日本オペレーションズ・リサーチ学会
2004年春季研究発表会

離散化Dirichlet分布に従うPerfect Sampler

01605000　東京大学　松井知己　MATSUI Tomomi
02103880　東京大学　*来嶋秀治　KIJIMA Shuji

1. はじめに

本稿では離散化Dirichlet分布に厳密に従うサンプリングの手法を提案する。提案するアルゴリズムはマルコフ連鎖を用いたサンプリング法で、monotone CFTP (Coupling From The Past) アルゴリズムに基づくLas Vegas型の乱択アルゴリズム（randomized algorithm）である。我々は離散化Dirichlet分布を唯一の極限分布に持つ新しいマルコフ連鎖を提案する。このマルコフ連鎖の各推移では離散化ベータ分布に従う確率変数を生成し、アルゴリズムは平均$O(n^3 \ln \Delta)$回の推移で終了して、Dirichlet分布に厳密に従う確率変数を返す。ただし、nは変数の次元（パラメータの個数）であり、$1/\Delta$は離散化のグリッド幅の大きさである。従って、計算時間はパラメータの数値の大きさに依存しない。Dirichlet分布に従うサンプリングは、生物情報学の分野で観測データから（通常）疾患の原因遺伝子を見つけるためによく用いられている統計的手法において、多項分布の事前あるいは事後確率としてしばしば現れる[2]。

マルコフ連鎖を用いたサンプリングは、EMアルゴリズム、マルコフ連鎖モンテカルロ法、Gibbsサンプラーといった手法に現れる。マルコフ連鎖の収束性は、これらの手法を用いる上で重要な問題であり、近年、様々なマルコフ連鎖に対する多くの研究がなされている。しかし、近似サンプリングでは、どんなに推移を繰り返しても定常分布に厳密に従うサンプリングは不可能である。それに対し、Propp and Wilsonによって提案されたCFTPアルゴリズムは、マルコフ連鎖のシミュレーションを工夫することで定常分布に厳密に従うサンプリング（Perfect Sampling）を可能とし、画期的アルゴリズムとして注目を浴びている[3]。Perfect samplingを行う利点は、定常分布に厳密に従うサンプリングを行うことで、誤差パラメータを考慮する必要が無くなる点である。特に精度の高いサンプリングを要する時、Perfect Samplingは近似サンプリングよりも速いアルゴリズムとなる。

しかし、CFTPアルゴリズムはそのままでは、マルコフ連鎖の全状態数に比例する計算量を必要とするため、状態数の多い対象に対して効率的ではない。対象とするマルコフ連鎖に'monotone'の性質がある時、効率的なCFTPアルゴリズムの設計が可能となる。これをmonotone CFTPアルゴリズムと呼ぶ。一般にmonotoneの性質をもつマルコフ連鎖の設計は困難で、これまで実際にmonotone CFTPアルゴリズムの設計された例は少ない。

（連続）Dirichlet分布からのサンプリングのひとつの方法としては棄却サンプリングがある。しかしパラメータが小さいと、$n=2$の場合（ベータ分布）でさえ棄却の確率はすぐに大きくなってしまいパラメータ値が小さいとき、効率的でなくなる。実用的な場面において、Dirichlet分布は様々な次元とパラメータをもって現れることから、任意の次元とパラメータを持つDirichlet分布に対する効率的なサンプリングアルゴリズムが望まれている。

2. サンプリングアルゴリズム

整数（非負整数、正整数）の集合をZ (Z_+, Z_{++})で表す。非負実数パラメータ$u_1,\ldots,u_n$を持つDirichlet分布は確率変数ベクトル$P=(P_1,P_2,\ldots,P_n)$に対する確率分布で、密度関数は定義域$\{(p_1,p_2,\ldots,p_n)\in \mathrm{R}^n \mid p_1+\cdots+p_n=1, p_1,p_2,\ldots,p_n>0\}$に対して、$\frac{\Gamma(\sum_{i=1}^n u_i)}{\prod_{i=1}^n \Gamma(u_i)}\prod_{i=1}^n p_i^{u_i-1}$で表される。ただし、$\Gamma(u)$はガンマ関数である。本稿では、$n\geq 2$を仮定する。任意の整数$\Delta \geq n$に対して、定義域を格子幅$1/\Delta$で離散化し、整数ベクトルの離散的集合$\Omega$を$\Omega \stackrel{\mathrm{def.}}{=} \{(x_1,x_2,\ldots,x_n)\in \mathrm{Z}_{++}^n \mid x_i>0\ (\forall i),\ x_1+\cdots+x_n=\Delta\}$で定義する。非負実数パラメータ$u_1,\ldots,u_n$を持つ離散化Dirichlet確率変数は確率ベクトル$X=(X_1,\ldots,X_n)\in\Omega$で確率分布$\Pr[X=(x_1,\ldots,x_n)] \stackrel{\mathrm{def.}}{=} C_\Delta \prod_{i=1}^n (x_i/\Delta)^{u_i-1}$をもつ。ただし、$C_\Delta$は分配関数（規格化定数）で$(C_\Delta)^{-1} \stackrel{\mathrm{def.}}{=} \sum_{x\in\Omega}\prod_{i=1}^n (x_i/\Delta)^{u_i-1}$で定義される。任意の整数$b\geq 2$に対して、2次元整数ベクトルの集合$\Omega(b) \stackrel{\mathrm{def.}}{=} \{(Y_1,Y_2)\in \mathrm{Z}^2 \mid Y_1,Y_2>0,\ Y_1+Y_2=b\}$を導入し、非負実数パラメータ$u_i,u_j$を持つ分布関数$f_b(Y_1,Y_2 \mid u_i,u_j):\Omega(b)\to[0,1]$を$f_b(Y_1,Y_2\mid u_i,u_j) \stackrel{\mathrm{def.}}{=} C(u_i,u_j,b)Y_1^{u_i-1}Y_2^{u_j-1}$とする。ただし、定数$C(u_i,u_j,b))^{-1} \stackrel{\mathrm{def.}}{=} \sum_{(Y_1,Y_2)\in\Omega(b)} Y_1^{u_i-1}Y_2^{u_j-1}$は分配関数である。また、ベクトル$(g_b(0|u_i,u_j),g_b(1|u_i,u_j),\ldots,g_b(b-1|u_i,u_j))$を

$$g_b(k|u_i,u_j) \stackrel{\mathrm{def.}}{=} \begin{cases} 0 & (k=0), \\ \sum_{l=1}^k C(u_i,u_j,b) l^{u_i-1}(b-l)^{u_j-1} & (k\in\{1,2,\ldots,b-1\}). \end{cases}$$

で定義する。明らかに $0 = g_b(0|u_i,u_j) < g_b(1|u_i,u_j) < \cdots < g_b(b-1|u_i,u_j) = 1$ が成り立つ。

状態空間 Ω を持つマルコフ連鎖 $\mathcal{M}$ について述べる。現在の状態を $X \in \Omega$ とする。このとき、推移 $X \mapsto X'$ は次のように実行される。まず、実数乱数 $\lambda \in [1,n)$ を生成し、$i := \lfloor \lambda \rfloor$, $b := X_i + X_{i+1}$ とする。次に、$k \in \{1,2,\ldots,b-1\}$ を $g_b(k-1|u_i,u_{i+1}) \le (\lambda - \lfloor \lambda \rfloor) < g_b(k|u_i,u_{i+1})$ を満たす唯一の値とする。最後に、

$$X'_j := \begin{cases} k & (j=i), \\ b-k & (j=i+1), \\ X_j & (\text{otherwise}), \end{cases}$$ とする。このマルコフ連鎖は明らかに既約で非周期的である。また、detailed balance equations が成り立つことから、マルコフ連鎖 $\mathcal{M}$ の定常分布は離散化 Dirichlet 分布となる。

このマルコフ連鎖に対して、*update function* $\phi : \Omega \times [1,n) \to \Omega$ を $\phi(X,\lambda) \stackrel{\text{def.}}{=} X'$ で定義する。ただし、X' は上記の手続きで決まる状態である。時刻 t_1 から t_2 $(t_1 < t_2)$ へのマルコフ連鎖の推移の結果は実数乱数ベクトル $\boldsymbol{\lambda} = (\lambda[t_1], \lambda[t_1+1], \ldots, \lambda[t_2-1]) \in [0,1)^{t_2-t_1}$ を用いて、$\Phi_{t_1}^{t_2}(x,\boldsymbol{\lambda}) \stackrel{\text{def.}}{=} \phi(\phi(\cdots(\phi(x,\lambda[t_1]),\ldots,\lambda[t_2-2]),\lambda[t_2-1])$ と定義すると、$\Phi_{t_1}^{t_2}(x,\boldsymbol{\lambda}) : \Omega \times [0,1)^{t_2-t_1} \to \Omega$ で表される。また、特別な状態として $X_{\mathrm{U}}, X_{\mathrm{L}} \in \Omega$ を $X_{\mathrm{U}} \stackrel{\text{def.}}{=} (\Delta - n + 1, 1, 1, \ldots, 1)$, $X_{\mathrm{L}} \stackrel{\text{def.}}{=} (1, 1, \ldots, 1, \Delta - n + 1)$. とする。これらを用いて、アルゴリズムを記述する。

アルゴリズム A

Step 1. シミュレーションの開始時刻を $T := -1$ とし、$\boldsymbol{\lambda}$ を空列とする。

Step 2. 実数乱数 $\lambda[T], \lambda[T+1], \ldots, \lambda[\lceil T/2 \rceil - 1] \in [1,n)$ を生成し、$\boldsymbol{\lambda} = (\lambda[T], \lambda[T+1], \ldots, \lambda[-1])$ とする。

Step 3. 2 本のマルコフ連鎖の初期状態を X_{U} と X_{L} とし、数列 $\boldsymbol{\lambda}$ を用いて、時刻 T から 0 に至るまで update function ϕ に従って推移させる。

Step 4. もし $\exists Y \in \Omega$, $Y = \Phi_T^0(X_{\mathrm{U}}, \boldsymbol{\lambda}) = \Phi_T^0(X_{\mathrm{L}}, \boldsymbol{\lambda})$ なら、Y を返し停止する。
そうでなければ、開始時刻を $T := 2T$ に更新し、Step 2 へ進む。

以下の 2 つの定理は、アルゴリズム A に対する重要な定理である。

定理 1 アルゴリズム *A* は確率 *1* で（有限時間で）終了し、状態を返す。得られる状態は、離散化 *Dirichlet* 分布に厳密に従う確率変数である。

条件 1 各パラメータは非増加順に並ぶ。すなわち、$u_1 \ge u_2 \ge \cdots \ge u_n$ を満たす。

定理 2 条件 *1* の下でアルゴリズム *A* の計算時間の期待値は $\mathrm{O}(n^3 \ln \Delta)$ で押さえられる。ただし、n は次元（パラメータの個数）を表し、$1/\Delta$ は離散化のグリッド幅を表す。

3. monotone CFTP

アルゴリズム A の Step 4 の一行目を [If $\exists Y \in \Omega, \forall x \in \Omega$, $Y = \Phi_T^0(x,\boldsymbol{\lambda})$, then return Y and stop.] のように置き換えたアルゴリズムをアルゴリズム B とする。アルゴリズム B は通常の CFTP アルゴリズムであり、任意のマルコフ連鎖に対しても定常分布に厳密に従うサンプリングを実現する。したがってアルゴリズム A と B が本質的に等価であることを示せば定理 1 は示される。以下で定理 1 の証明の鍵となる補題を導く。

まず、状態空間 Ω に半順序を導入する。任意のベクトル $X \in \Omega$ に対して、*cumulative sum vector* $\mathrm{c}_X = (\mathrm{c}_X(0), \mathrm{c}_X(1), \ldots, \mathrm{c}_X(n)) \in \mathrm{Z}_+^{n+1}$ を $\mathrm{c}_X(i) \stackrel{\text{def.}}{=} \begin{cases} 0 & (i=0), \\ X_1 + X_2 + \cdots + X_i & (i \in \{1,2,\ldots,n\}), \end{cases}$ と定義する。任意の状態対 $X, Y \in \Omega$ に対して、$X \succeq Y$ を $\mathrm{c}_X \ge \mathrm{c}_Y$ で定義する。明らかに、"$\succeq$" は Ω 上の半順序である。また、$\forall X \in \Omega$, $X_{\mathrm{U}} \succeq X \succeq X_{\mathrm{L}}$ であることも容易に分かる。

補題 1 （*monotone* マルコフ連鎖）定義した *update function* ϕ に対して、$\forall \lambda \in [1,n)$, $\forall X, \forall Y \in \Omega$, $X \succeq Y \Rightarrow \phi(X,\lambda) \succeq \phi(Y,\lambda)$ が成り立つ。

参考文献

[1] 鎌谷直之編著, 「ポストゲノム時代の遺伝統計学」, 羊土社, 2001.

[2] Matsui, T. and Kijima, S.,: Polynomial time perfect sampler for discretized Dirichlet distribution, METR 2003-17, Mathematical Engineering Technical Reports, University of Tokyo, 2003.
(available from http://www.keisu.t.u-tokyo.ac.jp/Research/techrep.0.html)

[3] Propp, J. and Wilson, D.: Exact sampling with coupled Markov chains and applications to statistical mechanics, *Random Structures and Algorithms*, **9** (1996), 232–252.