Metody Markov Chain Monte Carlo (NMTP539)

naposledy upraveno 4. ledna 2023

Obsah

Kapitola 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 markovský 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:

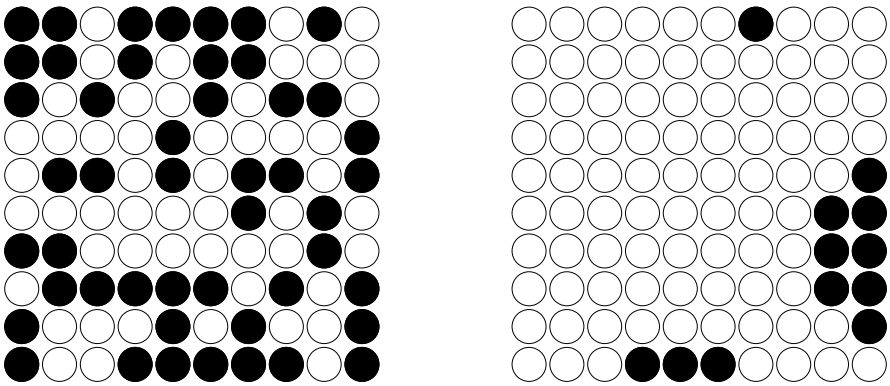
- První MCMC algoritmus byl vyvinut pro aplikace ve statistické fyzice – Metropolis a kol. (1953) [23] simulovali tekutinu v rovnováze s plynnou fází. Aby mohli zkoumat rovnovážný stav, simulovali dynamiku systému, který k němu vede. Nápad: mohu simulovat i jinou dynamiku se stejným rovnovážným stavem.
  Zobecnění algoritmu – Hastings (1970) [13], tzv. Metropolis-Hastingsův algoritmus.
- Řešení optimalizačních problémů metodou simulovaného žíhání – Kirkpatrick, Gelatt a Vecchi (1983) [19], Černý (1985) [3].
- Aplikace na statistické problémy poprvé až v 80. letech 20. století (Gibbsův výběrový plán, nezávisle na předchozím): Geman, Geman (1984) [8] – restaurování digitálních obrázků. Gelfand, Smith (1990) [7] – rozšíření mimo oblast prostorové statistiky do obecné bayesovské statistiky.
- Teorie MCMC až v 90. letech: Geyer (1992) [9], Tierney (1994) [37].

- Zobecnění pro simulování z rozdělení definovaného na sjednocení prostorů různé dimenze – Green (1995) [11], tzv. Metropolis-Hastings-Greenův algoritmus.
- Propp, Wilson (1996) [28] – tzv. perfektní simulace umožňuje simulovat vzorky přímo ze stacionárního rozdělení, ne jen z jeho aproximace.
- V posledních 25 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) [14] – 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 [27], 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.

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).

Ilustrační příklad: dekódování šifrovaných vězeňských zpráv dle [4]: hledá se správná dekódovací funkce f*:* $\{$*šifrovací abeceda*$\} \to \{$*normální abeceda*$\}$. Jazyk je modelován jako markovský řetězec. Pravděpodobnosti přechodu mezi písmeny odhadnuty pomocí četností přes nějaký standartní text a uloženy v matici M. Funkce $Pl(f) = \prod_i M(f(s_i), f(s_{i+1}))$, kde s_i postupně probíhá symboly zakódované zprávy, definuje tzv. přijatelnost funkce f. Dobrá dekódovací funkce f by měla mít vysoké hodnoty Pl. f maximalizující Pl se najde pomocí následujícího MCMC algoritmu:

- Začni s nějakou počáteční hodnotou f
- Spočti $Pl(f)$.
- Změň f na f^* tím, že prohodíš dekódování dvou náhodně zvolených symbolů.
- Spočti $Pl(f^*)$; pokud je větší než $Pl(f)$, přijmi f^*.
- Pokud je $Pl(f^*)/Pl(f) < 1$, přijmi f^* jen s pstí $Pl(f^*)/Pl(f)$, jinak zůstaň v f.

Funguje to. Proč a další podrobnosti viz další kapitoly.

Literatura: Literatura o MCMC je velmi rozsáhlá, zmíníme tedy jen několik základních zdrojů. Výborná je přehledová kniha [2], kde lze rovněž nalézt odkazy na další literaturu. Pro bayesovské aplikace jsou dobré [29], [6], pro použití v prostorové statistice [25], [26]. [18] obsahuje pěkný úvod do perfektní simulace, [12] je dobře čitelná kniha vysvětlující MCMC algoritmy pro problémy na konečných stavových prostorech.

Kapitola 2

Simulace

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

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

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

   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: $\mathbb{P}(F^{-1}(U) \leq x) = \mathbb{P}(U \leq F(x)) = F(x)$.

   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).

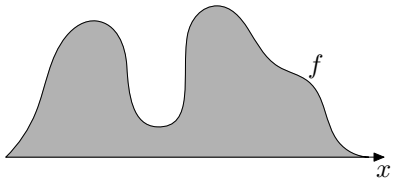
(b) využití interpretace daného rozdělení.
Příklad: Wishartovo rozdělení (zobecnění Γ-rozdělení) – když $X_1, \ldots, X_n$ je výběr z $N_d(\mu, \Sigma)$, tak $\sum_{i=1}^n (X_i - \mu)(X_i - \mu)^T \sim W_d(n/2, \Sigma^{-1}/2)$. Pro $d = 1$ se jedná o $\sigma^2 \chi^2_n$, obecně je $W_1(\alpha, \beta)$ 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 2.2.1. *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 $f(x, u) = f(u \mid x) f(x) = 1$.



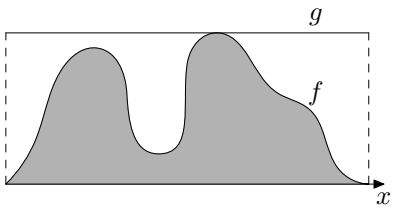
□

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 2.2.2. *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 2.2.3. *Hodnota Z z algoritmu 2.2.2 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.

Příklad: studie robustnosti: $X \sim N(\mu, \sigma^2)$ s pstí $(1-\epsilon)$ a $N(\mu + a, \sigma^2)$ nebo $N(\mu, k\sigma^2)$ s pstí ϵ.

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 a Importance sampling

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 a Sluckého lemma 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

$$\frac{1}{m}\sum_{j=1}^m h(X_j)\frac{f(X_j)}{g(X_j)}, \tag{2.1}$$

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 (2.1) 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 (2.1). Dá se ukázat (z Cauchyovy-Schwarzovy nerovnosti), že rozptyl (2.1) 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 $c\,h(x)f(x)$.
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*:

$$\frac{\sum_{j=1}^m h(X_j)\frac{f^*(X_j)}{g^*(X_j)}}{\sum_{j=1}^m \frac{f^*(X_j)}{g^*(X_j)}}. \tag{2.2}$$

Podle silného zákona velkých čísel konverguje (2.2) 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 (2.2).

Výhoda MCMC je, že místo generování nezávislých vzorků (= realizací X_j), což může být těžké, generujeme markovský řetězec. Závislosti v markovském řetězci způsobí větší rozptyl statistiky (2.1) resp. (2.2), ovšem ten může být snížen větším počtem m vygenerovaných vzorků.

Kapitola 3

Bayesovská statistika

Existuje speciální přednáška *Bayesovské metody.*

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 použít i nevlastní hustotu $\pi(\theta)$, tj. $\int \pi(\theta)\,\nu(\mathrm{d}\theta) = \infty$.

Pokud nemáme představu o apriorním rozdělení, je obvyklou volbou neurčité apriorní rozdělení ($\pi(\theta)$ je konstantní). Ne vždy je ale možné ho použít, protože neurčité nevlastní apriorní rozdělení může vést na nevlastní aposteriorní rozdělení.

Příklad: $X \mid \theta \sim R(0, \theta)$, $\pi(\theta) = 1, \theta > 0$, pak $\int f(x \mid \theta)\pi(\theta)\,\mathrm{d}\theta = \int \frac{1}{\theta}\,\mathrm{d}\theta$.

Jak vypadá neurčité rozdělení rovněž závisí na parametrizaci.

Příklad: Buď $X \mid \theta \sim Binom(\theta, n)$, $\pi(\theta) = 1, \theta \in (0, 1)$. Po reparametrizaci $\psi = \theta^2$ dostanu $\pi(\psi) = \frac{1}{2\sqrt{\psi}}$, což není rovnoměrné rozdělení.

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.

V bayesovské statistice je statistická inference 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 3.2.1. *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

$$f(x \mid \theta) = \exp\left\{\sum_{j=1}^{d} c_j(\theta)T_j(x) + A(\theta) + B(x)\right\}. \tag{3.1}$$

Tvrzení 3.2.1. *Systém hustot*

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

je konjugovaný s $f(x \mid \theta)$ daným (3.1).

Důkaz:

$$\begin{aligned}\pi(\theta \mid x) &\propto f(x \mid \theta)\pi(\theta) \propto \exp\left\{\sum_{j=1}^{d} c_j(\theta)T_j(x) + A(\theta)\right\}\exp\left\{\sum_{j=1}^{d} \alpha_j c_j(\theta) + \beta A(\theta)\right\} \\ &= \exp\left\{\sum_{j=1}^{d}(\alpha_j + T_j(x))c_j(\theta) + (\beta+1)A(\theta)\right\}.\end{aligned}$$

□

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 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 (hierarchical models) je apriorní rozdělení specifikováno ve více stupních.

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

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

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

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

kde $Y = (Y_1, \dots, Y_n)^T$, $\beta = (\beta_1, \dots, \beta_d)^T$, $X = (x_{ij})$ je matice typu $n \times d$, disperze $\tau = \sigma^{-2} > 0$ 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 $\tilde{x}_{ij}$, $i = 1, \dots, d$, $j = 1, \dots, \tilde{d}$ a regresními koeficienty $\tilde{\beta}_1, \dots, \tilde{\beta}_{\tilde{d}}$. Celý model (viz [21]) má potom maticový tvar

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

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.

Kapitola 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 4.0.1. *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

$$\pi(A) = \int_{\mathcal{X}} P(x, A)\, \pi(\mathrm{d}x) \quad \forall A \in \mathfrak{X}. \tag{4.1}$$

Postačující podmínka (nikoli však nutná) pro invarianci π je *reverzibilita (reversibility)* markovského 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

$$f(x)p(x, y) = f(y)p(y, x) \quad \forall x, y \in \mathcal{X}, \tag{4.2}$$

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 (4.2) je známá jako *detailní podmínka rovnováhy (detailed balance condition)*. Není těžké se přesvědčit, že (4.1) je splněna, když (4.2) platí:

$$\begin{aligned}\int_{\mathcal{X}} P(x, A)\, \pi(\mathrm{d}x) = \int_{\mathcal{X}} \int_A p(x, y)\, \mu(\mathrm{d}y)\, f(x)\, \mu(\mathrm{d}x) = \int_{\mathcal{X}} \int_A f(y)p(y, x)\, \mu(\mathrm{d}y)\, \mu(\mathrm{d}x) \\ = \int_A f(y) \left(\int_{\mathcal{X}} p(y, x)\, \mu(\mathrm{d}x) \right) \mu(\mathrm{d}y) = \int_A f(y)\, \mu(\mathrm{d}y) = \pi(A).\end{aligned}$$

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 4.1.1. *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 $x^{(0)}, \ldots, x^{(T)}$ markovského řetězce $X^{(t)}$. 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).*

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

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 $\mathbb{R}^2$ s disjunktními projekcemi. 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 $x^{(0)}$).

Lemma 4.1.2. *(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

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

což znamená, že

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

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 $p^*(y,x) = \prod_{i=1}^d f(x_i \mid y_1,\dots,y_{i-1},x_{i+1},\dots,x_d)$ 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 4.1.3. *Za předpokladů lemmatu 4.1.2 je π invariantní míra vzhledem ke Gibbsovu jádru, tedy splňuje* (4.1).

Důkaz: Využijeme toho, že $f(y)p^*(y,x) = f(x)p(x,y)$ pro každé $x,y \in \mathcal{X}^+$, viz lemma 4.1.2. Pro $A \in \mathfrak{X}$ je

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

□

Algoritmu 4.1.1 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 4.1.4. *Náhodné procházení v Gibbsově výběrovém plánu (random scan Gibbs sampler):*

1. *zvol počáteční stav $x^{(0)} = (x_1^{(0)},\dots,x_d^{(0)})$, polož $t = 0$,*
2. *vygeneruj k z rovnoměrného rozdělení na množině $\{1,\dots,d\}$ a simuluj $x_k^{(t+1)}$ z podmíněného rozdělení $X_k \mid x_1^{(t)},\dots,x_{k-1}^{(t)},x_{k+1}^{(t)},\dots,x_d^{(t)}$, 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 4.2.1. *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. Tedy p_0 splňuje detailní podmínku rovnováhy (4.2). 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 4.2.2. *Cílové rozdělení π s hustotou f je invariantní pro Metropolisovo-Hastingsovo jádro.*

Důkaz: Pro $A \in \mathfrak{X}$ je

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

Ve třetí rovnosti jsme použili Fubiniovu větu a ve čtvrté podmínku detailní rovnováhy pro p_0. □

Pozn.: Alternativní důkaz věty spočívá v důkazu, že Metropolis-Hastingsovo jádro P je reverzibilní vzhledem k π, neboť reverzibilita vzhledem k π implikuje invarianci vzhledem k π. Reverzibilita je ovšem splněna, neboť pro spojitou část jádra P plyne reverzibilita z podmínky detailní rovnováhy pro p_0, a diskrétní část jádra $r(x)\delta_x(\mathrm{d}y)$ je triviálně reverzibilní vzhledem k libovolnému pravděpodobnostnímu rozdělení (ověřte).

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. [23], kde lze rovněž nalézt heuristický důkaz konvergence. Přínos Hastingse ([13]) 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 (algoritmus 2.2.2) 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

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

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*.

Různá markovská jádra definovaná na tomtéž prostoru s tímtéž stacionárním rozdělením π lze kombinovat pomocí skládání a míchání (nezávislého na aktuálním stavu) a vytvářet tak nová markovská jádra (a jim odpovídající MCMC algoritmy) s tímtéž stacionárním rozdělením π. Podrobněji s k tomuto tématu vrátíme v dalších kapitole.

Ilustrační příklad: dekódování šifrovaných vězeňských zpráv, část 2:
Máme k dispozici řetězec $S = \{s_1, \dots s_l\}$ znaků šifrovací abecedy, který obsahuje zašifrované sdělení. Na prostoru $\mathcal{X}$ možných dekódovacích funkcí f*:* $\{$*šifrovací abeceda*$\} \to \{$*normální abeceda*$\}$ se snažíme najít tu správnou. Pokud je velikost šifrovací abecedy m a normální abecedy $n \geq m$, tak možných f z $\mathcal{X}$ je $n!/(n-m)!$. Tedy prostor $\mathcal{X}$ je velký. Pro nalezení té správné f zadefinujeme vhodný pravděpodobnostní model a provedeme jeho bayesovskou analýzu.
Model: pozorovaná zašifrovaná zpráva S (data) je kus anglického textu překódovaný pomocí f. Anglický text budiž realizace markovského řetězce s hodnotami v normální abecedě, počátečním rozdělením daným vektorem v a maticí pravděpodobností přechodu M. Pro M a v použijeme jejich odhady pomocí četností přes nějaký standartní (a dost dlouhý) text a budeme je považovat za známé. Dešifrovací funkce f budiž neznámý parametr. Věrohodnost pozorovaných dat S je pak rovna

$$L(S|f) = v(f(s_1)) \prod_{i=1}^{l} M(f(s_i), f(s_{i+1})).$$

Apriorní rozdělení pro f volíme rovnoměrné na $\mathcal{X}$. Aposteriorní rozdělení $f|S$ bude potom až na konstantu rovno $L(S|f)$.
V řešení dle [4] zanedbali počáteční pravděpodobnosti v, resp. podmínili data prvním pozorovaným znakem s_1. Tento rozdíl má ale na výsledek dekódování zanedbatelný vliv. Aposteriorní rozdělení f potom přesně odpovídá přijatelnosti $Pl(f) = \prod_i M(f(s_i), f(s_{i+1}))$.
Pro generování z rozdělení daného $Pl(f)$ se použije Metropolisův algoritmus symetrické náhodné procházky na $\mathcal{X}$ s návrhovým jádrem

$$Q(f, f^*) = \begin{cases} \frac{1}{m(m-1)(n-m+1)(n-m+2)} & \text{pro } f, f^* \text{ různé v max. dvou symbolech} \\ 0 & \text{jinak} \end{cases}$$

Pravděpodobnost přijetí pak bude $\alpha(f, f^*) = \min\{Pl(f^*)/Pl(f), 1\}$. Což přesně odpovídá algoritmu ze strany 3. Pokud je náš zašifrovaný řetězec S dostatečně dlouhý, bude aposteriorní rozdělení dané $PL(f)$ silně unimodální a koncentrované v okolí správné dešifrovací funkce f. Generovaný MCMC řetězec se tedy po nějaké době ustálí v okolí tohoto modu a popsaný postup nalezne správnou dešifrovací funkci f.

Kapitola 5

Markovské 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 5.1.1. *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 markovský 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 (homogeneous) řetězci.

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

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

$$\mathbb{P}(X_0 = i_0, \ldots, X_n = i_n) = p_{i_0} p_{i_0 i_1} \cdots p_{i_{n-1} i_n}, \tag{5.1}$$

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 řetězci 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

$$\sum_{j \in S} |p_{ij}^{(n)} - \pi_j| \leq V(i) \lambda^n \quad \forall i \in S, n \in \mathbb{N}. \tag{5.2}$$

Nejmenší λ, pro něž existuje funkce V splňující (5.2), se značí λ^* a nazývá se *geometrický řád konvergence (geometric rate of convergence)*.

Nechť existují vlastní čísla $|\lambda_0| \geq |\lambda_1| \geq \dots$ a vlastní levé vektory $e_0, e_1, \dots$ 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 v absolutní hodnotě – SLEM (second largest eigenvalue modulus). Toto tvrzení je důsledkem Perronovy-Frobeniovy věty, která udává tvar matice P^n.

Buď funkce h taková, že $\mathbb{E}_\pi h^2(X) < \infty$. Potom 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 [24]. Dobrým zdrojem asymptotických výsledků se zřetelem k jejich využití pro MCMC jsou také přehledové články [37, 15, 32].

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 (5.1), ve kterém jsou konečněrozměrná rozdělení vyjádřena pomocí pravděpodobností přechodu.

Věta 5.2.1. *Buď dáno markovské jádro P na $(\mathcal{X}, \mathfrak{X})$ a pravděpodobnostní rozdělení ϱ na $\mathfrak{X}$. Pak existuje náhodný proces $\{X_n, n \in \mathbb{N}_0\}$ takový, že*

$$\mathbb{P}(X_0 \in A_0, \dots, X_n \in A_n) = \int_{A_0} \cdots \int_{A_{n-1}} P(y_{n-1}, A_n) P(y_{n-2}, \mathrm{d}y_{n-1}) \cdots P(y_0, \mathrm{d}y_1) \varrho(\mathrm{d}y_0) \tag{5.3}$$

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

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

□

Definice 5.2.1. *Řekneme, že náhodný proces $\{X_n\}$ s obecnou množinou stavů $\mathcal{X}$ je homogenní markovský řetězec (homogeneous Markov chain) s přechodovým jádrem P a počátečním rozdělením ϱ, pokud jeho konečněrozměrná rozdělení splňují* (5.3) *pro každé $n \in \mathbb{N}_0$ a pro všechna $A_0, \dots, A_n \in \mathfrak{X}$.*

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 5.2.2. *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*

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

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

Definice 5.2.2. *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 5.2.3. *(Chapmanova-Kolmogorovova rovnost) Pro $n, m \in \mathbb{N}_0$ a $m \le n$ platí*

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

Důkaz: V (5.3) stačí položit $\varrho = \delta_x$, $A_i = \mathcal{X}$, $i = 0, \dots, n-1$ a $A_n = A$. Definice P^m a P^{n-m} se použije pro prvních m a posledních $n - m$ integrandů.

□

Definice 5.2.3. *Pravděpodobnostní rozdělení π na $\mathfrak{X}$ nazveme limitní rozdělení (limiting distribution) markovského řetězce $\{X_n\}$ generovaného P, jestliže*

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

Pro dané počáteční rozdělení ϱ je $\mathbb{P}(X_n \in A) = \int_{\mathcal{X}} P^n(x, A)\, \varrho(\mathrm{d}x)$, tedy pokud $\rho \ll \pi$, bude platit i $\lim_{n\to\infty} \mathbb{P}(X_n \in A) = \pi(A)$.

Na definici stacionárního rozdělení jsme již narazili, viz (4.1).

Definice 5.2.4. *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 markovského řetězce s přechodovým jádrem P.

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

Tvrzení 5.2.4. *Je-li π limitní rozdělení, potom je i 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 5.2.5. *Markovský řetězec generovaný přechodovým jádrem P je reverzibilní (reversible) vzhledem k π, jestliže pro každé $A, B \in \mathfrak{X}$ platí*

$$\int_A P(x, B)\, \pi(\mathrm{d}x) = \int_B P(x, A)\, \pi(\mathrm{d}x). \tag{5.4}$$

Tvrzení 5.2.5. *Je-li markovský řetězec reverzibilní vzhledem k π, potom π je stacionární rozdělení.*

Důkaz: Stačí položit $A = \mathcal{X}$ v (5.4).

□

Definice 5.2.6. *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 5.2.7. *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 $L(x, A) > 0$.*
Řetězec je nerozložitelný (irreducible), je-li φ-nerozložitelný pro nějaké φ.

Příklad: Řetězec se spočetně mnoha stavy, který není nerozložitelný v diskrétní definici, může být φ-nerozložitelný. Uvažujme náhodnou procházku s absorpčním stavem 0, tedy $p_{0,0} = 1$, $p_{i,i+1} = p \in (0, 1)$ 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 5.2.8. *Pro $0 < \varepsilon < 1$ a markovský řetězec s přechodovým jádrem P definujeme rezolventu (resolvent) jako $K_\varepsilon(x, A) = (1 - \varepsilon) \sum_{n=0}^{\infty} \varepsilon^n P^n(x, A)$.*

Pozn.: Snadno nahlédneme, že $K_\varepsilon(x, A) > 0$ právě tehdy, když $L(x, A) > 0$, a to platí pro všechna $\epsilon > 0, x \in \mathcal{X}, A \in \mathfrak{X}$.

Věta 5.2.6. *Nechť je markovský řetězec $\{X_n\}$ 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 $\bar{A}(k) = \{y : \sum_{n=1}^{k} P^n(y, A) > \frac{1}{k}\}$.

ad i) 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

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

Tedy $\psi(A) > 0$ implikuje existenci k takového, že $\varphi(\bar{A}(k)) > 0$. Potom z φ-nerozložitelnosti (pro každé x musí existovat m takové, že $P^m(x, \bar{A}(k)) > 0$) 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.

ad ii) Nechť $\{X_n\}$ je φ'-nerozložitelný. Je-li $\varphi'(A) > 0$, je $\sum_{n=0}^{\infty} P^n(y, A) > 0$ a tedy i $\psi(A) > 0$ 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.: Vždy, když budeme v dalším 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 5.2.7. *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. Navíc π je také maximální míra nerozložitelnosti $\{X_n\}$.*

Důkaz: [37]

Definice 5.2.9. *Markovský řetězec, který je ψ-nerozložitelný a ve kterém existuje stacionární rozdělení, se nazývá nenulový (positive).*

Pozn.: Pro obecnou definici aperiodicity viz [24], v případě existence stacionárního rozdělení π nám stačí následující:

Definice 5.2.10. *Markovský řetězec $\{X_n\}$ se stacionární rozdělením π 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).*

Pozn.: Je-li markovský řetězec π-nerozložitelný a periodický, pak nutně $\pi(\mathcal{X} \setminus \bigcup_{i=1}^{q} A_i) = 0$ neboť jinak dostaneme spor s nerozložitelností pro jakékoli $x \in \bigcup_{i=1}^{q} A_i$.

Definice 5.2.11. *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 5.2.8. *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: [37]

Tvrzení opravdu neplatí pro všechna x, jak ukazuje následující příklad.

Příklad: Buď $\mathcal{X} = \mathbb{N}$, $P(1,\{1\}) = 1$, $P(x,\{1\}) = \frac{1}{x^2}$, $P(x,\{x+1\}) = 1 - \frac{1}{x^2}, x \geq 2$. Stacionární rozdělení je zřejmě $\pi(\cdot) = \delta_1(\cdot)$ a řetězec je π-nerozložitelný. Ale pro $x_0 = x \geq 2$ je

$$P[X_n = x + n \ \forall n \ ] = \prod_{j=x}^{\infty} \left(1 - \frac{1}{j^2}\right) > 0,$$

takže $\|P^n(x,\cdot) - \pi(\cdot)\|_{TV} \not\to 0$. Tvrzení věty 5.2.8 tedy platí jen pro $x = 1$, což je ale π-skoro všude.

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

Definice 5.2.12. *Pro $A \in \mathfrak{X}$ položme $\eta_A = \sum_{n=1}^{\infty} I_{[X_n \in A]}$ počet navštívení množiny A, tzv. čas obsazení (occupation time). Řekneme, že množina A je trvalá (recurrent), jestliže*

$$U(x,A) = \mathbb{E}\left[\eta_A \mid X_0 = x\right] = \sum_{n=1}^{\infty} P^n(x,A) = \infty$$

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.

Věta 5.2.9. *Nenulový markovský řetězec je trvalý.*

Důkaz: [37]

Definice 5.2.13. *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.: Lze ukázat, že ekvivalentně je možné 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 $U(x,A) = \mathbb{E}\left[\eta_A \mid X_0 = x\right] = \infty$, a tudíž je A i trvalá.

Definice 5.2.14. *Je-li řetězec $\{X_n\}$ harrisovsky trvalý, neperiodický a nenulový, nazývá se ergodický (ergodic).*

Podle tvrzení 5.2.4 je stacionární rozdělení přirozený kandidát pro limitní rozdělení. Pro ergodický řetězec je stacionární rozdělení limitní. Na rozdíl od věty 5.2.8 máme konvergenci pro všechna x.

Věta 5.2.10. *Nechť je markovský řetězec $\{X_n\}$ se stacionárním rozdělením π ergodický, pak*

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

Důkaz: [24], Theorem 13.3.3.

Pozn.: Předpoklady jsou postačující a nutné – viz [37].

Harissovská trvalost nám zajistí neexistenci výjimečných bodů ve větě 5.2.10, je proto dobré umět ji identifikovat. V tom může být nápomocná následující věta.

Definice 5.2.15. *Nezáporná funkce h je harmonická (harmonic) pro markovské jádro P, pokud $h = Ph$.*

Věta 5.2.11. *Trvalý řetězec je harrisovsky trvalý, právě když každá omezená harmonická funkce pro přechodové jádro řetězce P je konstantní.*

Důkaz: [37]

Uvažujme nyní měřitelnou funkci $h : \mathcal{X} \to \mathbb{R}$. V praxi jsou často výstupem MCMC průměry $\bar{h}_n = \frac{1}{n}\sum_{i=1}^{n} h(X_i)$, 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 5.2.12. *(silný zákon velkých čísel, ergodická věta) Nechť $\{X_n\}$ je ergodický markovský řetězec, potom pro libovolnou funkci h splňující $\mathbb{E}_\pi|h(X)| < \infty$ platí*

$$\mathbb{P}\left(\lim_{n\to\infty} \bar{h}_n = \mathbb{E}_\pi h(X)\right) = 1 \quad \textit{pro libovolné počáteční rozdělení } \varrho. \tag{5.5}$$

Důkaz: [24]

Pozn.: Věta 5.2.12 platí i bez předpokladu aperiodicity. I ve větě 5.2.10 je možno aperiodicitu vypustit – dostaneme potom tvrzení pro $\|\frac{1}{q}\sum_{i=n}^{n+q-1} P^i(x,\cdot) - \pi(\cdot)\|_{TV}$ místo $\|P^n(x,\cdot) - \pi(\cdot)\|_{TV}$.

Pokud bychom chtěli kontrolovat i rychlost konvergence v (5.5) (tj. chtěli bychom mít CLV), potřebujeme nějaký silnější pojem ergodicity než z definice 5.2.14.

Definice 5.2.16. *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 5.2.17. *Řekneme, že ergodický markovský řetězec $\{X_n\}$ se stacionárním rozdělením π je geometricky ergodický (geometrically ergodic), existuje-li $\lambda \in [0,1)$ a reálná π-integrovatelná funkce V na $\mathcal{X}$ tak, že*

$$\|P^n(x,\cdot) - \pi(\cdot)\|_{TV} \le V(x)\lambda^n \quad \textit{pro každé } x \in \mathcal{X},\ n \in \mathbb{N}. \tag{5.6}$$

Infimum takových λ se nazývá geometrický řád konvergence (geometric rate of convergence) markovského 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 (a ano, stejnoměrná ergodicita zaručuje, že $\|P^n - \pi\|_{TV}$ jde k 0 geometricky rychle).

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

Věta 5.2.13. *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*

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

pro libovolné počáteční rozdělení.

Důkaz: [24]

Pozn.: Existence σ_h^2 je zajištěna geometrickou ergodicitou, a za tohoto předpokladu lze vyjádřit několika způsoby. Pokud platí CLV pro $\{h(X_n)\}$, pak

$$\sigma_h^2 = \lim_{n\to\infty} \frac{1}{n}\,\mathbb{E}\left[\left(\sum_{i=1}^{n}(h(X_i) - \pi(h))\right)^2\right], \tag{5.7}$$

a to pro libovolné počáteční rozdělení. Rovněž platí

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

pro silně stacionární řetězec s počátečním rozdělením rovným stacionárnímu. Alternativně lze psát $\sigma_h^2 = \tau_h \operatorname{var}_\pi h(X)$, kde

$$\tau_h = 1 + 2\sum_{i=1}^{\infty} \operatorname{cor}_\pi(h(X_0), h(X_i)),$$

se nazývá integrovaný autokorelační čas.

Pozn.: σ_h^2 určuje velikost takzvané MCMC chyby, které se dopouštíme, když používáme MCMC aproximaci $\bar{h}_n$ místo $\mathbb{E}_\pi h(X)$.

Že pro CLV nestačí obyčejná ergodicita, ale je třeba silnější vlastnost, ukazuje následující věta.

Věta 5.2.14. *Buď $\{X_n\}$ markovský řetězec s počátečným rozdělením rovným stacionárnímu rozdělení π a buď $r(x) = P[X_{n+1} = X_n | X_n = x]$. Pokud*

$$\lim_{n\to\infty} n \int_{\mathcal{X}} (h(x) - \pi(h))^2 r^n(x) \pi(\mathrm{d}x) = \infty,$$

pak neplatí CLV pro $\{h(X_n)\}$.

Důkaz: Spočítejme (5.7) (pokud platí CLV, tak tato limita musí existovat):

$$\begin{aligned}\sigma_h^2 &= \lim_{n\to\infty} \frac{1}{n} \mathbb{E}\left[\left(\sum_{i=1}^{n} (h(X_i) - \pi(h))\right)^2\right] \geq \lim_{n\to\infty} \frac{1}{n} \mathbb{E}\left[\left(\sum_{i=1}^{n} (h(X_i) - \pi(h))\right)^2 I(X_0 = X_1 = \cdots = X_n)\right] \\ &= \lim_{n\to\infty} \frac{1}{n} \mathbb{E}\left[(n(h(X_0) - \pi(h)))^2 (r(X_0))^n\right] = \lim_{n\to\infty} n \int_{\mathcal{X}} (h(x) - \pi(h))^2 r^n(x) \pi(\mathrm{d}x) = \infty,\end{aligned}$$

a tedy CLV nemůže platit. □

Pomocí této věty lze například dokázat, že MH-algoritmus s nezávislými návrhy pro generování z $Exp(1)$ s návrhy $Q(x, \cdot) \sim Exp(\lambda)$ s $\lambda > 2$ nesplňuje CLV pro $h(X_n) = X_n$ (viz [30]).

Pozn.: Pro stejnoměrnou ergodicitu stačí pro platnost CLV $\pi(h^2) < \infty$.

Abychom zformulovali postačující podmínky pro geometrickou a stejnoměrnou ergodicitu budeme potřebovat zavést minorizační podmínku a malou množinu.

Definice 5.2.18. *Řekneme, že φ-nerozložitelný markovský 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 5.2.19. *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 $\varepsilon > 0$.*

Věta 5.2.15. *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: [24]

Pozn.: Heuristicky minorizační podmínka říká, že všechny m-krokové přechody z C mají ϵ-ový překryv (komponentu míry ϵ společnou).

Pozn.: Uvědomme si, že pro spočetné $\mathcal{X}$ a pokud

$$\epsilon_{n_0} = \sum_{y\in\mathcal{X}} \inf_{x\in C} P^{n_0}(x, \{y\}) > 0,$$

pak C je $(n_0, \epsilon_{n_0}, \nu)$-malá, kde $\nu(\{y\}) = \epsilon_{n_0}^{-1} \inf_{x\in C} P^{n_0}(x, \{y\})$. Samozřejmě pro nerozložitelný a neperiodický řetězec na konečném prostoru máme vždy $\epsilon_{n_0} > 0$ pro n_0 dost velké.

Pozn.: Obdobně, pokud má P^{n_0} hustotu vzhledem k nějaké míře $\eta(\cdot)$ (tj. $P^{n_0}(x, \mathrm{d}y) = p_{n_0}(x, y)\,\eta(\mathrm{d}y)$), pak můžeme vzít $\epsilon_{n_0} = \int_{y\in\mathcal{X}}(\inf_{x\in C} p_{n_0}(x, y))\eta(\mathrm{d}y)$.

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 markovský řetězec. Protože $P(0, \{0\}) = F(0) > 0$, je $\psi(\{0\}) > 0$. 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

$$P(x, B) = \mathbb{P}(Z_k \in B - x) \ge \int\limits_{B-x} f(y)\,\mathrm{d}y \ge \delta\lambda^1(B), \tag{5.8}$$

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 (5.8), splňuje $M(1, \delta\beta, C, \frac{\lambda^1|_C}{\beta})$.

Vlastnost geometrické ergodicity závisí na exkurzích z centrální (nějaké malé) množiny.

Definice 5.2.20. *Nechť V je měřitelná nezáporná funkce na $\mathcal{X}$. Operátor driftu (drift operator) Δ pro V a markovský řetězec s přechodovým jádrem P je definován jako $\Delta V(x) = PV(x) - V(x)$. Hodnota $\Delta V(x)$ se nazývá drift.*

Definice 5.2.21. *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*

$$\Delta V(x) \le (\lambda - 1)V(x) + bI_C(x), \tag{5.9}$$

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

Pozn.: Podmínka (5.9) říká, že mimo C se V chová jako supermartingal ($\mathbb{E}V(x)$ klesá), $\Delta V(x)$ je mimo C záporná. V se tedy může obnovit jen v C (neboť $V \ge 1$ a nemůže tedy klesat pořád) a $\{X_n\}$ je proto nucen malou množinu C opakovaně navštěvovat.

Pozn.: Pro $\mathcal{X} = \mathbb{R}^d$ jsou často omezené podmnožiny $\mathcal{X}$ malé pro P. Potom pro splnění (5.9) stačí ukázat, že pro nějakou V platí $\lim_{|x|\to\infty} \frac{PV(x)}{V(x)} < 1$.

Věta 5.2.16. *Markovský 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: [24], Chapter 15 nebo pomocí couplingových metod v [32].

Věta 5.2.17. *Markovský řetězec je stejnoměrně ergodický právě tehdy, když $\mathcal{X}$ je malá množina. Řád konvergence je menší nebo roven $(1 - \varepsilon)^{1/m}$.*

Důkaz: [37]

5.3 Ergodicita MCMC algoritmů

Víme, že pro ergodický markovský řetězec máme konvergenci ke stacionárnímu rozdělení. V kapitole 4 jsme uvedli příklady konstrukce řetězců s předepsaným stacionárním rozdělením. Abychom měli zajištěnu konvergenci tohoto řetězce, potřebujeme ověřit nerozložitelnost a neperiodicitu. Ty se u většiny MCMC algoritmů ověří snadno. O dost horší je to u ověřování geometrické (resp. stejnoměrné) ergodicity, které nám zajišťují platnost CLV pro ergodické průměry.

Začněmě některými postačujícími podmínkami, za kterých dostaneme konvergenci k cílovému rozdělení v případě Gibbsova výběrového plánu a Metropolisova-Hastingsova algoritmu.

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

Věta 5.3.1. *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 4.1.1) je μ-nerozložitelný a neperiodický.*

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

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

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 4.1.3 je π stacionární rozdělení a podle věty 5.2.8 je rovněž limitní.

□

Pozn.: Existují i jiné postačující podmínky pro nerozložitelnost Gibbsova výběrového plánu. Například Gibbsův výběrový plán pro cílové rozdělení definované na $\mathcal{X}^+ \subseteq \mathbb{R}^d$, které má na $\mathcal{X}^+$ kladnou hustotu vzhledem k Lebesgueově míře a pro které je $Int(\mathcal{X}^+)$ souvislá množina, je π-nerozložitelný.

Věta 5.3.2. *Uvažujme markovský řetězec $\{X_n\}$ generovaný Metropolisovým-Hastingsovým algoritmem (algoritmus 4.2.1). 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 je návrhové jádro Q neperiodické nebo nechť $\pi(\{x : P(x, \{x\}) > 0\}) > 0$, potom je $\{X_n\}$ neperiodický markovský řetězec.

(ii) Pokud je jádro Q 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 je π 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. Např. pokud je $\mathcal{X} = \mathbb{R}^d$ a q_0 je kladná všude na $\mathcal{X}$, nebo 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 věty 5.2.11.

Důsledek 5.3.3. *Nechť $\{X_n\}$ je φ-nerozložitelný markovský 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: [37]

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 5.3.4. *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 5.2.17.

□

Tvrzení 5.3.5. *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 5.2.17:

$$\alpha(x, y)q(x, y) = q_0(y) \min\left\{1, \frac{f(y)q_0(x)}{q_0(y)f(x)}\right\} = \begin{cases} q_0(y) & \text{pro } w(y) \geq w(x) \\ f(y)\frac{1}{w(x)} & \text{pro } w(y) < w(x) \end{cases} \geq f(y)\frac{1}{\sup_{x \in \mathcal{X}^+} w(x)}$$

□

Pro ověřování geometrické ergodicity lze použít podmínku geometrického driftu nebo couplingových metod. Pro mnohé standardní algoritmy je dokázáno, kdy je geometrická ergodicita splněna. Např. v [22] bylo dokázáno, že Metropolisův algoritmus symetrické náhodné procházky na $\mathbb{R}^d$ je v principu geometricky ergodický právě když má π konečné exponenciální momenty. Pro algoritmy, které nejsou geometricky ergodické, je možné studovat tzv. polynomiální ergodicitu, která umí rovněž zajistit platnost CLV.

Pozn.: Je dobré mít na paměti, že geometrická ergodicita je kvalitativní vlastnost a jako taková má jen omezenou vypovídací schopnost o konvergenci daného MCMC algoritmu. Vždy záleží na konstantách v (5.6). Uvažme třeba následující příklady z [32].

Příklad: MH-algoritmus s nezávislými návrhy pro cílové rozdělení $\pi \sim Exp(1)$ s návrhovým rozdělením $Q(x, \cdot) \sim Exp(\lambda), \lambda > 0$. Chování algoritmu závisí na hodnotě λ (viz např. [31]). Pro $0 < \lambda \leq 1$ je algoritmus geometricky ergodický s geometrickým řádem konvergence $(1 - \lambda)$, platí CLV a simulace se chovají

rozumně i pro malé hodnoty λ. Pro $\lambda > 1$ není algoritmus geometricky ergodický a pro $\lambda > 2$ dokonce ani neplatí CLV pro ergodické průměry $\{X_n\}$. Tedy algoritmus je prakticky nepoužitelný. Na simulacích to ale nemusí být vidět, pokud nenecháme řetězec běžet dostatečně dlouho. Pokud ano, objevuje se "zasekávání"ve stavech s vyšší hodnotou x_n a to na libovolně velký počet iterací (viz cvičení). V tomto případě tedy geometrická ergodicita odpovídá rozumnému chování MH-algoritmu.

Příklad: "Čarodějnický klobouk". Buď $\mathcal{X} = [0, 1]$, $\delta = 10^{-100}$ (řekněme), $0 < a < 1-\delta$ a uvažujme cílovou hustotu $\pi(x) \propto \delta + I_{[a,a+\delta]}(x)$. V tom případě $\pi([a, a + \delta]) \approx \frac{1}{2}$. Uvažujme Metropolisův algoritmus se symetrickou náhodnou procházkou. Pokud není $X_0 \in [a, a+\delta]$ nebo pokud nemá algoritmus to štěstí, že se návrh trefí do miniaturního intervalu $[a, a+\delta]$ pro nějaké rozumně malé n, pak celý (konečně dlouhý) běh algoritmu zůstane mimo interval $[a, a + \delta]$. Bude se tedy prostému oku (i jakémukoli statistickému testu aplikovanému na výstup z algoritmu) jevit jako dobře mixující algoritmus konvergující k rovnoměrnému rozdělení na $[0, 1]$, což je ale rozdělení velmi rozdílné od π. Přestože je algoritmus geometricky ergodický (dokonce stejnoměrně ergodický), v tomto případě nefunguje dobře.

Příklad: Buď $\mathcal{X} = \mathbb{R}$, $\pi(x) \propto \frac{1}{1+x^2}$ a uvažujme Metropolisův algoritmus s náhodnou procházkou (např. $X_0 = 0$, $Q(x, \cdot) \sim R[x-1, x+1]$). Ten je ergodický, ale není geometricky ergodický. A opravdu, simulace vypadají, že konvergují velmi pomalu. Pokud ale zadefinujeme nenormované cílové rozdělení $\pi'(x) = \pi(x)I_{|x|<10^{100}}$, pak ten samý algoritmus pro generování z π' bude geometricky (dokonce stejnoměrně) ergodický, ovšem běhy obou algoritmů budou (v každém reálně dosažitelném počtu iterací) nerozlišitelné.

A potom samozřejmě existuje mnoho MCMC algoritmů na konečných stavových prostorech (a tedy stejnoměrně ergodických) (např. Gibbsův výběrový plán pro Isingův model s nízkou teplotou a velkou mříží), které konvergují extrémně pomalu. Bylo by tedy dobré mít k dispozici kvantitativní výsledky o rychlosti konvergence – viz např. [15]. Bohužel, pro většinu reálných aplikací nejsou takové výsledky zatím dostupné.

Kapitola 6

Praktické aspekty MCMC

Algoritmy

K dispozici máme několik MCMC algoritmů. Je vhodné uvažovat různé algoritmy a vybrat ten, který je nejlepší pro daný problém.

Konvergence

Je důležité ověřit, že příslušný markovský řetězec je nerozložitelný a neperiodický. To zajišťuje konvergenci ke stacionárnímu rozdělení.

Pokud to je možné, je žádoucí nalézt algoritmus, o kterém lze ukázat, že je geometricky ergodický. To zajišťuje geometricky rychlou konvergenci a platnost CLV pro ergodické průměry. Rovněž to typicky odpovídá „rozumnému“ chování algoritmu (viz konec předchozí kapitoly).

Diagnostika konvergence, ladění algoritmu

Neexistuje žádné absolutní číslo, které by udávalo, kolik iterací je potřeba k dosažení stacionárního rozdělení (s danou přesností). Ověření konvergence trvá (mnohem) déle než samotná konvergence řetězce. Když nelze určit řád konvergence přesnými ani přibližnými metodami, užívá se nějaká metoda pro diagnostiku konvergence. V literatuře lze nalézt různé empirické metody (viz např. [2, Chapter 6]), které však nemůžou zaručit dostatečně dobrou konvergenci, protože jsou založeny na pozorováních řetězce. Diagnostika založena na teoretických výsledcích bývá velmi obtížná a v praxi se příliš nepoužívá (ale srovnej např. [15]).

Protože žádná „diagnostika konvergence“ nedokáže zaručit skutečnou konvergenci, zahrnujeme dále popsané jednoduché metody spíše pod pojem ladění MCMC algoritmu. A je dobré, před tím, než spustíme finální běh řetězce, takovéto ladění provést. Důležitou empirickou metodou je monitorování výstupu řetězce – tj. vykreslení průběhů celého řetězce, jeho marginálů nebo jiných charakteristik, pokud je stavový prostor příliš složitý. Pro představu o autokorelacích v řetězci se odhadne klasickým způsobem autokorelační funkce a kroskorelační funkce mezi jednotlivými marginály $\{X_n\}$. Pomalý pokles autokorelací implikuje pomalé mixování řetězce, což není vhodné pro odhady (je potřeba dlouhý běh řetězce, aby byl prozkoumán celý prostor). Pro představu o tom, jak dlouho nechat řetězec běžet, je také možné spustit nejprve více kratších řetězců z různých startovacích hodnot a provést vizuální kontrolu výstupů. To dává lepší náhled na konvergenci a mixování řetězce.

Pro ty, kteří budou MCMC skutečně používat, doporučujeme přečíst velmi pěknou diskusi o konvergenci, pseudokonvergenci a „black box MCMC“ od Ch. Geyera v [2, Chapter 1.11]. Alternativně jsou tyto úvahy dostupné i na webu Ch. Geyera.

Nebezpečí MCMC

Pomocí MCMC je možné počítat s prakticky libovolným modelem, který si vymyslíme, bez ohledu na to, jestli je statisticky rozumný. „Nerozumnost“ použitého modelu, resp. jeho nevhodnost pro analyzovaná data se může projevit i špatnou konvergencí MCMC algoritmu (zbytečně multimodální aposteriorní rozdělení atp.). Je-li to možné, je vhodné porovnat výsledky naší analýzy třeba s předchozími výsledky v jednodušším modelu, nebo provést „fake data check“.

Druhé velké nebezpečí spočívá v tom, že MCMC algoritmus neříká nic o částech stavového prostoru nenavštívených řetězcem – takže jsme zpátky u otázky konvergence a pseudokonvergence.

Reparametrizace, rozšíření dat, škálování návrhové hustoty, ...

Rychlost konvergence MCMC algoritmu může být ovlivněna i parametrizací statistického problému, který je analyzován. Silné korelace v aposteriorním rozdělení totiž u mnoha algoritmů (Gibbsův výběrový plán, MH s náhodnou procházkou ...) implikují i silné autokorelace v generovaném MCMC řetězci a tedy i pomalou konvergenci. Reparametrizace původního problému, pak automaticky zlepší i konvergenční vlastnosti výsledného MCMC řetězce.

Příklad: Uvažujme centrovaný lineární model, kdy pozorovaná data $Y_i \sim N(\theta_i, \sigma_e^2),\ i = 1, \dots, n$ jsou nezávislá při daných θ_i, která jsou rovněž nezávislá a $\theta_i \sim N(\mu, \sigma_\alpha^2),\ i = 1, \dots, n$ a parametr μ má nevlastní neinformativní rozdělení na $\mathbb{R}$. Použijeme-li Gibbsův výběrový plán pro simulování z aposteriorního rozdělení pro $(\{\theta_i\}_1^n, \mu)$, pak bude geometricky ergodický s řádem $\kappa = \frac{\sigma_e^2}{\sigma_e^2+\sigma_\alpha^2}$ (viz [33]). Takže bude konvergovat pomalu, pokud jsou chyby měření σ_e^2 velké v porovnání s variabilitou σ_α^2. Pokud použijeme necentrovanou parametrizaci $Y_i \sim N(\theta_i + \mu, \sigma_e^2),\ i = 1, \dots, n,\ \theta_i \sim N(0, \sigma_\alpha^2),\ i = 1, \dots, n$, pak bude řád toho samého Gibbsova algoritmu $(1 - \kappa)$.

Pro případ Gibbsova výběrového plánu (ale lze s výhodou užít i v situacích dalších MCMC algoritmů) je ještě třeba zmínit pojem rozšíření dat. To se hodí v případě, že máme chybějící data, a rozdělení pozorovaných dat by bylo nutné získat integrací věrohodnosti přes data chybějící, což ale může být v praxi špatně proveditelné. A nebo v situaci, kdy věrohodnost není výpočetně zvladatelná, ale podmíněně na něčem nepozorovaném se stane jednoduchou. Pak je dobré ony chybějící/nepozorované veličiny zahrnout do modelu jako další odhadované parametry, čímž se zjednoduší simulované hustoty a tedy i použité MCMC algoritmy.

Obecně budiž $\mathbf{y}_{pozor}$ pozorovaná data, $\mathbf{y}_{chyb}$ chybějící data, θ neznámý (obecně vektorový) parametr s apriorní hustotou $p(\theta)$, a budiž hustota $\pi(\mathbf{y}_{pozor}|\mathbf{y}_{chyb}, \theta)$ dostupná, ale hustota $\pi(\mathbf{y}_{pozor}|\theta)$ nedostupná. Pokud bereme $\mathbf{y}_{chyb}$ jako další parametr, pak

$$\pi(\theta, \mathbf{y}_{chyb}|\mathbf{y}_{pozor}) \propto \pi(\mathbf{y}_{pozor}|\mathbf{y}_{chyb}, \theta)\pi(\mathbf{y}_{chyb}|\theta)p(\theta),$$

a Gibbsův výběrový plán umožňuje generovat přibližný vzorek ze sdružené hustoty. Vzorek z $\pi(\theta|\mathbf{y}_{pozor})$ potom dostaneme prostě použitím jen θ_n z vygenerovaného MCMC řetězce — to odpovídá vyintegrování

$$\pi(\theta|\mathbf{y}_{pozor}) = \int \pi(\theta, \mathbf{y}_{chyb}|\mathbf{y}_{pozor})\mu(\mathrm{d}\mathbf{y}_{chyb}).$$

Příklad: Buď $\{Y_i\}_{i=0}^n$ pozorovaná časová řada, kde $Y_0 = 0$ a podmíněné rozdělení Y_i za podmínky předchozích $Y_0, \dots Y_{i-1}$ je $Y_{i-1} + Z_i$, kde $\{Z_i\}_{i=1}^n$ jsou vzájemně nezávislé a nezávislé na $Y_0, \dots Y_{i-1}$ s $Beta(\theta, \theta)$ rozdělením. Apriorní rozdělení parametru θ budiž rovnoměrné na $(0, 1)$. Pozoruji-li kompletní časovou řadu $\{Y_i\}_{i=0}^n$, je aposteriorní hustota pro θ jednoduchá (součin $Beta$-přírůstků). Ale pokud např. pozorování Y_{i^*} chybí?
Potom je dobré zahrnout Y_{i^*} do neznámých (odhadovaných) parametrů, neboť úplné podmíněné rozdělení Y_{i^*} má při daném θ hustotu $\propto ((y_{i^*} - y_{i^*-1})(y_{i^*-1} + 1 - y_{i^*})(y_{i^*} - y_{i^*+1} - 1)(y_{i^*+1} - y_{i^*}))^{\theta-1}$, na množině $(y_{i^*-1}, y_{i^*-1} + 1) \cap (y_{i^*+1} - 1, y_{i^*+1})$, a Gibbsův výběrový plán je jednoduchý.

Příklad: Buďte $\{Y_i\}_{i=1}^n$ iid data z hustoty $\frac{1}{2\sqrt{2\pi}}\left(\mathrm{e}^{-x^2/2} + \mathrm{e}^{-(x-\theta)^2/2}\right)$, tj. směsové hustoty kde data pochází s pravděpodobností $\frac{1}{2}$ ze standardního normálního rozdělení a s pravděpodobností $\frac{1}{2}$ z $N(\theta, 1)$. Apriorní rozdělení pro θ budiž standardní normální. Věrohodnost

$$L(\theta) \propto \prod_{i=1}^n \left(\mathrm{e}^{-y_i^2/2} + \mathrm{e}^{-(y_i-\theta)^2/2}\right)$$

se špatně maximalizuje, stejně tak jako aposteriorní hustota pro θ – zlogaritmovaná hustota má $2n$ členů, což je pro velká n výpočetně nezvladatelné.
Zaveďme ale $\mathbf{Y}_{chyb}$ posloupnost alternativních veličin, které budou indikovat, jestli bylo Y_i generováno

z $N(\theta, 1)$ nebo ne, a jejich apriorní rozdělení budiž neinformativní (rovnoměrné na $\{0,1\}^n$). Potom je podmíněné aposterirní rozdělení pro θ rovno

$$N\left(\frac{\sum_{i\in A} Y_i}{n+1}, \frac{1}{n+1}\right), \qquad \text{kde} \quad A = \{1 \le i \le n : Y_{chyb,i} = 1\},$$

a podmíněná aposteriorní pravděpodobnost pro $i \in A$ je

$$\frac{\mathrm{e}^{-(y_i-\theta)^2/2}}{\mathrm{e}^{-(y_i-\theta)^2/2} + \mathrm{e}^{-y_i^2/2}},$$

a Gibbsův výběrový plán umožňuje snadno generovat přibližný vzorek ze sdružené hustoty.

Pro určité algoritmy existují pokyny, jak volit parametry měřítka v návrhovém rozdělení (např. četnost přijatých návrhů kolem 25% pro vícerozměrné problémy) (viz např. [37] a citace v něm). Ovšem přílišné ladění návrhového rozdělení není nutné.

Tvorba vzorku

Burn-in – tento pojem by měl označovat úsek, kde se ještě projevuje startovací hodnota. Vynechání tohoto úseku znamená, že vlastně generovaný řetězec startujeme z jiné hodnoty X_0. Určuje se na základě vizuální inspekce výstupů generovaných v ladicí fázi. V konečném důsledku je to jedno z řešení problému volby vhodného počátečního bodu pro simulace. Není optimální, jsou i jiná řešení – např. nějaký bod, kam jsme se dostali během ladicích simulací, hodnota odhadů pro daná data a statistický problém s jednodušším modelem z předchozí analýzy atp. Zhruba řečeno, vhodný startovací bod je takový, který se nachází v oblasti, kde cílové rozdělení má nezanedbatelně velké množství hmoty (vzhledem k cílovému rozdělení lze startovací bod nazvat typickým, nikoli výjimečným). Ale v principu to může být jakýkoli bod, který nám nevadí mít ve výsledném vzorku. Pro podrobnější diskusi této otázky viz např. [2, Chapter 1.11].

Statistickou analýzu výstupů MCMC zakládáme na vzorku vzniklém jako výstup z jednoho dostatečně dlouhého běhu zvoleného MCMC algoritmu. Tento vzorek je korelovaný, ale jeho použití je založeno na ergodických větách pro markovské řetězce. Vyšší korelace je třeba kompenzovat větší délkou řetězce. Pokud by pro nás potřebně dlouhý řetězec byl příliš dlouhý z hlediska nároků na jeho skladování v paměti, je možné ukládat jen každou řekněme k-tou iteraci (uvědomme si, že to vlastně odpovídá použití jiného markovského jádra P^k). Skladované iterace z podřetězce pak budou méně korelované. Nicméně vydatnost odhadu se tímto postupem snižuje (vygenerovat musíme stále stejné množství hodnot, jen ne všechny ukládáme a použijeme).

Použití vzorku

Ergodická věta 5.2.12 zajišťuje konzistenci ergodického průměru. Chyba, které se dopouštíme použitím MCMC aproximace, tedy rozdíl $\bar{h}_n - \mathbb{E}_\pi h(X)$, se nazývá MCMC chyba. Abychom dostali MCMC chybu pod kontrolu potřebujeme CLV 5.2.13. Vždy, když používáme MCMC je dobré mít představu o tom, jak velké MCMC chyby se dopouštíme. S využitím centrální limitní věty pro geometricky ergodické řetězce lze sestavit i přibližné intervaly spolehlivosti. Je však třeba odhadnout neznámý limitní rozptyl.

A také je důležité zopakovat, že pokud nemáme teoretickou vlastnost MCMC algoritmu, která CLV implikuje (např. geometrickou ergodicitu), tak nemáme žádnou jistotu, že MCMC chyba konverguje k 0. Samozřejmě můžeme „odhadnout“ asymptotický rozptyl podle metod níže a nějaká čísla nám vyjdou. Nicméně pokud CLV neplatí, pak tato čísla nemají žádný smysl!

Velikost „dostatečně malé“ MCMC chyby také závisí na problému, který chceme pomocí MCMC vyřešit. V praxi se vyskytují dva typy úkolů:

a) statistická inference o aposteriorním rozdělení θ

b) odhad integrálu = neznámé konstanty $\mathbb{E}\,\theta$.

Úkol typu (a) se typicky vyskytuje v bayesovské statistice, úkol typu (b) ve statistické fyzice, ve statistice při výpočtech např. normalizačních konstant nebo v grafových aplikacích při odhadu počtu prvků nějaké (velké) množiny. Pro úkol typu (b) je potřeba mnohem větší přesnost než pro úkol typu (a).

Příklad: Mějme skalární parametr θ s aposteriorním rozdělením které je přibližně normální s (pomocí MCMC) odhadnutou střední hodnotou 3.47 a odhadnutou směrodatnou odchylkou 1.83. A budiž odhadnutá MCMC směrodatná odchylka pro odhad střední hodnoty 0.1. Potom pro úkol typu (a) můžeme naše simulace ukončit. Odhadnutá MCMC chyba je zanedbatelná vzhledem k nejistotě o θ v aposteriorním

rozdělení. Pokud nás ale zajímá přesná hodnota $\mathbb{E}\,\theta$ (plníme úkol typu (b)), musíme použít delší běh MCMC řetězce, abychom zjistili, jestli je $\mathbb{E}\,\theta$ rovno 3.53 nebo 3.53840 nebo něčemu jinému.

Odhad limitního rozptylu

Předpokládejme, že platí centrální limitní věta 5.2.13 pro ergodické průměry $\bar{h}_n = \frac{1}{n}\sum_{i=1}^{n} h(X_i)$ markovského řetězce $\{X_n\}$. Limitní rozptyl je

$$\sigma_h^2 = \sigma^2 \left(1 + 2\sum_{k=1}^{\infty} \rho_k\right),$$

kde $\sigma^2 = \mathrm{var}_\pi\, h(X)$ je rozptyl při limitním rozdělení π a $\rho_k = \mathrm{cor}(h(X_t), h(X_{t+k}))$ jsou autokorelace řádu k. Existuje několik metod odhadu limitního rozptylu.

Přímá metoda

Použijeme klasické výběrové odhady pro odhad rozptylu $\sigma^2 = \gamma_0$ a kovariancí $\gamma_k = \mathrm{cov}(h(X_t), h(X_{t+k}))$:

$$\hat{\gamma}_k = \frac{1}{n}\sum_{j=1}^{n-k} (h(X_j) - \bar{h}_n)(h(X_{j+k}) - \bar{h}_n), \qquad k \geq 0.$$

Položíme $\hat{\rho}_k = \frac{\hat{\gamma}_k}{\hat{\gamma}_0}$ pro $k \leq m$, $\hat{\rho}_k = 0$ pro $k > m$ a

$$\widehat{\sigma}_h^2 = \hat{\gamma}_0 \left(1 + 2\sum_{k=1}^{\infty} \hat{\rho}_k\right) = \hat{\gamma}_0 + 2\sum_{k=1}^{m} \hat{\gamma}_k.$$

Hodnota m by měla být volena tak, abychom zahrnuli do součtu všechny podstatné členy. Pro velké hodnoty k jsou totiž odhady $\hat{\gamma}_k$ velmi nepřesné a je proto lepší je nahradit nulou. Na druhou stranu příliš malá hodnota m může pro silně korelovaný řetězec $\{X_n\}$ znamenat velké záporné vychýlení (velké podhodnocení) odhadu σ_h^2.

Batch means

Rozdělíme simulovaný řetězec o délce $n = mk$ na k podřetězců (batch) o m po sobě jdoucích hodnotách. Uvažujeme průměry podřetězců, tedy:

$$\bar{h}^{(i)} = \frac{1}{m}\sum_{j=1}^{m} h(X_{m(i-1)+j}), \qquad i = 1, \ldots, k.$$

Snahou je volit k tak, aby průměry $\bar{h}^{(1)}, \ldots, \bar{h}^{(k)}$ byly přibližně nezávislé. Hodnota k se obvykle volí mezi 20 a 30. Aby byl předpoklad nezávislosti oprávněný, je u pomalu mixujících řetězců třeba brát dostatečně dlouhé batche. Jinak se opět objevuje záporné vychýlení.

Pro m dost velké je (z CLV pro markovké řetězce) rozptyl $\bar{h}^{(i)}$ přibližně $\frac{\sigma_h^2}{m}$ a můžeme ho odhadnout výběrovým rozptylem, tedy

$$\widehat{\sigma}_h^2 = \frac{m}{k-1}\sum_{i=1}^{k} \left(\bar{h}^{(i)} - \bar{h}_n\right)^2.$$

Metoda počáteční sekvence

Pro reverzibilní řetězce platí, že posloupnost $\{\Gamma_j\}_{k=0}^{\infty}$, definovaná jako $\Gamma_k = \gamma_{2k} + \gamma_{2k+1}$, je striktně kladná, striktně klesající a striktně konvexní (důkaz viz [9]). Toto tvrzení umožňuje zavést tři další odhady limitního rozptylu: pomocí počáteční kladné, počáteční monotónní a počáteční konvexní sekvence. Všechny jsou konzistentními overestimates limitního rozptylu – tj. pro libovolné $\epsilon > 0$ jde pravděpodobnost, že záporné vychýlení odhadu bude menší než $-\epsilon$, s rostoucí délkou řetězce k nule. Nejlepší z těchto tří odhadů (protože nejmenší) je odhad pomocí počáteční konvexní posloupnosti $\hat{\sigma}^2_{conv}$. Ovšem nejhůře se počítá (lze ale použít funkci `initseq` z `R` knihovny `mcmc`). Odhad se spočítá následovně: najde se maximální m splňující $\hat{\Gamma}_k > 0$ pro všechny $k \leq m$. Položí se $\hat{\Gamma}_{m+1} = 0$ a $k \to \tilde{\Gamma}_k$ se definuje jako největší konvexní minoranta od $k \to \hat{\Gamma}_k$, přes $0, \ldots m+1$. Potom $\hat{\sigma}^2_{conv} = -\hat{\gamma}_0 + 2\sum_{k=0}^{m} \tilde{\Gamma}_k$. Pro podrobnosti viz [9] nebo [2, Chapter 1.10]. Tuto metodu je možno kombinovat s metodou batch means v případě, že batche nejsou nezávislé.

Implementace a numerika

Implementace: Interpretované jazyky jako `R` nebo Java jsou pomalé. Pro rozsáhlejší výpočty je nutné užít nějaký vyšší programovací jazyk (např. C++), kde celý kód je kompilován najednou. Kód zkompilovaný v C lze volat v `R`, viz `help(.C)`.

Kapitola 7

Metropolisův-Hastingsův-Greenův algoritmus

Metropolisův-Hastingsův-Greenův (MHG) algoritmus je zobecnění Metropolis-Hastingsova algoritmu, které umožňuje generovat markovský řetězec s požadovaným cílovým rozdělením i v případě, kdy nemáme k dispozici hustotu cílového rozdělení a přechodovou hustotu vzhledem k referenčním mírám μ a $\mu \times \mu$ (když je návrhové rozdělení singulární vzhledem k cílovému rozdělení). To se často děje, když je stavový prostor sjednocení prostorů různé dimenze. Původní myšlenka byla publikována v [11], kde bylo také představeno míchání jader závislé na stavu, které se s MHG algoritmem často kombinuje.

7.1 Míchání závislé na stavu

Mějme konečnou nebo spočetnou množinu jader $P_i(x, A), i \in I$ na $\mathcal{X} \times \mathfrak{X}$ a míchací pravděpodobnosti závislé na stavu $p_i(x)$. Definujme celkové jádro

$$P(x, A) = \sum_{i \in I} p_i(x) P_i(x, A) \tag{7.1}$$

a substochastická jádra $K_i(x, A) = p_i(x)P_i(x, A)$. Pak máme jednoznačné přiřazení mezi K_i a dvojicí (p_i, P_i), neboť $p_i(x) = K_i(x, \mathcal{X})$ a $P_i(x, A) = K_i(x, A)/K_i(x, \mathcal{X})$.

Pokud je každé K_i reverzibilní vzhledem k danému rozdělení π, pak je i součet (7.1) reverzibilní vzhledem k π, a pokud je P stochastické jádro, plyne z tvrzení 5.2.5, že π je stacionární rozdělení pro P. Ale i pokud se K_i nesečtou na kompletní stochastické jádro, tj. $\sum_{i \in I} K_i(x, \mathcal{X}) < 1$ pro nějaká x, můžeme součet doplnit na stochastické jádro následujícím způsobem: definujeme defekt

$$d(x) = 1 - \sum_{i \in I} K_i(x, \mathcal{X}), \qquad x \in \mathcal{X},$$

a nové jádro

$$\tilde{K}(x, A) = d(x)\delta_x(A), \qquad x \in \mathcal{X}, A \in \mathfrak{X}. \tag{7.2}$$

Lemma 7.1.1. *Jádro $\tilde{K}$ je reverzibilní vzhledem k libovonému rozdělení π.*

Důkaz: Pro libovolné $A, B \in \mathfrak{X}$ platí

$$\int_A \tilde{K}(x, B)\pi(\mathrm{d}x) = \int_{\mathcal{X}} d(x) I(x \in A, x \in B)\pi(\mathrm{d}x),$$

takže po prohození pořadí A a B se hodnota výrazu nezmění a podmínka definice 5.2.5 je splněna. □

Algoritmus 7.1.2. *MCMC s mícháním závislým na stavu:*

0. Zvol $x^{(0)}$ a polož $t = 0$.

1. *Vyber index i s pravděpodobností $p_i(x^{(t)})$, s pravděpodobností $1 - \sum_{i\in I} p_i(x^{(t)})$ přeskoč krok 2. a polož $x^{(t+1)} = x^{(t)}$.*

2. *Simuluj $x^{(t+1)}$ z rozdělení $P_i(x^{(t)}, .)$.*

3. *Polož $t = t + 1$ a jdi na 1.*

Tento algoritmus má přechodové jádro rovné $P = \tilde{K} + \sum_{i\in I} K_i$ a pokud je každé K_i reverzibilní vzhledem k π, je i P reverzibilní vzhledem k π a π je tedy jeho stacionární rozdělení.

7.2 Metropolisův-Hastingsův-Greenův algoritmus

MHG algoritmus je zobecnění MH algoritmu s mírami místo hustot. Uvažujeme tedy nenormovanou cílovou míru π na $\mathfrak{X}$ a návrhové jádro $Q(x, .)$. Navíc potřebujeme symetrickou míru ξ na $\mathfrak{X} \times \mathfrak{X}$, aby nahradila součinovou míru $\mu \times \mu$ z obyčejného MH algoritmu. ξ musí dominovat $\pi(\mathrm{d}x)Q(x, \mathrm{d}y)$, takže existuje Radon-Nikodýmova derivace

$$f(x, y) = \frac{\pi(\mathrm{d}x)Q(x, \mathrm{d}y)}{\xi(\mathrm{d}x, \mathrm{d}y)},$$

která nahrazuje $f(x)q(x, y)$ z obyčejného MH algoritmu. Definujeme tzv. Greenův poměr $R = \frac{f(y,x)}{f(x,y)}$.

Algoritmus 7.2.1. *jeden krok MHG algoritmu:*

1. *simuluj $y \sim Q(x, .)$,*

2. *spočti Greenův poměr R,*

3. *přijmi y s pravděpodobností $\min(1, R)$.*

Typicky se používá v kombinaci s mícháním závislým na stavu. Mějme tedy množinu (konečnou nebo spočetnou) návrhových jader $Q_i(x, .), i \in I$, která mohou být substochastická. Požadavky, která by měla jádra splňovat:

- $Q_i(x, .)$ je známo pro všechna i,
- $\sum_{i\in I} Q_i(x, \mathcal{X}) \leq 1$ platí pro všechna $x \in \mathcal{X}$,
- pro všechna i je
$$f_i(x, y) = \frac{\pi(\mathrm{d}x)Q_i(x, \mathrm{d}y)}{\xi_i(\mathrm{d}x, \mathrm{d}y)} \tag{7.3}$$
známé a vyčíslitelné pro všechna $x, y \in \mathcal{X}$ (pro různá i je možné použít různé ξ_i),
- pro každé $x \in \mathcal{X}$ a $i \in I$ umíme simulovat z návrhového rozdělení s přechodovým jádrem
$$P_i(x, .) = \frac{Q_i(x, .)}{Q_i(x, \mathcal{X})}.$$

Algoritmus 7.2.2. *Metropolisův-Hastingsův-Greenův algoritmus:*

0. *Zvol $x^{(0)}$ a polož $t = 0$.*

1. *Vyber jádro Q_i s pravděpodobností $p_i(x^{(t)}) = Q_i(x^{(t)}, \mathcal{X})$, s pravděpodobností $1-\sum_{i\in I} p_i(x^{(t)})$ polož $x^{(t+1)} = x^{(t)}$ a jdi na 5.*

2. *Generuj $y \sim P_i(x^{(t)}, .)$.*

3. *Vyčísli Greenův poměr $R = \frac{f_i(y,x^{(t)})}{f_i(x^{(t)},y)}$, kde f_i je dáno v (7.3).*

4. *Přijmi y s pravděpodobností $\min(1, R)$ a polož $x^{(t+1)} = y$.*

5. *Polož $t = t + 1$ a jdi na 1.*

Tvrzení 7.2.3. *Markovský řetězec generovaný algoritmem 7.2.2 má stacionární rozdělení π.*

Důkaz: Stačí nám ukázat reverzibilitu příslušného jádra vzhledem k π. Označme

$$\alpha_i(x,y) = \min\left(1, \frac{f_i(y,x)}{f_i(x,y)}\right)$$

a

$$r_i(x) = \int_{\mathcal{X}} (1 - \alpha_i(x,y))\; Q_i(x, \mathrm{d}y),$$

pravděpodobnost, že algoritmus zůstane v x když se používá jádro Q_i. Potom je jádro MHG algoritmu rovno $(\tilde{K} + \sum_{i\in I} K_i)$, kde

$$K_i(x,A) = r_i(x)\,\delta_x(A) + \int_A \alpha_i(x,y)\, Q_i(x,\mathrm{d}y), \qquad x \in \mathcal{X}, A \in \mathfrak{X}$$

a jádro $\tilde{K}$ je dáno rovnicí (7.2). Potom

$$\begin{aligned}\int_A \int_B K_i(x,\mathrm{d}y)\pi(\mathrm{d}x) &= \int_{A\cap B} r_i(x)\pi(\mathrm{d}x) + \int_A \int_B \alpha_i(x,y) Q_i(x,\mathrm{d}y)\pi(\mathrm{d}x) \\ &= \int_{A\cap B} r_i(x)\pi(\mathrm{d}x) + \int_A \int_B f_i(x,y)\alpha_i(x,y)\xi_i(\mathrm{d}y,\mathrm{d}x)\end{aligned}$$

První člen je triviálně symetrický pro A, B a druhý člen je symetrický, protože ξ_i je symetrická míra a MHG algoritmus je navržený tak, aby platilo

$$f_i(x,y)\alpha_i(x,y) = f_i(y,x)\alpha_i(y,x) \quad \text{pro všechna } x,y \in \mathcal{X}.$$

Podmínka (5.2.5) je tedy splněna a reverzibilita platí. □

Metropolisův-Hastingsův-Greenův algoritmus lze použít např. pro bayesovský výběr modelu ([11], [29, Chapter 7], [2, Chapter 1.17]) nebo pro simulování bodových procesů na omezeném okně. V těchto případech je totiž stavový prostor roven sjednocení prostorů různé dimenze a pro přechody mezi nimi je potřeba použít konstrukci se symetrickými referenčními mírami ξ_i.

Příklad: Simulace bodového procesu na omezeném okně
Mějme dán bodový proces X na omezené množině $E \subset \mathbb{R}^d$, s hustotou $p(x)$ vzhledem ke standardnímu Poissonovu procesu (viz definice z kapitoly 8.1). Označme $n = n(x) = x(E)$ počet bodů konfigurace x v množině E. Pravděpodobnostní rozdělení procesu X lze pak popsat následovně:

$$\mathbb{P}(X \in F) = \frac{1}{c}\int_F p(x)\Pi(\mathrm{d}x) = \frac{1}{c}\sum_{n=0}^{\infty} \frac{\mathrm{e}^{-\lambda(E)}}{n!} \int_{F\cap E^n} p(x)\lambda^n(\mathrm{d}x),$$

kde λ je Lebesgueova míra na $E \subset \mathbb{R}^d$ a λ^n je Lebesgueova míra na E^n.
Odvodíme MHG algoritmus pro generování z X. Použijeme míchání závislé na stavu a $i = n(x)$ počet bodů konfigurace x. Jádro Q_i použijeme jenom pro ty konfigurace x, kde $n(x) = i$ (navrhneme přidat jeden bod rovnoměrně náhodně na E, ostatní body x se nemění) nebo $n(x) = i+1$ (navrhneme jeden bod umazat – a pro určitost budeme mazat poslední bod konfigurace, ostatní body x se nemění). Míchací pravděpodobnosti budou

$$p_i(x) = \begin{cases} \frac{1}{2} & \text{pro } n(x) = i, \\ \frac{1}{2} & \text{pro } n(x) = i+1, \\ 0 & \text{jinak.} \end{cases}$$

Pro pevné x je $\sum_i p_i(x) = 1$, jen pro $n(x) = 0$ ne. V tom případě s pravděpodobností $(1 - \sum_i p_i(x)) = \frac{1}{2}$ nebude algoritmus dělat nic, tj. zůstane v současném stavu.
Uvažujme nyní přidání bodu do x. Sdružené rozdělení současného stavu a návrhu (x,y) je koncentrováno na množině

$$D_i^+ = \{(x,y) \in E^i \times E^{i+1} : x_j = y_j, j \le i\},$$

což ale není symetrická množina. Výměna x a y dá

$$D_i^- = \{(x,y) \in E^{i+1} \times E^i : x_j = y_j, j \le i\}.$$

D_i^+ je izomorfní s E^{i+1} přes zobrazení $(x, y) \to y$ a D_i^- je izomorfní s E^{i+1} přes zobrazení $(x, y) \to x$. Buď ξ_i symetrická míra na $\mathcal{X} \times \mathcal{X}$ koncentrovaná na $D_i^+ \cup D_i^-$ a odpovídající λ^{i+1} na E^{i+1}. Můžeme nyní spočítat Radon-Nikodýmovy derivace (7.3). Pokud navrhujeme přidat bod, pak je

$$f_i(x, y) = \frac{p(x)\,\mathrm{e}^{-\lambda(E)}}{c\,i!\lambda(E)} \propto \frac{p(x)}{i!\lambda(E)}, \tag{7.4}$$

neboť rovnoměrně náhodně přidaný bod má hustotu $\frac{\lambda(.)}{\lambda(E)}$. Pokud navrhujeme odebrat (poslední) bod, pak je

$$f_i(x, y) = \frac{p(x)\,\mathrm{e}^{-\lambda(E)}}{c\,(i+1)!} \propto \frac{p(x)}{(i+1)!}, \tag{7.5}$$

protože odebrání nenáhodného bodu má hustotu 1. Greenův poměr pro přidání bodu bude potom roven podílu (7.5) s prohozenými x a y a (7.4)

$$R = \frac{\lambda(E)\,p(y)}{(i+1)\,p(x)}. \tag{7.6}$$

Pro umazání bodu bude Greenův poměr roven převrácené hodnotě (7.6).

Pozn.: i-tice bodů se u bodového procesu nahlížejí jako neuspořádané, $p(x)$ je symetrická vzhledem k permutaci bodů. Pokud ovšem složím výše uvedený mazací krok s krokem, který jen permutuje body v konfiguraci x, tak se Greenův poměr nezmění, ale vymažu rovnoměrně náhodně vybraný bod. Upravený algoritmus bude potom rychleji mixovat.

Pozn.: Uvědomme si, že kdybychom v našem příkladě použili obecné pravděpodobnosti $Q(x)$, resp. $1 - Q(y)$ pro přidání, resp. vymazání bodu, obecnou hustotu $b(x, .)$ (normovanou) pro rozdělení nově navrhovaného bodu a obecné pravděpodobnosti $d(y, .)$ (normované) pro výběr vymazávaného bodu, dostaneme místo (7.6)

$$R = \frac{f_i(y, x)}{f_i(x, y)} = \frac{1 - Q(y)}{Q(x)} \frac{p(y)}{p(x)} \frac{d(y, v)}{b(x, v)},$$

kde $y = x \cup v$. Takže dostáváme algoritmus totožný s Metropolis-Hastingsovým algoritmem rození a zániku 8.2.1.

Kapitola 8

Bodové procesy

Kromě bayesovské statistiky (kapitola 3) se metody MCMC na obecných stavových prostorech často využívají v prostorové statistice. Detailněji jsou této problematice věnovány přednášky *Prostorové modelování* a *Prostorová statistika*.

8.1 Bodové procesy

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 8.1.1. *Bodový proces (point process) na E je náhodný element v měřitelném prostoru $(\mathcal{N}, \mathfrak{N})$. Nechť Λ je difúzní (tedy $\Lambda(\{u\}) = 0$ pro $u \in E$) 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 ($F \in \mathfrak{N}$):

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

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 8.1.2. *Mějme reálné parametry $\beta > 0$, $0 \le \gamma \le 1$ a $R > 0$. Straussův proces (Strauss process) je bodový proces X s hustotou $p(x) = \alpha\beta^{x(E)}\gamma^{S(x)}$, kde $S(x) = \sum_{i\neq j} I_{[\varrho(x_i,x_j)<R]}$.*

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 [36], 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 $\gamma \ge 0$. Tedy pro $\gamma > 1$ můžeme dostat model pro shlukování bodů, který ovšem není moc vhodný, v praxi se používají lepší modely.

8.2 Metropolisův-Hastingsův algoritmus rození a zániku

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

Algoritmus 8.2.1. *Metropolisův-Hastingsův algoritmus rození 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 u s hustotou $b(x, u)$ vzhledem k Λ, s pravděpodobností $1 - Q(x)$ navrhni ubrání bodu v s pravděpodobností $d(x, v), v \in x$,*
2. *návrh přijmi (buď $X_{t+1} = x\cup u$, nebo $X_{t+1} = x\setminus v$) s pravděpodobností $\alpha(x, x\cup u) = \min(1, h(x, u))$, $\alpha(x \cup u, x) = \min\left(1, \frac{1}{h(x,u)}\right)$, kde*

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

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

Dále volíme speciálně $Q(\cdot) = \frac{1}{2}$, $b(\cdot, \cdot) = \frac{1}{\Lambda(E)}$, $d(x \cup u, \cdot) = \frac{1}{x(E)+1}$, tedy

$$h(x, u) = \lambda(x, u)\frac{\Lambda(E)}{x(E) + 1}, \tag{8.1}$$

kde $\lambda(x, u) = \frac{p(x\cup u)}{p(x)}$ je *podmíněná intenzita (conditional intensity).*

Uvědomme si, že algoritmus 8.2.1 neodpovídá definici 4.2.1 obyčejného MH algoritmu. Není to jeho speciální případ – nemáme totiž k dispozici přechodovou hustotu q pro výpočet Hastingsova poměru. Reverzibilita tohoto algoritmu vzhledem k cílovému rozdělení se proto musí dokázat - viz např. důkaz Theorem 7.2 v [26]. A nebo je možné odvodit tento algoritmus jako speciální případ Metropolisova-Hastingsova-Greenova algoritmu 7.2.2 – viz příklad z předchozí kapitoly.

Nyní definujeme podmínku stability, která je postačující pro geometrickou ergodicitu algoritmu 8.2.1.

Definice 8.2.1. *Řekneme, že konečný bodový proces X s hustotou p je lokálně stabilní (locally stable), když $\lambda(x, u) \le K$ pro nějakou konstantu K, neboli*

$$p(x \cup u) \le Kp(x), \quad \text{pro všechna } x \in \mathcal{N}, u \in E. \tag{8.2}$$

Pozn.: Z podmínky (8.2) plyne, že pro $p(x) = 0$ je i $p(x \cup u) = 0$, podmíněná intenzita $\lambda(x, u)$ 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 Metropolisovým-Hastingsovým algoritmem rození 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 8.2.2. *Pokud p splňuje podmínku (8.2) lokální stability, potom je markovský řetězec $\{X_t\}$ φ-nerozložitelný na $\mathcal{N}^+$ a pro každé $k \in \mathbb{N}_0$ je $C = \{x \in \mathcal{N}^+ : x(E) \leq k\}$ malá množina.*

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

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

za předpokladu, že K z definice 8.2.1 je zvoleno dostatečně velké, aby $\frac{1}{K\Lambda(E)} < 1$. Tedy pro $x \in C$ a $m > k$ je $P^m(x, \{\emptyset\}) \geq c^m$. Zvolme míru $\nu = \delta_\emptyset$, potom $P^m(x, A) \geq c^m \nu(A)$ pro každé $x \in C$ a $A \in \mathfrak{N}$. Pro $n = x(E) = 0$ je $P^m(\emptyset, A) \geq (1 - Q(\emptyset))^m \delta_\emptyset(A) = (1/2)^m \nu(A) \geq c^m \nu(A)$. Celkem tak dostáváme, že C je malá ($\varepsilon = c^m$). Podobně položíme-li $\varphi = \delta_\emptyset$, tak $P^m(x, A) \geq c^m > 0$ kdykoli $m \geq x(E)$ a $\varphi(A) > 0$. Řetězec je proto φ-nerozložitelný.

□

Věta 8.2.3. *Markovský řetězec pro simulaci bodového procesu s lokálně stabilní hustotou $p(x)$ MH-algoritmem rození a zániku 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 5.2.17 a tvrzení 8.2.2. Naopak nechť řetězec je stejnoměrně ergodický, tedy $\mathcal{N}$ je malá (věta 5.2.17). Existuje proto míra ν a $m \in \mathbb{N}$ tak, že $P^m(x, F) \geq \nu(F)$ 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 $P^m(x, \mathcal{N}_k) \geq \nu(\mathcal{N}_k) > 0$, což je spor, neboť $P^m(x, \mathcal{N}_k) = 0$.

□

Věta 8.2.4. *Je-li p lokálně stabilní hustota, je příslušný Metropolisův-Hastingsův algoritmus rození a zániku aperiodický a geometricky ergodický.*

Důkaz: Zřejmě $P(\emptyset, \{\emptyset\}) \geq 1 - Q(\emptyset) = \frac{1}{2} > 0$ a odtud plyne aperiodicita. K ověření geometrické ergodicity použijeme větu 5.2.16 s $V(x) = c^n$, kde $n = x(E)$ a $c > 1$ je konstanta. Z předpokladu věty je $\lambda(x, u) \leq K$. Pro $n \geq 1$ 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:

$$\begin{aligned} P(x, \mathcal{N}_{n+1}) &= Q(x) \int_E I_{[x \cup u \in \mathcal{N}_{n+1}]} b(x, u)\alpha(x, x \cup u)\, \Lambda(\mathrm{d}u) \\ &= \frac{1}{2} \int_E I_{[x \cup u \in \mathcal{N}_{n+1}]} \min \left\{ 1, \lambda(x, u) \frac{\Lambda(E)}{n+1} \right\} \frac{\Lambda(\mathrm{d}u)}{\Lambda(E)} \leq \frac{1}{2} \min \left\{ 1, \frac{K\Lambda(E)}{n+1} \right\} \leq \frac{\varepsilon}{2}, \end{aligned}$$

pro $n + 1 \geq \frac{K\Lambda(E)}{\varepsilon}$,

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

při $n \geq K\Lambda(E)$,

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

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

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

Protože $\frac{1}{2c} + \frac{1}{2} < 1$, existuje $\beta < 1$ tak, že pro ε dost malé je $\int V(y)\, P(x, \mathrm{d}y) \leq \beta V(x)$ pro $x \notin C$, kde $C = \{x \in \mathcal{N} : x(E) < N_{K,\varepsilon}\}$ je malá množina (tvrzení 8.2.2). Dále pro $x \in C$ je $\int V(y)\, P(x, \mathrm{d}y) \leq c^{N_{k,\varepsilon}+1} = b$. Je proto splněna podmínka geometrického driftu.

□

Kapitola 9

Další algoritmy založené na MCMC metodách

9.1 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 markovský 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 9.1.1. *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 9.1.1. *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: Buď $M = \{s \in S : h(s) = \min_{i \in S} h(i)\}$ a označme $\min_{s \in S} = a = a_i = h(i)$, $i \in M$,

$b = \min_{s\in S\setminus M} h(s)$. Potom

$$\sum_{i\in M}\pi_{h,T}(i) = \sum_i \frac{1}{Z_{h,T}}\mathrm{e}^{-\frac{a_i}{T}} = \frac{\sum_i \mathrm{e}^{-\frac{a_i}{T}}}{\sum_{s'\in S}\mathrm{e}^{-\frac{h(s')}{T}}} = \frac{\sum_i \mathrm{e}^{-\frac{a_i}{T}}}{\sum_i \mathrm{e}^{-\frac{a_i}{T}} + \sum_{s'\in S\setminus M}\mathrm{e}^{-\frac{h(s')}{T}}}$$
$$\geq \frac{\sum_i \mathrm{e}^{-\frac{a_i}{T}}}{\sum_i \mathrm{e}^{-\frac{a_i}{T}} + |S\setminus M|\,\mathrm{e}^{-\frac{b}{T}}} = \frac{|M|}{|M| + |S\setminus M|\,\mathrm{e}^{\frac{a-b}{T}}},$$

a tedy $\lim_{T\to 0+}\alpha(T) = 1$.

□

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

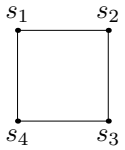
$$T^{(n)} \geq \frac{k\,(\max_{s\in S} h(s) - \min_{s\in S} h(s))}{\log n}$$

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: [8]

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 případu. 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

$$P_T = \begin{pmatrix} 1-e^{-1/T} & \frac{1}{2}\mathrm{e}^{-1/T} & 0 & \frac{1}{2}\mathrm{e}^{-1/T} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & \frac{12}{\mathrm{e}}^{-2/T} & 1-\mathrm{e}^{-2/T} & \frac{1}{2}\mathrm{e}^{-2/T} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \end{pmatrix}.$$

Nechť nehomogenní markovský 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

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

Pro $0 \leq u_i < 1$ platí $\prod_{i=1}^{\infty}(1-u_i) > 0 \Leftrightarrow \sum_{i=1}^{\infty} u_i < \infty$ (viz [35], 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).

9.2 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 [28]. 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 ([16], 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 9.2.1. *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.*

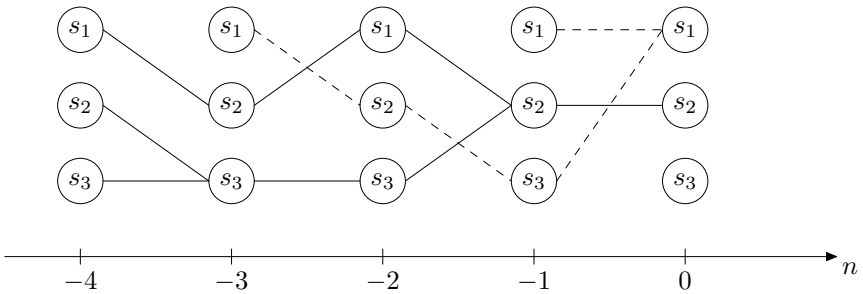
Pokud všechny řetězce skončí ve stejném stavu, říkáme, že došlo ke *koalescenci (coalescence)*. V druhém kroku se vždycky užívá stejná posloupnost U_i, což vyžaduje jisté nároky na paměť počítače.

Příklad: (na koalescenci) Buď $S = \{s_1, s_2, s_3\}$.

$N_1 = 1$, 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 9.2.2. *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

$$\begin{aligned}|\mathbb{P}(Y = s_i) - \pi_i| = |\mathbb{P}(Y = s_i) - \mathbb{P}(\tilde{Y} = s_i)| &\leq \max\Big((\mathbb{P}(Y = s_i) - \mathbb{P}(\tilde{Y} = s_i)), (\mathbb{P}(\tilde{Y} = s_i) - \mathbb{P}(Y = s_i))\Big) \\ &\leq \max\Big(\mathbb{P}(Y = s_i, \tilde{Y} \neq s_i), \mathbb{P}(\tilde{Y} = s_i, Y \neq s_i)\Big) \leq \mathbb{P}(Y \neq \tilde{Y}) \leq \varepsilon.\end{aligned}$$

□

Příklad: Ukážeme, že metoda nefunguje, pokud provádíme coupling dopředu a chceme použít vzorek X_N, kde N je čas koalescence. Buď

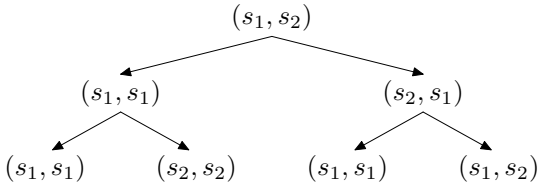
$$P = \begin{pmatrix} 1/2 & 1/2 \\ 1 & 0 \end{pmatrix}.$$

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.

Obdobně lze nahlédnout, že je důležité nezaměňovat výstup algoritmu X_0 s $X_{\text{čas koalescence}}$.

Příklad: CFTP algoritmus nefunguje, pokud nepoužíváme stále stejná U_t: Uvažujme řetězec z předchozího příkladu a označme $M = \max\{m : \text{algoritmus se rozhodne simulovat řetězec startující v čase } -N_m\}$. Nechť se generují vždy nová U_t a Y je výstup algoritmu. Potom

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



Pro daný markovský řetězec může existovat několik různých přechodových funkcí. Pro obyčejnou MCMC simulaci je výběr přechodové funkce nepodstatný, ale pro perfektní simulaci je často velmi závažný. S různými přechodovými funkcemi může docházet ke koalescenci různě rychle nebo třeba vůbec ne.

Příklad: Uvažujme markovský řetězec se stavovým prostorem $S = \{s_1, s_2\}$ a maticí přechodu

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

Dvě možné volby přechodové funkce jsou

$$\Phi_1(s_1, x) = \begin{cases} s_1 & \text{pro } x \in [0, 1/2), \\ s_2 & \text{pro } x \in [1/2, 1], \end{cases} \qquad \Phi_1(s_2, x) = \begin{cases} s_2 & \text{pro } x \in [0, 1/2), \\ s_1 & \text{pro } x \in [1/2, 1], \end{cases}$$

$$\Phi_2(s_i, x) = \begin{cases} s_1 & \text{pro } x \in [0, 1/2), \\ s_2 & \text{pro } x \in [1/2, 1], \end{cases} \qquad \text{pro } i = 1, 2.$$

Pro algoritmus perfektní simulace volíme $N_k = 2^{k-1}, k \in \mathbb{N}$.

Při použití přechodové funkce Φ_1 algoritmus nikdy neskončí, zatímco při použití přechodové funkce Φ_2 algoritmus skončí (s pravděpodobností 1) hned v prvním běhu řetězce z času $-N_1 = -1$ do času 0.

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

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

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

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. Pro další čtení doporučujeme např. [17], [18] a [2, Chapter 8].

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] S. Brooks, A. Gelman, G. L. Jones and X. Meng (2011): *Handbook of Markov Chain Monte Carlo*, Chapman & Hall/CRC, Boca Raton.

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

[4] P. Diaconis (2009): The Markov Chain Monte Carlo revolution, *Bulletin of AMS* **46**, 179–205.

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

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

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

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

[9] C. Geyer (1992): Practical Monte Carlo Markov chain (with discussion), *Statistical Science* **7**, 473–511.

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

[11] P. Green (1995): Reversible jump MCMC computation and Bayesian model determination, *Biometrika*, **82**, 711-732.

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

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

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

[15] G. L. Jones and J. P. Hobert (2001): Honest exploration of intractable probability distributions via Markov chain Monte Carlo, *Statistical Science* **16**, 312–334.

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

[17] W. S. Kendall (2015): Introduction to CFTP using R, v *Stochastic Geometry, Spatial Statistics, and Random Fields: Models and Algorithms*, editor V. Schmidt, Springer Lecture Notes in Mathematics /**2120**, 405-439.

[18] W. S. Kendall (2005): Notes on perfect simulation, v *Markov Chain Monte Carlo: Innovations and Applications*, editoři W. S. Kendall, F. Liang a J. S. Wang, World Scientific, Singapore, 93–146.

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

[20] D. A. Levin, Y. Peres and E. L. Wilmer (2009): *Markov Chains and Mixing Times*, Providence, RI:Amer. Math. Soc.

[21] 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.

[22] K. L. Mengersen and R. L. Tweedie (1996): Rates of convergence of the Hastings and Metropolis algorithms, *Ann. Statist.* **24**, 101–121.

[23] 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.

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

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

[26] J. Møller and R. P. Waagepetersen (2003): *Statistical Inference and Simulation for Spatial Point Processes*, Chapman & Hall/CRC, Boca Raton.

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

[28] 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.

[29] C. P. Robert (2007): *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*, Second Edition, Springer, New York.

[30] G. O. Roberts (1999): A Note on Acceptance Rate Criteria for CLTs for Metropolis-Hastings Algorithms, *Journal of Applied Probebility* **36**, 1210–1217.

[31] G. O. Roberts and J. S. Rosenthal (1998): Markov chain Monte Carlo: some practical implications of theoretical results, *The Canadian Journal of Statistics* **26**, 5–20.

[32] G. O. Roberts and J. S. Rosenthal (2004): General state space Markov chains and MCMC algorithms, *Probability Surveys* **1**, 20–71.

[33] G. O. Roberts and S. K. Sahu (1997): Updating schemes, correlation structure, blocking and parametrization for the Gibbs sampler, *JRSS B* **59**, 291–317.

[34] J. S. Rosenthal (2002): Quantitative convergence rates of markov chains: a simple account, *Elect. Comm. in Probab.* **7**, 123–128.

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

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

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