Stochastik-Praktikum

Markov Chain Monte Carlo

Peter Frentrup

Humboldt-Universität zu Berlin

16. Januar 2018

Übersicht

Problemstellung

- Zur Untersuchung einer Verteilung $\mathbb{P}_f$ mit Dichte f ist es i.A. nötig, (viele) Stichproben $X \sim \mathbb{P}_f$ zu generieren, um so Kenngrößen von $\mathbb{P}_f$, wie Erwarrtungswert $\mathbb{E}_f[X]$, Varianz $\mathbb{V}\mathrm{ar}_f(X)$, etc. nach Monte-Carlo-Methode zu approximieren: $\mathbb{E}_f[h(X)] \sim \frac{1}{N}\sum_{i=1}^N h(X_i)$.
- Verschiedene Methoden für konkrete f:
  - Inversionsmethode
  - spezielle Methoden (Normalverteilung, diskrete Verteilungen)
  - **Verwerfungsmethode**

Verwerfungsmethode

- **Ziel:** Erzeuge $X \sim \mathbb{P}_f$ mit Wahrscheinlichkeitsdichte f.
- Gegeben: Kandidatendichte g, sodass $Y \sim \mathbb{P}_g$ *leicht* zu simulieren ist, mit $f(x)/g(x) \leq M \; \forall x$ für eine Konstante M.
- Algorithmus:
  - Schritt 1: Erzeuge $U \sim \mathcal{U}_{[0,1]}$ und $Y \sim \mathbb{P}_g$, sodass $U \perp\!\!\!\perp Y$ (unabh.)
  - Schritt 2:
    - Falls $U \leq f(Y)/(Mg(Y))$, so akzeptiere: $X := Y$;
    - sonst: lehne Y ab, kehre zu Schritt 1 zurück.

Verwerfungsmethode – Eigenschaften

- Korrektheit:

$$\mathbb{P}[X \le x] = \mathbb{P}\left[Y \le x \,\middle|\, U \le \frac{f(Y)}{Mg(Y)}\right] = \frac{\mathbb{P}\left[Y \le x, U \le \frac{f(Y)}{Mg(Y)}\right]}{\mathbb{P}[U \le \frac{f(Y)}{Mg(Y)}]}$$

$$= \frac{\int_{-\infty}^{x} \int_0^{f(y)/(Mg(y))} \mathrm{d}u\, g(y)\,\mathrm{d}y}{\int_{-\infty}^{\infty} \int_0^{f(y)/(Mg(y))} \mathrm{d}u\, g(y)\,\mathrm{d}y} = \frac{\frac{1}{M}\int_{-\infty}^{x} f(y)\,\mathrm{d}y}{\frac{1}{M}\int_{-\infty}^{\infty} f(y)\,\mathrm{d}y}$$

$$= \mathbb{P}_f\left[(-\infty, x]\right].$$

- Unabhängig von der Dimension.
- Akzeptanzwahrscheinlichkeit: $\mathbb{P}\left[U \le \frac{f(Y)}{Mg(Y)}\right] = 1/M$.
  Je kleiner M, desto eher wirkt akzeptiert.
- Problem: g schwierig zu finden.

Verwerfungsmethode – Eigenschaften

- Korrektheit:

$$\mathbb{P}[X \le x] = \mathbb{P}\left[Y \le x \,\middle|\, U \le \frac{f(Y)}{Mg(Y)}\right] = \frac{\mathbb{P}\left[Y \le x, U \le \frac{f(Y)}{Mg(Y)}\right]}{\mathbb{P}[U \le \frac{f(Y)}{Mg(Y)}]}$$
$$= \frac{\int_{-\infty}^{x} \int_0^{f(y)/(Mg(y))} \mathrm{d}u \, g(y) \, \mathrm{d}y}{\int_{-\infty}^{\infty} \int_0^{f(y)/(Mg(y))} \mathrm{d}u \, g(y) \, \mathrm{d}y} = \frac{\frac{1}{M} \int_{-\infty}^{x} f(y) \, \mathrm{d}y}{\frac{1}{M} \int_{-\infty}^{\infty} f(y) \, \mathrm{d}y}$$
$$= \mathbb{P}_f\left[(-\infty, x]\right].$$

- Unabhängig von der Dimension.
- Akzeptanzwahrscheinlichkeit: $\mathbb{P}\left[U \le \frac{f(Y)}{Mg(Y)}\right] = 1/M$.
  Je kleiner M, desto eher wirkt akzeptiert.
- Problem: g schwierig zu finden.

Verwerfungsmethode – Beispiel

- Erzeuge $X \sim \text{Beta}(4,3)$, d.h.
  $f(x) = \frac{1}{B(4,3)} x^{4-1}(1-x)^{3-1} = 60x^3(1-x)^2$, $x \in [0,1]$
- Inversionsmethode nicht analytisch möglich:
  $u = \int_0^y f(x)\,\mathrm{d}x = 15y^4 - 24y^5 + 10y^6$ nicht analytisch invertierbar.
- Für Verwerfungsmethode, wähle $g(y) = 1$, also $Y \sim \mathcal{U}_{[0,1]}$.
  Konstante M:

$$
\begin{aligned}
f(x) &\leq Mg(x) = M \\
\Leftrightarrow 60x^3(1-x)^2 &\leq M
\end{aligned}
$$

  für $x \in [0,1]$, d.h. $M \approx 2{,}1$.

Ursprüngliches Problem

- Viele Stichproben $X_i \sim \mathbb{P}_f$ für Monte Carlo generieren:

$$\mathbb{E}_f[h(X)] \approx \frac{1}{N} \sum_{i=1}^{N} h(X_i)\,.$$

- Verschiedene Methoden für konkrete f:
  - Inversionsmethode
  - spezielle Methoden
  - Verwerfungsmethode
- Gemeinsamkeiten obiger Methoden:
  1. Erzeugen i.i.d. Samples X_i.
  2. Basieren auf relativ starken Annahmen an f.

Ursprüngliches Problem

- Viele Stichproben $X_i \sim \mathbb{P}_f$ für Monte Carlo generieren:

$$\mathbb{E}_f[h(X)] \approx \frac{1}{N}\sum_{i=1}^{N} h(X_i)\,.$$

- Verschiedene Methoden für konkrete f:
  - Inversionsmethode (i.A. teuer: muss $F(x) = \int_{-\infty}^{x} f(z)\,\mathrm{d}z$ invertieren)
  - spezielle Methoden (nur spezielle f)
  - Verwerfungsmethode (benötige Dichte g mit $f(x)/g(x) \leq$ konst., für die $\mathbb{P}_g$ leicht zu samplen ist)
- Gemeinsamkeiten obiger Methoden:
  1. Erzeugen i.i.d. Samples X_i.
  2. Basieren auf relativ starken Annahmen an f.

Ursprüngliches Problem

- Viele Stichproben $X_i \sim \mathbb{P}_f$ für Monte Carlo generieren:

$$\mathbb{E}_f[h(X)] \approx \frac{1}{N} \sum_{i=1}^{N} h(X_i)\,.$$

- Verschiedene Methoden für konkrete f:
  - Inversionsmethode (i.A. teuer: muss $F(x) = \int_{-\infty}^{x} f(z)\,\mathrm{d}z$ invertieren)
  - spezielle Methoden (nur spezielle f)
  - Verwerfungsmethode (benötige Dichte g mit $f(x)/g(x) \leq$ konst., für die $\mathbb{P}_g$ leicht zu samplen ist)

- Gemeinsamkeiten obiger Methoden:
  1. Erzeugen i.i.d. Samples X_i.
  2. Basieren auf relativ starken Annahmen an f.
     Häufig nicht erfüllt! Beispiel: Bayes a posteriori Dichte
     $f^{\theta|X} = \dfrac{f^{X|\theta} f^{\theta}}{f^{X}} \propto f^{X|\theta} f^{\theta}\,.$

Ursprüngliches Problem

- Viele Stichproben $X_i \sim \mathbb{P}_f$ für Monte Carlo generieren:

$$\mathbb{E}_f[h(X)] \approx \frac{1}{N}\sum_{i=1}^{N} h(X_i)\,.$$

- Verschiedene Methoden für konkrete f:
  - Inversionsmethode (i.A. teuer: muss $F(x) = \int_{-\infty}^{x} f(z)\,\mathrm{d}z$ invertieren)
  - spezielle Methoden (nur spezielle f)
  - Verwerfungsmethode (benötige Dichte g mit $f(x)/g(x) \leq$ konst., für die $\mathbb{P}_g$ leicht zu samplen ist)

- Gemeinsamkeiten obiger Methoden:
  1. Erzeugen i.i.d. Samples X_i. $\leftarrow$ Das ist nicht nötig!
  2. Basieren auf relativ starken Annahmen an f.
     Häufig nicht erfüllt! Beispiel: Bayes a posteriori Dichte
     $f^{\theta|X} = \dfrac{f^{X|\theta} f^{\theta}}{f^X} \propto f^{X|\theta} f^{\theta}\,.$

Übersicht

Idee

[Christian P. Robert, George Casella – *Monte Carlo Statistical Methods*]

Satz (Ergodensatz von Birkhoff)

Ist $(X_n)_{n\geq 0}$ eine Harris-rekurrente Markov-Kette mit invariantem W-Maß $\mathbb{P}_f$, so gilt für alle $h \in L^1(\mathbb{P}_f)$, dass

$$\frac{1}{N}\sum_{n=0}^{N-1} h(X_n) \to \int h(x)f(x)\,\mathrm{d}x = \mathbb{E}_f[h(X)], \quad \textit{für } N \to \infty.$$

- Für „große" n erzeugt die Markovkette annähernd Samples bezüglich f.
- Diese Samples sind *nicht* i.i.d.!

Markov-Ketten

Definition (Überganskern)

Eine Abbildung $K : \mathbb{R}^d \times \mathcal{B}(\mathbb{R}^d) \to \mathbb{R}$ heißt Überganskern, falls

1. $K(x, \cdot)$ für alle $x \in \mathbb{R}$ ein Wahrscheinlichkeitsmaß auf $\mathcal{B}(\mathbb{R}^d)$ ist und
2. $K(\cdot, B)$ für alle $B \in \mathcal{B}(\mathbb{R}^d)$ messbar ist.

Definition (Markov-Kette)

Ein Prozess $(X_n)_{n\geq 0}$ heißt Markov-Kette (MC), falls

$$\mathbb{P}[X_{n+1} \in B \mid X_0, \ldots, X_n] = \mathbb{P}[X_{n+1} \in B \mid X_n] = \int_B K(X_n, \mathrm{d}x)$$

für $n \in \mathbb{N}_0$, $B \in \mathcal{B}$, mit Übergangskern K.

Beispiel: AR(1)-Prozess, Random Walk, ...

Markov-Ketten

Eine Markov-Kette $(X_n)_{n\geq 0}$ mit Überganskern K heißt

- Zeit-homogen, falls $\mathbb{P}[X_{n+1} \in B \mid X_n] = \mathbb{P}[X_1 \in B \mid X_0]$;
- ν-irreduzibel für ein Maß ν auf $\mathcal{B}$, falls $\forall x \in \mathbb{R}^d, A \in \mathcal{B}$ mit $\nu(A) > 0$ $\exists n$: $K^n(x, A) > 0$;
- Harris-rekurrent, falls $\exists \nu : (X_n)$ ν-irreduzibel und $\forall A, \nu(A) > 0 \, \forall x \in A : \mathbb{P}[\exists n \geq 1 : X_n \in A \mid X_0 = x] = 1$

Definition

Ein Maß μ heißt zu einem Übergangskern K invariant, falls $\mu K = \mu$, also $\int \mu(\mathrm{d}x) K(x, B) = \mu(B) \; \forall B \in \mathcal{B}$.
Ist μ ein W-Maß, so heißt es auch stationär ((X_n) mit $X_0 \sim \mu$ ist stationär).

Für Existenz von μ genügt die Detailed Balance Bedingung:

$$\mu(\mathrm{d}x) K(x, \mathrm{d}y) = \mu(\mathrm{d}y) K(y, \mathrm{d}x) .$$

Der Metropolis-Hastings-Algorithmus

Ziel: Markov-Kette $(X_n)_{n\geq 0}$ mit invariantem W-Maß $\mathbb{P}_f$ simulieren.

1. Wähle X_0 zufällig/beliebig, sodass $f(X_0) > 0$.
2. Angenommen, wir haben schon X_n erzeugt.
   Erzeuge $Y_n \sim \mathbb{P}_{q(\cdot|x)}$.
   Setze $r(X_n, Y_n) := \min\left\{1, \frac{f(Y_n)q(X_n \mid Y_n)}{f(X_n)q(Y_n \mid X_n)}\right\}$ die Akzeptanzwkt.
   Setze

$$X_{n+1} := \begin{cases} Y_n & \text{mit Wahrscheinlichkeit } r(X_n, Y_n), \\ X_n & \text{mit Wahrscheinlichkeit } 1 - r(X_n, Y_n). \end{cases}$$

Spezialfall: Symmetrische Sampling-Dichte $q(y|x) = q(x|y)$
$\Rightarrow r(X_n, Y_n) = \min\{1, f(Y_n)/f(X_n)\}$.
Interpretation:

- Akzeptiere Y_n immer, falls es im Bereich größerer Dichte liegt.
- Ansonsten vertraue dem Sample weniger und „wirf eine Münze".

Der Metropolis-Hastings-Algorithmus

Ziel: Markov-Kette $(X_n)_{n\geq 0}$ mit invariantem W-Maß $\mathbb{P}_f$ simulieren.

1. Wähle X_0 zufällig/beliebig, sodass $f(X_0) > 0$.
2. Angenommen, wir haben schon X_n erzeugt.
   Erzeuge $Y_n \sim \mathbb{P}_{q(\cdot|x)}$.
   Setze $r(X_n, Y_n) := \min\left\{1, \dfrac{f(Y_n)q(X_n \mid Y_n)}{f(X_n)q(Y_n \mid X_n)}\right\}$ die Akzeptanzwkt.
   Setze

$$X_{n+1} := \begin{cases} Y_n & \text{mit Wahrscheinlichkeit } r(X_n, Y_n), \\ X_n & \text{mit Wahrscheinlichkeit } 1 - r(X_n, Y_n). \end{cases}$$

Spezialfall: Symmetrische Sampling-Dichte $q(y|x) = q(x|y)$
$\Rightarrow r(X_n, Y_n) = \min\{1, f(Y_n)/f(X_n)\}$.
Interpretation:

- Akzeptiere Y_n immer, falls es im Bereich größerer Dichte liegt.
- Ansonsten vertraue dem Sample weniger und „wirf eine Münze“.

Eigenschaften der Metropolis-Hastings MC (X_n)

- Übergangskern der Markov-Kette ist

$$K(x,B) = \mathbb{P}[X_{n+1} \in B \mid X_n = x] = \int_B r(x,y)q(y \mid x)\,\mathrm{d}y + \mathbb{1}_B(x)\int \big(1 - r(x,z)\big)q(z \mid x)\,\mathrm{d}z$$

für Borelmenge B.
K erfüllt Detailed Balance: $K(x,\,\mathrm{d}y)f(x)\,\mathrm{d}x = K(y,\,\mathrm{d}x)f(y)\,\mathrm{d}y$ (nachrechnen!)
$\Rightarrow \mu = \mathbb{P}_f$ ist invariante Verteilung von (X_n), da

$$\mathbb{P}[X_1 \in B \mid X_0 \sim \mu] = \int K(x,B)f(x)\,\mathrm{d}x = \iint \mathbb{1}_B(y)K(x,\,\mathrm{d}y)f(x)\,\mathrm{d}x = \iint \mathbb{1}_B(y)K(y,\,\mathrm{d}x)f(y)\,\mathrm{d}y = \int f(y)\mathbb{1}_B(y)\,\mathrm{d}y = \mu(B)\,.$$

Weitere Eigenschaften von (X_n)

- Falls $q(y \mid x) > 0$ für alle $(x, y) \in \mathcal{E} \times \mathcal{E}$, wobei $\mathcal{E} = \operatorname{supp}(f)$, so ist $(X_n)_{n \geq 0}$ irreduzibel (bzgl. Lebesgue-Maß), d.h. jeder Punkt im Support $\mathcal{E}$ von f kann in einem Schritt erreicht werden, denn $K(x, y) > 0$.

- Man kann zeigen: falls $\mathbb{P}[X_{n+1} = X_n] > 0$, so ist $(X_n)_{n \geq 0}$ *aperiodisch* und somit (da irreduzibel) Harris-rekurrent, d.h. der Ergodenzatz ist anwendbar.

- Es sollte leicht sein, von $q(\cdot \mid x)$ zu samplen.

- Erzeugte Samples hängen stark von der Konvergenzgeschwindigkeit gegen die stationäre Verteilung $\mathbb{P}_f$ ab.

- Samples sind *nicht* unabhängig; Approximation ist erst nach Burn-in-Phase gut; verwende $\frac{1}{N} \sum_{n=b+1}^{b+N} h(X_n)$.

Anwendungsbeispiel

- 28. Januar 1986: Explosion der Raumfähre Challenger wegen Materialermüdung en Dichtungsringen
- Wahrscheinlicher Grund: ungewöhnlich niedrige Außentemperatur von 31 $^\circ F$ (ca. 0 $^\circ C$)

Probleme	1	1	1	1	0	0	0	0	0	0	0	0
Temperatur	53	57	58	63	66	67	67	67	68	69	70	70

Probleme	1	1	0	0	0	1	0	0	0	0	0
Temperatur	70	70	72	73	75	75	76	76	78	79	81

Anwendungsbeispiel

- **Modelliere** Materialprobleme mit *logistischer Regression*:
  - Annahme: Beobachte $Y_i \overset{iid}{\sim} \text{Bernoulli}(p(x_i))$
  - $X_i = \text{Temperatur}$, $Y_i = \text{Materialproblem Ja/Nein}$
  - $p(x_i) = \mathbb{P}[Y_i = 1 \mid X_i = x_i] = \dfrac{\exp(\alpha + x_i\beta)}{1 + \exp(\alpha + x_i\beta)}$ für Parameter $\alpha, \beta \in \mathbb{R}$
  - Dies ist ein *verallgemeinertes lineares Modell*
- **Ziel:** Bestimme α, β anhand der Daten und mache Vorhersagen für ungesehene Temperaturen.

Anwendungsbeispiel

- **hier: Bayes-Analyse**
- a-Priori-Dichte: $\pi_\alpha(\alpha \mid b) = \frac{1}{b} e^\alpha e^{-e^\alpha / b}$, $\pi_\beta(\beta) = 1$, $b =$ Hyperparameter (für datengetriebene Wahl siehe [Robert & Casella, 2004])
- Likelihood-Funktion: Für $\mathbf{x} = (x_i)_{i=1}^{23}$, $\mathbf{y} = (y_i)_{i=1}^{23}$,

$$L(\alpha, \beta \mid \mathbf{x}, \mathbf{y}) = \prod_{i=1}^{23} p(x_i)^{y_i} (1 - p(x_i))^{1-y_i}$$

- a-Posteriori-Dichte: $f(\alpha, \beta) \propto L(\alpha, \beta \mid \mathbf{x}, \mathbf{y}) \pi(\alpha, \beta)$
- Vorschlagsdichte (unabhängig): $q(\alpha, \beta) = \pi_\alpha(\alpha \mid b) \varphi(\beta)$, mit $\mathcal{N}(0, 1)$-Dichte φ.
- Von q lässt sich leicht samplen.
- Akzeptanzwahrscheinlichkeit von (α', β'):

$$r\big((\alpha, \beta), (\alpha', \beta')\big) = \min\left\{1, \frac{L(\alpha', \beta' \mid \mathbf{x}, \mathbf{y}) \varphi(\beta)}{L(\alpha, \beta \mid \mathbf{x}, \mathbf{y}) \varphi(\beta')}\right\}$$