$, jakmile $\mu(A) > 0$. Protože π je absolutně spojitá vůči μ, je pochopitelně řetězec i π-nerozložitelný. Podle věty je π stacionární rozdělení a podle věty 14 je rovněž limitní. □

Věta 24. *Uvažujme Markovův řetězec $\{X_n\}$ generovaný Metropolisovým-Hastingsovým algoritmem (algoritmus 4). Označme P příslušné přechodové jádro řetězce a π stacionární rozdělení. Předpokládejme, že hustota q návrhového jádra Q je nulová mimo $\mathcal{X}^+ \times \mathcal{X}^+$.*

(i) Nechť návrhové jádro Q je neperiodické nebo $\pi(\{x : P(x, \{x\}) > 0\}) > 0$, potom je $\{X_n\}$ neperiodický Markovův řetězec.

(ii) Pokud jádro Q je π-nerozložitelné a $q(x, y) = 0$, právě když $q(y, x) = 0$, potom $\{X_n\}$ je π-nerozložitelný.

Důkaz: Část (i) je zřejmá. V části (ii) podmínka $q(x, y) = 0 \Leftrightarrow q(y, x) = 0$ znamená, že $\alpha(x, y) > 0$ pro všechna $x, y \in \mathcal{X}^+$. Připomeňme, že $P(x, A) = \int_A q(x, y)\alpha(x, y)\, \mu(\mathrm{d}y)$ pro $x \notin A$ a $P(x, \{x\}) = r(x) = 1 - \int_{y \neq x} q(x, y)\alpha(x, y)\, \mu(\mathrm{d}y)$. Podobně přechodové jádro n-tého řádu má absolutně spojitou a atomickou část ($P^n(x, \{x\}) > 0$). Nyní si stačí uvědomit, že $Q^n(x, A) > 0$ implikuje $P^n(x, A) > 0$ pro libovolné $n \in \mathbb{N}$, $x \in \mathcal{X}^+$ a $A \in \mathfrak{X}$ takové, že $\pi(A) > 0$. Pro $n = 1$ je to zřejmé díky tomu, že $\alpha(x, y) > 0$. Pro $n > 1$ je třeba si rozmyslet, že skládání jader nic nezkazí: $Q^n(x, A)$ znamená, že s kladnou pravděpodobností se ze stavu x po n návrzích dostaneme do množiny A, přitom pravděpodobnost, že přijmeme všech n návrhů je kladná, proto i pravděpodobnost přechodu řetězce z x do A je kladná.

□

Pozn.: Pokud $q(x, y) > 0$ pro každé $x, y \in \mathcal{X}$, potom $\{X_n\}$ je π-nerozložitelný.

Příklad: Nechť π je rovnoměrné rozdělení na $[0, 1]^2 \cup [1, 2]^2$ a q je definována na $[0, 2]^2$ následovně:

$$q((x_1, x_2), (y_1, y_2)) = \begin{cases} \frac{1}{4} & \text{pro } y_1 = x_1 \text{ a } 0 \leq y_2 \leq 2, \\ \frac{1}{4} & \text{pro } y_2 = x_2 \text{ a } 0 \leq y_1 \leq 2. \end{cases}$$

Lehce se přesvědčíme, že Metropolisův-Hastingsův algoritmus je v této situaci rozložitelný. Proto v předchozí větě předpokládáme, že nosič hustoty $q(x, \cdot)$ je obsažen v nosiči cílové hustoty.

V tomto případě i Gibbsův výběrový plán dává rozložitelný řetězec.

Příklad: Příkladem nerozložitelného a neperiodického Metropolisova-Hastingsova algoritmu je symetrická náhodná procházka s hustotou q_0, která je kladná všude na $\mathbb{R}^d$ (např. mnohorozměrné normální rozdělení). Pokud je $\mathcal{X}^+$ otevřená souvislá podmnožina $\mathbb{R}^d$ a q_0 je kladná na nějakém okolí nuly, potom Metropolisův-Hastingsův algoritmus symetrické náhodné procházky dává π-nerozložitelný a neperiodický řetězec.

Příklad: Nezávislý Metropolisův-Hastingsův algoritmus je π-nerozložitelný a neperiodický, právě když $q_0(x) > 0$ pro μ-s.v. $x \in \mathcal{X}^+$.

Abychom měli zajištěnu konvergenci algoritmu pro všechna počáteční rozdělení, potřebujeme ještě harrisovskou trvalost. Naštěstí v našich aplikacích se dá ukázat, že většina φ-nerozložitelných Gibbsových výběrových plánů a všechny φ-nerozložitelné Metropolisovy-Hastingsovy algoritmy jsou harrisovsky trvalé. K ověření tohoto tvrzení se využívá níže uvedené tvrzení.

Definice 25. Nezáporná funkce h je *harmonická (harmonic)* pro markovské jádro P, pokud $h = Ph$.

Věta 25. *Trvalý řetězec je harrisovsky trvalý, právě když každá omezená harmonická funkce pro přechodové jádro řetězce je konstantní.*

Důkaz: [24]

□

Důsledek 26. *Nechť $\{X_n\}$ je φ-nerozložitelný Markovův řetězec se stacionárním rozdělením π. Pokud přechodové jádro P je*
(i) Gibbsovo a $P(x,\cdot)$ je absolutně spojité vzhledem k π pro všechna $x \in \mathcal{X}$,
(ii) Metropolisovo-Hastingsovo,
potom je řetězec harrisovsky trvalý.

Důkaz: [24]

□

Ke zkoumání rychlosti konvergence se v konkrétních situacích použijí výsledky uvedené v předchozí podkapitole. Například Metropolisův-Hastingsův algoritmus pro simulaci z hustot na kompaktní množině je stejnoměrně ergodický.

Tvrzení 27. *Pokud $\mu(\mathcal{X}^+) < \infty$ a q i f jsou omezené a odražené od nuly na $\mathcal{X}^+$, pak řetězec získaný Metropolisovým-Hastingsovým algoritmem je stejnoměrně ergodický.*

Důkaz: Existují konstanty $0 < c_1 \leq c_2 < \infty$ takové, že $c_1 \leq f(x) \leq c_2$ a $c_1 \leq q(x,y) \leq c_2$ pro všechna $x, y \in \mathcal{X}^+$. Proto $\alpha(x,y)q(x,y) \geq \frac{c_1^2}{c_2}$ a $P(x,A) \geq \frac{c_1^2}{c_2}\mu(A)$ pro všechna $x \in \mathcal{X}^+$. Odtud vidíme, že řetězec splňuje minorizační podmínku $M(1, \varepsilon, \mathcal{X}^+, \nu)$, kde $\varepsilon = \frac{c_1^2 \mu(\mathcal{X}^+)}{c_2}$ a $\nu = \frac{\mu(\cdot)}{\mu(\mathcal{X}^+)}$, což znamená, že $\mathcal{X}^+$ je malá množina a tvrzení plyne z věty 22.

□

Tvrzení 28. *Nezávislý Metropolisův-Hastingsův algoritmus s omezenou váhovou funkcí $w = f/q_0$ splňuje minorizační podmínku $M(1, \varepsilon, \mathcal{X}^+, \pi)$, kde $\varepsilon = \frac{1}{\sup_{x \in \mathcal{X}^+} w(x)}$, a je tudíž stejnoměrně ergodický. Řád konvergence je nanejvýš $1 - \varepsilon$.*

Důkaz: Vše plyne z věty 22 a faktu, že $P(x,A) \geq \frac{1}{w(x)}\pi(A)$ pro libovolné $x \in \mathcal{X}^+$ a $A \in \mathfrak{X}$.

□

6. Další algoritmy založené na MCMC metodách

6.1 Simulované žíhání

Simulované žíhání (simulated annealing) je příkladem stochastického optimalizačního algoritmu. Název je odvozen z toho, že snahou je simulovat fyzikální proces žíhání. Tato technika se používá v metalurgii, kde kontrolované chlazení materiálů vede k zvětšení velikostí krystalů a zmenšení jejich defektů. Při velké teplotě se atomy pohybují víceméně volně, zatímco při menších teplotách jsou pohyby spíše do míst s nižší energií.

Naším cílem je nalezení globálního minima (nebo maxima) reálné funkce h na prostoru $\mathcal{X}$. Budeme postupovat tak, že necháme běžet Markovův řetězec, jehož stacionární rozdělení je soustředěno na stavech s malou (nebo velkou) hodnotou h. Po nějaké době přeskočíme na řetězec, jehož stacionární rozdělení je ještě více koncentrováno na stavech s malou (nebo velkou) hodnotou h a takto pokračujeme dále. Volbu řetězců kontrolujeme pomocí parametru T, který ve fyzikální interpretaci označuje teplotu. Při dané teplotě máme kladnou pravděpodobnost přechodu do horšího stavu (stavu s větší hodnotou funkce h), tato pravděpodobnost však klesá se snižující se teplotou. Možnost pohybů do horších stavů je důležitá v tom, že nám zabraňuje v uvíznutí v lokálním minimu.

Otázkou zůstává, jak při dané teplotě volit markovský řetězec a jak příslušné stacionární rozdělení. Pokud je h nezáporná a funkce $h^{1/T}$ je integrovatelná, lze uvažovat stacionární rozdělení s hustotou f úměrnou $h^{1/T}$. Pro velká T je většina pravděpodobnostní hmoty rozložena kolem maxima hustoty f.

Jiná možnost (pro úlohu minimalizace) je dána následující definicí.

Definice 26. *Boltzmannovo rozdělení (Boltzmann distribution)* $\pi_{h,T}$ na $\mathcal{X}$ s funkcí energie $h : \mathcal{X} \to \mathbb{R}$ a parametrem teploty $T > 0$ je dáno hustotou $f_{h,T}(x) = \frac{1}{Z_{h,T}} \exp\{-h(x)/T\}$, kde $Z_{h,T}$ je normující konstanta.
Pozn.: Předpokládáme, že funkce $\exp\{-h(x)/T\}$ je integrovatelná. Boltzmannovo rozdělení lze jednoduše modifikovat, aby se dalo použít pro úlohu maximalizace h, stačí uvažovat hustotu $f_{h,T}(x) = \frac{1}{Z_{h,T}} \exp\{h(x)/T\}$.

Následující věta říká, že pro konečný prostor $\mathcal{X}$ Boltzmannovo rozdělení s malou hodnotou T má požadovanou vlastnost, že umisťuje nejvíc pravděpodobnosti na prvky minimalizující h.

Věta 29. *Nechť S je konečná, $h : S \to \mathbb{R}$ libovolná funkce. Pro $T > 0$ buď $\alpha(T)$ pravděpodobnost, že náhodný element Y na S s Boltzmannovým rozdělením $\pi_{h,T}$ splňuje $h(Y) = \min_{s\in S} h(s)$. Potom $\lim_{T\to 0+} \alpha(T) = 1$.*

Důkaz: Předpokládejme, že minimum h je jediné, označme ho s. Nechť $a = h(s)$, $b = \min_{s'\neq s} h(s')$. Je $a < b$, tedy $\lim_{T\to 0+} \exp\{\frac{a-b}{T}\} = 0$. Potom

$$\pi_{h,T}(s) = \frac{1}{Z_{h,T}} \exp(-a/T) = \frac{\exp(-a/T)}{\sum_{s'\in S} \exp(-h(s')/T)} \geq \frac{\mathrm{e}^{-a/T}}{\mathrm{e}^{-a/T} + (k-1)\mathrm{e}^{-b/T}} = \frac{1}{1 + (k-1)\exp\{\frac{a-b}{T}\}},$$

kde k je počet prvků množiny S. Odsud $\lim_{T\to 0+} \pi_{h,T}(s) = 1$. Pro případ, že minimum se nabývá ve více bodech není těžké důkaz příslušně modifikovat.

□

Algoritmus simulovaného žíhání spočívá v konstrukci řetězce pro simulaci z $\pi_{h,T}$. Většinou se užívá Metropolisův-Hastingsův algoritmus, jeho výhoda je, že není nutné znát normující konstantu $Z_{h,T}$. Zvolí se klesající posloupnost kladných čísel (teplot) T_n, $\lim_{n\to\infty} T_n = 0$ a posloupnost přirozených čísel N_n – *schéma žíhání (annealing schedule)*. Řetězec běží (z libovolného počátečního stavu) N_1 časových jednotek při teplotě T_1, N_2 při T_2 atd., dokud není splněna podmínka ukončení (např. proběhl zadaný počet iterací nebo nedošlo k žádnému zlepšení po daném počtu iterací). Stav řetězce, ve kterém byla dosažena nejmenší hodnota, považujeme za řešení naší optimalizační úlohy. Existují věty uvádějící, jak rychle musí jít T_n k nule (jak rychle musíme ochlazovat), abychom měli zaručenu konvergenci.

Věta 30. *Nechť $S = \{s_1, \ldots, s_k\}$ a $h : S \to \mathbb{R}$. Je-li $T^{(n)}$ teplota v čase n a*

$$T^{(n)} \geq \frac{k\,(\max_{s\in S} h(s) - \min_{s\in S} h(s))}{\log n}$$

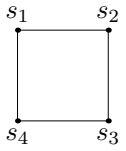
pro dostatečně velká n, potom $\lim_{n\to\infty} \alpha(T^{(n)}) = 1$, kde $\alpha(T^{(n)}) = \mathbb{P}(h(Y_n) = \min_{s\in S} h(s))$ a Y_n je stav řetězce v čase n.

Důkaz: [7]

□

Typicky ovšem splnění podmínky z předchozí věty vede na extrémně pomalé algoritmy. V praxi užití rychlejšího ochlazování nese nebezpečí, že algoritmus skončí v lokálním a nikoliv globálním minimu. Nutné kompromisy mezi pomalým a rychlým ochlazováním se hledají případ od případů. K volbě schématu žíhání většinou není lepší doporučení než metoda pokusu a omylu.

Příklad: Tento příklad by měl varovat před rychlým ochlazovacím schématem. Nechť $S = \{s_1, s_2, s_3, s_4\}$, $h(s_1) = 1$, $h(s_2) = 2$, $h(s_3) = 0$ a $h(s_4) = 2$.



Hledejme minimum metodou simulovaného žíhání. Návrh v Metropolisově-Hastingsově algoritmu volíme tak, že rovnoměrně náhodně vybereme stav mezi sousedy současného stavu, tedy $q_{ij} = \frac{1}{d_i}$, pokud s_j je soused s_i, kde d_i je počet sousedů s_i. V našem případě je $d_i = 2$ pro všechna i a pravděpodobnosti přijetí závisí pouze na podílu $\pi_{h,T}(s_j)/\pi_{h,T}(s_i)$. Celkově Dostaneme matici pravděpodobností přechodu

$$P_T = \begin{pmatrix} 1 - e^{-1/T} & \frac{1}{2}\mathrm{e}^{-1/T} & 0 & \frac{1}{2}\mathrm{e}^{-1/T} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2}\mathrm{e}^{-2/T} & 1 - \mathrm{e}^{-2/T} & \frac{1}{2}\mathrm{e}^{-2/T} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \end{pmatrix}.$$

Nechť nehomogenní Markovův řetězec $\{X_n\}$ startuje v $X_0 = s_1$ a běží dle nějakého žíhacího schématu. Značíme $T^{(n)}$ teplotu v čase n a A jev, že řetězec zůstane v s_1 navždy. Potom je

$$\begin{aligned}\mathbb{P}(A) &= \mathbb{P}(X_1 = s_1, X_2 = s_1, \ldots) = \lim_{n\to\infty} \mathbb{P}(X_1 = s_1, \ldots, X_n = s_1) \\ &= \lim_{n\to\infty} \mathbb{P}(X_1 = s_1 \mid X_0 = s_1)\mathbb{P}(X_2 = s_1 \mid X_1 = s_1) \cdot \mathbb{P}(X_n = s_1 \mid X_{n-1} = s_1) \\ &= \lim_{n\to\infty} \prod_{i=1}^{n} \left(1 - \mathrm{e}^{-1/T^{(i)}}\right) = \prod_{i=1}^{\infty} \left(1 - \mathrm{e}^{-1/T^{(i)}}\right).\end{aligned}$$

Pro $0 \leq u_i < 1$ platí $\prod_{i=1}^{\infty}(1-u_i) > 0 \Leftrightarrow \sum_{i=1}^{\infty} u_i < \infty$ (viz [22], věta 15.5). Tedy pokud jde $T^{(n)}$ k nule dost rychle (tak, že $\sum_{i=1}^{\infty} \mathrm{e}^{-1/T^{(i)}} < \infty$), potom je $\mathbb{P}(A) > 0$ a řetězec může zůstat v s_1 navždy (např. pro $T^{(n)} = 1/n$). Stav s_1 je lokální minimum (ne globální).

Příkladem použití simulovaného žíhání jsou různé NP-těžké kombinatorické optimalizační problémy (např. problém obchodního cestujícího nebo bisekce grafu).

6.2 Perfektní simulace

Jedná se o algoritmus, který dává na výstupu přesně stacionární rozdělení a navíc je schopen určit, kdy je stacionární rozdělení dosaženo (kdy algoritmus zastavit). Tedy není třeba zabývat se řádem konvergence ani otázkou, jak dlouho nechat řetězec běžet.

Vyložíme metodu perfektní simulace založenou na myšlence CFTP (coupling from the past), kterou navrhli Propp a Wilson [20]. Při algoritmu neběží pouze jeden ale více řetězců (*coupling*). Navíc řetězce neběží od času 0 dopředu, ale běží z minulosti do času 0 (*from the past*).

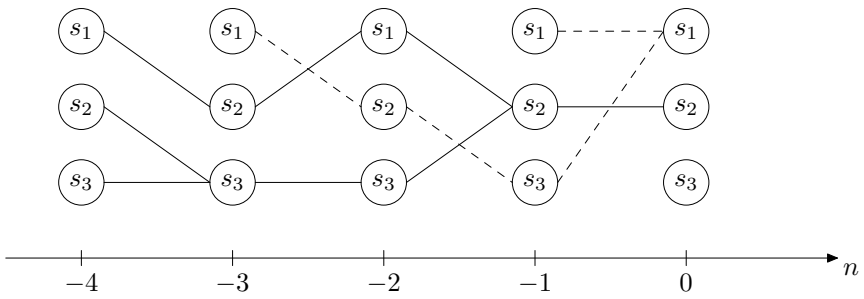
Cílem je simulovat z rozdělení π na konečném stavovém prostoru $S = \{s_1, \ldots, s_k\}$. Nechť $P = \{p_{ij}\}$ je matice pravděpodobností přechodu nerozložitelného, neperiodického a reverzibilního řetězce vzhledem k π. Funkce $\Phi : S \times [0,1] \to S$ splňující $p_{ij} = \mathbb{P}(\Phi(s_i, U) = s_j)$, kde $U \sim R(0,1)$, se nazývá *přechodová funkce (update function)*. Takováto funkce existuje pro každý homogenní Markovův řetězec ([12], Proposition 8.6). Přechodová funkce se dá využít při simulaci markovského řetězce $\{X_n\}$, pro $n \in \mathbb{N}$ totiž platí $X_n = \Phi(X_{n-1}, U_n)$, kde U_n je posloupnost nezávislých náhodných veličin s rovnoměrným rozdělením na $[0,1]$. Stačí tedy specifikovat počáteční stav X_0, simulovat z $R(0,1)$ a dopočítávat hodnoty vygenerovaného řetězce. Dále ještě uvažujme rostoucí posloupnost přirozených čísel $N_1, N_2, \ldots$ (běžně se volí $N_k = 2^{k-1}$, $k \in \mathbb{N}$) a $U_0, U_{-1}, U_{-2}, \ldots$ posloupnost nezávislých náhodných veličin s rovnoměrným rozdělením na $[0,1]$.

Algoritmus 5. *CFTP perfektní simulace (CFTP perfect simulation)*:

1. polož $m = 1$,
2. pro každý stav $s \in S$ simuluj Markovův řetězec s maticí pravděpodobností přechodu P, který startuje v čase $-N_m$ ze stavu s a běží do času 0 užitím Φ a $U_{-N_m+1}, \ldots, U_0$ (stejné pro všech k řetězců), tj. $X_{-N_m} = s$ a $X_t = \Phi(X_{t-1}, U_t)$, $t = -N_m + 1, \ldots, 0$,
3. pokud všechny řetězce jsou v čase 0 ve stejném stavu, je tento stav výstupem a algoritmus končí, jinak zvětši m o jedna a jdi na 2.

Pokud všechny řetězce skončí ve stejném stavu, říkáme, že došlo ke *koalescenci (coalescence)*. V druhém kroku se vždycky užívá stejná posloupnost U_i, což vyžaduje jisté nároky na paměť počítače.

Příklad: $S = \{s_1, s_2, s_3\}$, $N_1 = 1$, chod z času -1 do 0: nechť $\Phi(s_1, U_0) = s_1$, $\Phi(s_2, U_0) = s_2$, $\Phi(s_3, U_0) = s_1$, tedy nedošlo ke koalescenci. Jdeme na $N_2 = 2$, nechť $\Phi(\Phi(s_1, U_{-1}), U_0) = \Phi(s_2, U_0) = s_2$, $\Phi(\Phi(s_2, U_{-1}), U_0) = \Phi(s_3, U_0) = s_1$, $\Phi(\Phi(s_3, U_{-1}), U_0) = \Phi(s_2, U_0) = s_2$, opět není koalescence. Jdeme na $N_3 = 4$ a dostáváme koalescenci (viz obrázek), výstupem je stav s_2.



Kdybychom začínali v časech $-8, -16, \ldots$, vždy bude stejný výstup (jde o výstup z π).

Problém je, že nemáme zaručeno, že algoritmus skončí v konečném čase (jsou třeba nějaké dodatečné podmínky na Φ). Pokud ovšem skončí v konečném čase, dává správný výstup.

Věta 31. *Předpokládejme, že algoritmus skončí s pravděpodobností 1, buď Y výstup algoritmu. Potom pro každé $i \in \{1, \ldots, k\}$ je $\mathbb{P}(Y = s_i) = \pi_i$, kde π je stacionární rozdělení.*

Důkaz: Pro $s_i \in S$ ukážeme, že $|\mathbb{P}(Y = s_i) - \pi_i| \leq \varepsilon$ pro libovolné $\varepsilon > 0$. Z předpokladu existuje M tak, že $\mathbb{P}(\text{algoritmus nepotřebuje startovat z času menšího než } -N_M) \geq 1 - \varepsilon$. Uvažujme Markovův řetězec od času $-N_M$ do 0 se stejnou Φ a $U_{-N_M+1}, \ldots, U_0$, ale s počátečním rozdělením π. Buď $\tilde{Y}$ jeho stav v 0 (má rozdělení π). Potom z našeho předpokladu plyne $P(Y \neq \tilde{Y}) \leq \varepsilon$, a proto $\mathbb{P}(Y = s_i) - \pi_i = \mathbb{P}(Y = s_i) - \mathbb{P}(\tilde{Y} = s_i) \leq \mathbb{P}(Y = s_i, \tilde{Y} \neq s_i) \leq \mathbb{P}(Y \neq \tilde{Y}) \leq \varepsilon$. Podobně se ukáže $\pi_i - \mathbb{P}(Y = s_i) \leq \varepsilon$.

□

Příklad: Ukážeme, že metoda nefunguje, pokud provádíme coupling dopředu, ani pokud neužíváme stejná U_i. Buď

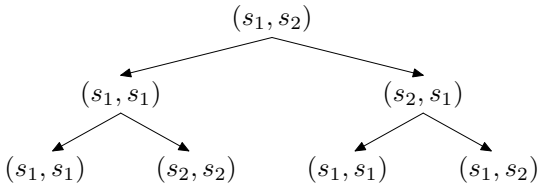
$$P = \begin{pmatrix} 1/2 & 1/2 \\ 1 & 0 \end{pmatrix}.$$

Lehce zjistíme, že $\pi = (2/3, 1/3)^T$.

Uvažujme dva řetězce (ze stavu s_1 a ze stavu s_2) startující čase v nule. Nechť koalescence nastane v čase N. V čase $N - 1$ jsou různé, tedy jeden z nich je ve stavu s_2, z nějž jde s pravděpodobností 1 do stavu s_1. Tedy s pravděpodobností 1 je koalescence ve stavu s_1, což neodpovídá stacionárnímu rozdělení.

Označme $M = \max\{m : \text{algoritmus se rozhodne simulovat řetězec startující v čase } -N_m\}$. Nechť se generují vždy nová U_i a Y je výstup algoritmu. Potom

$$\begin{aligned}\mathbb{P}(Y = s_1) &= \sum_{m=1}^{\infty} \mathbb{P}(M = m, Y = s_1) \geq \mathbb{P}(M = 1, Y = s_1) + \mathbb{P}(M = 2, Y = s_1) \\ &= \mathbb{P}(M = 1)\mathbb{P}(Y = s_1 \mid M = 1) + \mathbb{P}(M = 2)\mathbb{P}(Y = s_1 \mid M = 2) = \frac{1}{2} \cdot 1 + \frac{3}{8} \cdot \frac{2}{3} = \frac{3}{4} > \frac{2}{3}.\end{aligned}$$



Pokud je prostor S obrovský, je náročné nechat běžet řetězec ze všech stavů. U některých úloh to není nutné, počet simulovaných řetězců lze výrazně snížit. Uvedeme si tzv. sendvičovou vlastnost, která se uplatňuje pro markovské řetězce s uspořádáním na množině stavů.

Příklad: Uvažujme žebříkovou náhodnou procházku na $S = \{1, \ldots, k\}$, tj. pravděpodobnosti přechodu jsou $p_{i,i+1} = p_{i+1,i} = 1/2$ pro $i = 1, \ldots, k-1$ a $p_{11} = p_{kk} = 1/2$. Stacionární rozdělení je $\pi_i = \frac{1}{k}$, $i = 1, \ldots, k$. Přechodovou funkci lze vzít tvaru

$$\Phi(1, x) = \begin{cases} 1 & x \in [0, 1/2), \\ 2 & x \in [1/2, 1], \end{cases} \qquad \Phi(k, x) = \begin{cases} k-1 & x \in [0, 1/2), \\ k & x \in [1/2, 1], \end{cases}$$

$$\Phi(i, x) = \begin{cases} i-1 & x \in [0, 1/2), \\ i+1 & x \in [1/2, 1], \end{cases} \quad i = 1, \ldots, k.$$

Takto definována Φ má vlastnost monotonie: pro každé $x \in [0, 1]$, $i, j \in \{1, \ldots, k\}$ platí $i \leq j \Rightarrow \Phi(i, x) \leq \Phi(j, x)$. To znamená, že řetězec, který startuje ze stavu $i \in \{2, \ldots, k-1\}$ vždy zůstane mezi řetězci startujícími v 1 a k. Této vlastnosti se říká *sendvičová vlastnost (sandwich property)*. Když dojde ke koalescenci dvou krajních řetězců, nastává koalescence všech a algoritmus můžeme ukončit. Stačí tedy spustit jen dva řetězce – ze stavu 1 a k.

Sendvičová vlastnost umožňuje využití algoritmu v problémech s obecnou množinou stavů, na které existuje částečné uspořádání. Kromě Proppova-Wilsonova algoritmu existují v literatuře různé další varianty CFTP metod. Jiný algoritmus perfektní simulace je tzv. *přerušitelná metoda (interruptible method)* – viz [4].

7. Bodové procesy

Kromě bayesovské statistiky (kapitola 3) se metody MCMC na obecných stavových prostorech často využívají v prostorové statistice. Detailněji je této problematice věnována přednáška prof. Beneše *Prostorové modelování, prostorová statistika*.

Buď (E, ϱ) separabilní úplný metrický prostor, $\mathcal{B}$ borelovská σ-algebra na E a $\mathcal{B}_0 \subseteq \mathcal{B}$ systém omezených množin. Nechť $\mathcal{N} = \{x \subseteq E : x(B) < \infty \ \forall B \in \mathcal{B}_0\}$, kde $x(B)$ označuje počet bodů $x \cap B$. Symbol $\mathcal{N}$ tedy označuje systém všech lokálně konečných podmnožin prostoru E. Na $\mathcal{N}$ lze zavést σ-algebru následovně: $\mathfrak{N} = \sigma\{\{x \in \mathcal{N} : x(B) = m\}, m \in \mathbb{N}_0, B \in \mathcal{B}_0\}$.

Definice 27. *Bodový proces (point process)* na E je náhodný element v měřitelném prostoru $(\mathcal{N}, \mathfrak{N})$. Nechť Λ je difúzní (tedy $\Lambda(\{\xi\}) = 0$ pro $\xi \in E$) a lokálně konečná míra na E (tedy $\Lambda(B) < \infty$ pro $B \in \mathcal{B}_0$). Bodový proces X takový, že
(i) $X(B)$ má Poissonovo rozdělení s parametrem $\Lambda(B)$ pro každé $B \in \mathcal{B}_0$,
(ii) $X(B_1), \ldots, X(B_n)$ jsou nezávislé pro každé $n \in \mathbb{N}$ a $B_1, \ldots, B_n \in \mathcal{B}_0$ po dvou disjunktní,
nazveme *Poissonův bodový proces (Poisson point process)* s *mírou intenzity (intensity measure)* Λ.
Pozn.: Obecněji lze bodový proces definovat jako náhodnou celočíselnou lokálně konečnou míru. Tento přístup připouští, že některé body započítáváme s větší násobností. Pokud má každý bod míru nanejvýš 1, nazývá se bodový proces jednoduchý. V naší definici uvažujeme jenom jednoduché bodové procesy.

Mějme Poissonův bodový proces X s difúzní (neatomickou) mírou intenzity Λ takovou, že $\Lambda(E) < \infty$. Lze vyjádřit rozdělení Poissonova procesu ($F \in \mathfrak{N}$):

$$\begin{aligned}\Pi(F) = \mathbb{P}(X \in F) &= \sum_{n=0}^{\infty} \mathbb{P}(X(E) = n)\mathbb{P}(X \in F \mid X(E) = n) \\ &= \sum_{n=0}^{\infty} \frac{\Lambda(E)^n}{n!} \mathrm{e}^{-\Lambda(E)} \int_E \cdots \int_E \mathbf{1}_{[\{x_1,\ldots,x_n\} \in F]} \frac{\Lambda(\mathrm{d}x_1)}{\Lambda(E)} \cdots \frac{\Lambda(\mathrm{d}x_n)}{\Lambda(E)} \\ &= \mathrm{e}^{-\Lambda(E)} \left[\mathbf{1}_{[\emptyset \in F]} + \sum_{n=1}^{\infty} \frac{1}{n!} \int_E \cdots \int_E \mathbf{1}_{[\{x_1,\ldots,x_n\} \in F]} \, \Lambda(\mathrm{d}x_1) \cdots \Lambda(\mathrm{d}x_n) \right].\end{aligned}$$

Budeme se zabývat bodovými procesy X s hustotou p vzhledem k Π, tj. platí $\mathbb{P}(X \in F) = \int_F p(x) \, \Pi(\mathrm{d}x)$. Takový proces X je konečný (díky podmínce $\Lambda(E) < \infty$) a jednoduchý (díky tomu, že Λ je neatomická). Často se uvažuje, že E je omezená podmnožina $\mathbb{R}^d$ a Λ je Lebesgueova míra, hustota p je potom vzhledem ke standardnímu Poissonovu procesu (homogenní proces s jednotkovou intenzitou na E). Nejznámějším příkladem konečného bodového procesu s hustotou vzhledem k rozdělení Poissonova procesu je Straussův proces, který je modelem pro odpudivé interakce mezi body.

Definice 28. Mějme reálné parametry $\beta > 0$, $0 \leq \gamma \leq 1$ a $R > 0$. *Straussův proces (Strauss process)* je bodový proces X s hustotou $p(x) = \alpha \beta^{x(E)} \gamma^{S(x)}$, kde $S(x) = \sum_{i \neq j} \mathbf{1}_{[\varrho(x_i, x_j) < R]}$.
Pozn.: Normující konstanta $\alpha = \left(\int_{\mathcal{N}} \beta^{x(E)} \gamma^{S(x)} \, \Pi(\mathrm{d}x) \right)^{-1}$ je většinou neznámá. Lze ji spočítat například pro limitní případ $\gamma = 1$, který odpovídá Poissonovu procesu s mírou intenzity $\beta\Lambda$. Případ $\gamma = 0$ znamená, že $S(x) = 0$ (pokládáme $0^0 = 1$) a výsledkem je bodový proces s pevným jádrem (hard-core process), tj. žádné dva body v x nemůžou být blíž než R. Strauss nazval tento proces modelem shlukování [23], to by odpovídalo případu $\gamma > 1$, pro který však $p(x)$ není integrovatelná.

Pokud bychom však uvažovali podmíněný Straussův proces (podmíněně při daném počtu $x(E)$ bodů procesu), hustota $p(x)$ už nezávisí na parametru β a je integrovatelná pro všechna $\gamma \geq 0$. Tedy pro $\gamma > 1$ můžeme dostat model pro shlukování bodů, který ovšem není moc vhodný, v praxi se používají lepší modely.

Pro simulaci bodových procesů s hustotou vzhledem k Poissonovu procesu se s výhodou užije Metropolisův-Hastingsův algoritmus. Normující konstanta se zkrátí a návrh lze volit změnou jediného bodu v realizaci současného stavu.

Algoritmus 6. *Metropolisův-Hastingsův algoritmus zrození a zániku (birth-death Metropolis-Hastings algorithm)*
Pro $t = 0, 1, \ldots$ a dané $X_t = x \in \mathcal{N}$, generuj X_{t+1} následovně:

1. s pravděpodobností $Q(x)$ navrhni přidání bodu ξ s hustotou $b(x,\xi)$ vzhledem k Λ, s pravděpodobností $1-Q(x)$ navrhni ubrání bodu η s pravděpodobností $d(x,\eta)$, $\eta \in x$,
2. návrh přijmi (buď $X_{t+1} = x \cup \xi$, nebo $X_{t+1} = x \setminus \eta$) s pravděpodobností $\alpha(x, x\cup\xi) = \min(1, h(x,\xi))$, $\alpha(x\cup\xi, x) = \min\left(1, \frac{1}{h(x,\xi)}\right)$, kde

$$h(x,\xi) = \frac{p(x\cup\xi)}{p(x)} \cdot \frac{1-Q(x\cup\xi)}{Q(x)} \cdot \frac{d(x\cup\xi,\xi)}{b(x,\xi)}.$$

Polož $X_{t+1} = x$, pokud je návrh zamítnut.

Dále volíme speciálně $Q(\cdot) = \frac{1}{2}$, $b(\cdot,\cdot) = \frac{1}{\Lambda(E)}$, $d(x\cup\xi,\cdot) = \frac{1}{x(E)+1}$, tedy $h(x,\xi) = \lambda(x,\xi)\frac{\Lambda(E)}{x(E)+1}$, kde $\lambda(x,\xi) = \frac{p(x\cup\xi)}{p(x)}$ je *podmíněná intenzita (conditional intensity)*.

Definujeme podmínku stability, která je postačující pro geometrickou ergodicitu algoritmu.

Definice 29. Řekneme, že konečný bodový proces X s hustotou p je *lokálně stabilní (locally stable)*, když $\lambda(x,\xi) \le K$ pro nějakou konstantu K, neboli

$$p(x\cup\xi) \le Kp(x), \quad \text{pro všechna } x \in \mathcal{N}, \xi \in E. \tag{12}$$

Pozn.: Z podmínky (12) plyne, že pro $p(x) = 0$ je i $p(x\cup\xi) = 0$, podmíněná intenzita $\lambda(x,\xi)$ je tedy dobře definována (pokládáme $0/0 = 0$). Není těžké ukázat, že lokální stabilita implikuje integrovatelnost hustoty p vzhledem k rozdělení Poissonova procesu.

Uvažujme markovský řetězec $\{X_t\}$ generovaný algoritmem zrození a zániku popsaným výše. Protože řetězec přechází vždy jen do přípustných stavů, stavový prostor je $\mathcal{N}^+ = \{x \in \mathcal{N} : p(x) > 0\}$.

Tvrzení 32. *Pokud p splňuje podmínku (12) lokální stability, potom Markovův řetězec $\{X_t\}$ je φ-nerozložitelný na $\mathcal{N}^+$ a pro každé $k \in \mathbb{N}_0$ je $C = \{x \in \mathcal{N}^+ : x(E) \le k\}$ malá množina.*

Důkaz: Mějme dáno $k \in \mathbb{N}_0$ a $x \in \mathcal{N}^+$, $0 < x(E) = n \le k$. Označme $\mathcal{N}_n = \{x \subseteq E, x(E) = n\}$. Pravděpodobnost ubrání bodu z x je

$$P(x, \mathcal{N}_{n-1}) = (1-Q(x)) \sum_{\eta\in x} d(x,\eta)\alpha(x, x\setminus\eta) = \frac{1}{2}\sum_{\eta\in x}\frac{1}{n}\min\{1, \frac{n}{\lambda(x\setminus\eta,\eta)\Lambda(E)}\} \ge \frac{1}{2K\Lambda(E)} = c$$

za předpokladu, že K z definice je zvoleno dostatečně velké, aby $\frac{1}{K\Lambda(E)} < 1$. Tedy pro $x \in C$ a $m > k$ je $P^m(x, \{\emptyset\}) \ge c^m$. Zvolme míru $\nu = \delta_\emptyset$, potom $P^m(x, A) \ge c^m\nu(A)$ pro každé $x \in C$ a $A \in \mathfrak{N}$. Pro $n = x(E) = 0$ je $P^m(\emptyset, A) \ge (1-Q(\emptyset))^m\delta_\emptyset(A) = (1/2)^m\nu(A) \ge c^m\nu(A)$. Celkem tak dostáváme, že C je malá ($\varepsilon = c^m$). Podobně položíme-li $\varphi = \delta_\emptyset$, tak $P^m(x, A) \ge c^m > 0$ kdykoli $m \ge x(E)$ a $\varphi(A) > 0$. Řetězec je proto φ-nerozložitelný.

□

Věta 33. *Markovův řetězec pro simulaci bodového procesu s lokálně stabilní hustotou $p(x)$ MH-algoritmem je stejnoměrně ergodický právě tehdy, když existuje m tak, že $\mathcal{N} = \cup_{n=0}^m \mathcal{N}_n$.*

Důkaz: Je-li $\mathcal{N} = \cup_{n=0}^m \mathcal{N}_n$, pak je řetězec stejnoměrně ergodický podle věty 22 a tvrzení 32. Naopak nechť řetězec je stejnoměrně ergodický, tedy $\mathcal{N}$ je malá (věta 22). Existuje proto pravděpodobnostní míra ν a $m \in \mathbb{N}$ tak, že $P^m(x, F) \ge \nu(F)$ pro každé $x \in \mathcal{N}$ a $F \in \mathfrak{N}$. Předpokládejme, že neexistuje m takové, že $\mathcal{N} = \cup_{n=0}^m \mathcal{N}_n$. Ukážeme, že pak $\nu(\mathcal{N}_k) = 0$ pro každé k, což bude spor. Kdyby $\nu(\mathcal{N}_k) > 0$, tak vezmeme $x \in \mathcal{N}_{k+m+1}$ a $P^m(x, \mathcal{N}_k) \ge \nu(\mathcal{N}_k) > 0$, což je spor, neboť $P^m(x, \mathcal{N}_k) = 0$.

□

Věta 34. *Je-li p lokálně stabilní hustota, je příslušný Metropolisův-Hastingsův algoritmus neperiodický a geometricky ergodický.*

Důkaz: Zřejmě $P(\emptyset, \{\emptyset\}) \ge 1 - Q(\emptyset) = \frac{1}{2} > 0$ a odtud plyne neperiodicita. K ověření geometrické ergodicity použijeme větu 21 s $V(x) = c^n$, kde $n = x(E)$ a $c > 1$ je konstanta. Z předpokladu věty je $\lambda(x,\xi) \le K$. Pro $n \ge 1$ platí

$$PV(x) = \int_{\mathcal{N}} V(y)\, P(x, \mathrm{d}y) = c^{n+1}P(x, \mathcal{N}_{n+1}) + c^{n-1}P(x, \mathcal{N}_{n-1}) + c^n P(x, \{x\}).$$

Odhadneme jednotlivé členy:

$$P(x,\mathcal{N}_{n+1}) = Q(x)\int_E \mathbf{1}_{[x\cup\xi\in\mathcal{N}_{n+1}]} b(x,\xi)\alpha(x,x\cup\xi)\,\Lambda(\mathrm{d}\xi)$$
$$= \frac{1}{2}\int_E \mathbf{1}_{[x\cup\xi\in\mathcal{N}_{n+1}]} \min\left\{1,\lambda(x,\xi)\frac{\Lambda(E)}{n+1}\right\}\frac{\Lambda(\mathrm{d}\xi)}{\Lambda(E)} \le \frac{1}{2}\min\left\{1,\frac{K\Lambda(E)}{n+1}\right\} \le \frac{\varepsilon}{2},$$

kde $n+1 \ge \frac{K\Lambda(E)}{\varepsilon}$,

$$P(x,\mathcal{N}_{n-1}) = (1-Q(x))\sum_{\eta\in x} d(x,\eta)\alpha(x,x\setminus\eta) = \frac{1}{2}\sum_{\eta\in x}\frac{1}{n}\min\left\{1,\frac{n}{\Lambda(E)\lambda(x\setminus\eta,\eta)}\right\}$$
$$= \frac{1}{2}\sum_{\eta\in x}\min\left\{\frac{1}{n},\frac{1}{K\Lambda(E)}\right\} = \frac{1}{2}$$

při $n \ge K\Lambda(E)$,

$$P(x,\{x\}) = 1 - P(x,\mathcal{N}_{n+1}) - P(x,\mathcal{N}_{n-1}) \le 1 - P(x,\mathcal{N}_{n-1}) = \frac{1}{2} \quad \text{při } n \ge K\Lambda(E).$$

Položme $N_{K,\varepsilon} = \frac{K\Lambda(E)}{\varepsilon}$, potom pro $n \ge N_{K,\varepsilon}$ a $0 < \varepsilon < 1$ je

$$\int V(y)\,P(x,\mathrm{d}y) \le c^{n+1}\frac{\varepsilon}{2} + c^{n-1}\frac{1}{2} + c^n\frac{1}{2} = c^n\left(\frac{c\varepsilon}{2}+\frac{1}{2c}+\frac{1}{2}\right).$$

Protože $\frac{1}{2c}+\frac{1}{2} < 1$, existuje $\beta < 1$ tak, že pro ε dost malé je $\int V(y)\,P(x,\mathrm{d}y) \le \beta V(x)$ pro $x \notin C$, kde $C = \{x \in \mathcal{N} : x(E) < N_{K,\varepsilon}\}$ je malá množina (tvrzení 32). Dále pro $x \in C$ je $\int V(y)\,P(x,\mathrm{d}y) \le c^{N_{k,\varepsilon}+1} = b$. Je proto splněna podmínka geometrického driftu.

□

8. Různé na závěr

8.1 Rozšíření dat

Rozšíření dat (data augmentation) je technika, na kterou se dá dívat jako na speciální případ Gibbsova výběrového plánu. Dá se využít při práci s chybějícími pozorováními (missing data) nebo v situacích, kdy věrohodnost je v komplikovaném tvaru, ale podmíněně při nepozorovaných datech se stává jednoduchou.

V praxi se ve statistických problémech často setkáváme s chybějícími pozorováními. Nechť x jsou pozorovaná data, y jsou chybějící data a jejich sdružená hustota je $f(x,y \mid \theta)$. Chybějící data mohou být skutečná chybějící pozorování nebo přidané hodnoty (např. skryté nepozorované veličiny), které zjednoduší statistickou inferenci.

Pro dané pozorování x je marginální hustota $f(x \mid \theta)$ funkce neznámého parametru θ (věrohodnostní funkce). Problém je, že vyintegrování chybějících pozorování

$$f(x \mid \theta) = \int f(x,y \mid \theta)\,\mu(\mathrm{d}y)$$

bývá složité nebo většinou nemožné. Pokud apriorní rozdělení parametru θ je $\pi(\theta)$, můžeme však psát

$$\pi(\theta,y \mid x) \propto f(x,y \mid \theta)\pi(\theta), \tag{13}$$

tedy y považujeme za další neznámé parametry a pomocí MCMC generujeme z $\pi(\theta,y \mid x)$, aproximaci aposteriorního rozdělení $\pi(\theta \mid x)$ dostaneme „vynecháním“ y. Příslušný Gibbsův výběrový plán pro rozšíření dat je založen na plně podmíněných rozdělení $\pi(\theta \mid x,y)$ a $\pi(y \mid \theta,x)$, která dostaneme z (13).

Příklad: Směšovací model (mixture model): Nechť $f_1,\dots,f_k$ jsou pravděpodobnostní hustoty a $p_1,\dots,p_k$ jsou nezáporná čísla, $p_1+\cdots+p_k = 1$. Uvažujme náhodnou veličinu X s hustotou $f(x) = p_1f_1(x) + \cdots + p_kf_k(x)$, neboli X má hustotu f_j s pravděpodobností p_j. Pro náhodný výběr $X_1,\dots,X_n$ tak dostáváme sdruženou hustotu $f(x_1)\cdots f(x_n)$, která po roznásobení obsahuje k^n členů, což znemožňuje přímý výpočet pro větší rozsah výběru.

Na celou situaci můžeme pohlížet jako na problém chybějících dat. Chybějící pozorování v tomto případě odpovídají indexu použité hustoty, neboli $X \mid Y = j \sim f_j(x)$ a $P(Y=j) = p_j$. Kdybychom měli informaci o chybějícím pozorování Y, byla by inference přímočará. Náhodný výběr se v takovém případě rozpadá na podvýběry z příslušných hustot f_j.

8.2 MCMC a maximální věrohodnost

V klasické statistice existuje mnoho přístupů k inferenci založené na věrohodnosti (maximálně věrohodné odhady, testy poměrem věrohodností apod.), MCMC nabízí jeden z možných přístupů.

Předpokládejme, že máme pozorování x z hustoty $f_\theta(x) = \frac{1}{c(\theta)} h_\theta(x)$, kterou známe až na normující konstantu $c(\theta)$. Parametr θ je neznámý. Potom logaritmická věrohodnost je $l(\theta) = \log h_\theta(x) - \log c(\theta)$. Výhodnější bude uvažovat poměr věrohodnosti vzhledem k nějakému pevnému parametru ψ:

$$l(\theta) = \log \frac{h_\theta(x)}{h_\psi(x)} - \log \frac{c(\theta)}{c(\psi)}. \tag{14}$$

Zatímco první člen je známý, druhý člen obsahuje neznámé normující konstanty. Pokud $h_\theta(x) = 0$ kdykoli $h_\psi(x) = 0$, tak

$$\begin{aligned} \frac{c(\theta)}{c(\psi)} &= \frac{1}{c(\psi)} \int h_\theta(x)\, \mu(\mathrm{d}x) = \int \frac{h_\theta(x)}{h_\psi(x)} \frac{h_\psi(x)}{c(\psi)}\, \mu(\mathrm{d}x) \\ &= \int \frac{h_\theta(x)}{h_\psi(x)} f_\psi(x)\, \mu(\mathrm{d}x) = \mathbb{E}_\psi \frac{h_\theta(X)}{h_\psi(X)}. \end{aligned}$$

Tuto střední hodnotu můžeme přibližně spočítat pomocí MCMC metod. Pokud $\{X_n\}$ je markovský řetězec s limitním rozdělením daným hustotou f_ψ, pak (14) aproximujeme pomocí

$$l_n(\theta) = \log \frac{h_\theta(x)}{h_\psi(x)} - \log \left(\frac{1}{n} \sum_{i=1}^{n} \frac{h_\theta(X_i)}{h_\psi(X_i)} \right).$$

Maximalizace $l_n(\theta)$ vzhledem k θ dává MCMC aproximaci $\hat{\theta}_n$ maximálně věrohodného odhadu $\hat{\theta}$ parametru θ.

Pozn.: Je zde vidět spojitost s importance samplingem. Procedura funguje dobře, když ψ je blízko $\hat{\theta}$. Jako ψ se většinou volí hrubý odhad $\hat{\theta}$ získaný nějakou jednodušší ale méně efektivní metodou. Celou proceduru lze iterativně opakovat.

8.3 Výběr modelu

Úlohou je na základě pozorovaných dat x, které považujeme za realizaci náhodného elementu X, vybrat jeden model z konečné množiny M možných modelů. Každý model m určuje rozdělení X pomocí vektoru parametrů θ_m, příslušnou věrohodnost označíme $f(x \mid m, \theta_m)$. Pokud apriorní rozdělení pro model m je $\pi(m)$, pak aposteriorní rozdělení je

$$\pi(m \mid x) = \frac{\pi(m) f(x \mid m)}{\sum_{m \in M} \pi(m) f(x \mid m)}, \quad m \in M,$$

kde marginální věrohodnost $f(x \mid m) = \int f(x \mid m, \theta_m) \pi(\theta_m \mid m)\, \nu_m(\mathrm{d}\theta_m)$.

Při výpočtu aposteriorních pravděpodobností $\pi(m \mid x)$ je hlavním problémem nutnost výpočtu integrálu v marginálních věrohodnostech $f(x \mid m)$. Pokud pracujeme s velkým počtem modelů, stává se tento výpočet pro všechny modely prakticky neproveditelný. Metody MCMC nabízejí vhodný nástroj pro vypořádání se s těmito problémy, navíc je pomocí nich možné zkonstruovat algoritmy pro hledání modelu založené na generování z aposteriorního rozdělení $\pi(m, \theta_m \mid x)$. Uvedeme jeden příklad takového algoritmu, který „skáče“ mezi modely s různou dimenzí parametrů a přitom zaručuje splnění detailní podmínky rovnováhy.

Algoritmus 7. *Reverzibilní skok (reversible jump)*:
Nechť současný stav řetězce je (m, θ_m), kde dimenze θ_m je $d(\theta_m)$.

1. navrhni nový model m' s pravděpodobností $j(m, m')$,
2. generuj y (může být nižší dimenze než $\theta_{m'}$) z návrhové hustoty $q(y \mid \theta_m, m, m')$,
3. polož $(\theta'_{m'}, y') = g_{m,m'}(\theta_m, y)$, kde $g_{m,m'}$ je daná invertovatelná funkce (odtud $d(\theta_m) + d(y) = d(\theta_{m'}) + d(y')$ a $g_{m',m} = g^{-1}_{m,m'}$),

4. přijmi přechod do m' s pravděpodobností

$$\alpha = \min\left(1, \frac{f(y \mid m', \theta'_{m'})\pi(\theta'_{m'} \mid m')\pi(m')j(m', m)q(y' \mid \theta'_{m'}, m', m)}{f(y \mid m, \theta_m)\pi(\theta_m \mid m)\pi(m)j(m, m')q(y \mid \theta_m, m, m')} \left|\frac{\partial g_{m,m'}(\theta_m, y)}{\partial(\theta_m, y)}\right|\right).$$

Pozn.: Existuje několik jednodušších verzí algoritmu. Například pokud transformační funkce $g_{m,m'}$ je identita, tak $(\theta'_{m'}, y') = (y, \theta_m)$, $d(\theta_m) = d(y')$, $d(\theta'_{m'}) = d(y)$ a člen s jakobiánem v pravděpodobnosti přijetí je roven 1.

Všimněme si, že pro $m = m'$ se jedná o krok klasického Metropolisova-Hastingsova algoritmu.

Algoritmus 7 (podobně jako příbuzný algoritmus 6 a na rozdíl od Gibbsova výběrového plánu nebo klasického Metropolisova-Hastingsova algoritmu) funguje pro situace s měnící se dimenzí stavů.

Literatura

[1] J. E. Besag (1974): Spatial interaction and the statistical analysis of lattice systems (with discussion), *J. Roy. Statist. Soc. Ser. B* **36**, 192–236.

[2] P. Brémaud (1998): *Markov Chain: Gibbs Fields, Monte Carlo Simulation, and Queues*, Springer, New York.

[3] V. Černý (1985): Thermodynamical approach to the traveling salesman problem: an efficient simulation algorithm, *J. Optim. Theory Appl.* **45**, 41–51.

[4] J. A. Fill (1998): An interruptible algorithm for perfect sampling via Markov chains, *Ann. Appl. Probab.* **8**, 131–162.

[5] D. Gamerman and H. F. Lopes (2006): *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*, Second Edition, Chapman & Hall/CRC, Boca Raton.

[6] A. E. Gelfand and A. F. M. Smith (1990): Sampling-based approaches to calculating marginal densities, *J. Amer. Math. Soc.* **85**, 398–409.

[7] S. Geman and D. Geman (1984): Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images, *IEEE Trans. PAMI* **6**, 721–741.

[8] W. Gilks, S. Richardson and D. Spiegelhalter (1996): *Markov Chain Monte Carlo in Practice*, Chapman & Hall, London.

[9] O. Häggström (2002): *Finite Markov Chains and Algorithmic Applications*, Cambridge University Press, Cambridge.

[10] W. K. Hastings (1970): Monte Carlo sampling methods using Markov chains and their applications, *Biometrika* **57**, 97–109.

[11] E. Ising (1925): Beitrag zur Theorie des Ferromagnetismus, *Z. Physik* **31**, 253–258.

[12] O. Kallenberg (2002): *Foundations of Modern Probability*, Second Edition, Springer-Verlag, New York.

[13] W. S. Kendall, F. Liang and J. S. Wang (2005): *Markov Chain Monte Carlo: Innovations and Applications*, World Scientific, Singapore.

[14] S. Kirkpatrick, C. D. Gelatt, Jr. and M. P. Vecchi (1983): Optimization by simulated annealing, *Science* **220**, 671–680.

[15] D. V. Lindley and A. F. M. Smith (1972): Bayes estimates for the linear model (with discussion), *J. Roy. Statist. Soc. Ser. B* **34**, 1–41.

[16] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller and E. Teller (1953): Equation of state calculations by fast computing machine, *J. Chem. Phys.* **21**, 1087–1091.

[17] S. P. Meyn and R. L. Tweedie (1993): *Markov Chains and Stochastic Stability*, Springer-Verlag, New York.

[18] J. Møller (2003): *Spatial Statistics and Computational Methods*, Lecture Notes in Statistics 173, Springer, New York.

[19] L. Onsager (1944): Crystal statistics. I. A two-dimensional model with an order-disorder transition, *Phys. Rev.* **65**, 117–149.

[20] J. G. Propp and D. B. Wilson (1996): Exact sampling with coupled Markov chains and applications to statistical mechanics, *Random Structures Algorithms* **9**, 223–252.

[21] C. P. Robert (2001): *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*, Second Edition, Springer, New York.

[22] W. Rudin (1977): *Analýza v reálném a komplexním oboru*, Academia, Praha.

[23] D. J. Strauss (1975): A model for clustering, *Biometrika* **62**, 467–475.

[24] L. Tierney (1994): Markov chains for exploring posterior distributions (with discussion), *Ann. Statist.* **22**, 1701–1762.