4H4-5

決定的サンプリング法 Herded Gibbs の連続分布への拡張

Efficient Deterministic Sampling: Herded Gibbs for Continuous Distributions

山下 洋史 *1　鈴木 秀幸 *2
Hiroshi Yamashita　Hideyuki Suzuki

*1東京大学大学院情報理工学系研究科
Graduate School of Information Science and Technology, the University of Tokyo

*2大阪大学大学院情報科学研究科
Graduate School of Information Science and Technology, Osaka University

Herded Gibbs (HG) and discretized herded Gibbs (DHG) are deterministic sampling algorithms designed to be used instead of random Gibbs sampling. These two algorithms use herding, which is a dynamical system that can generate samples deterministically, and are efficient in estimating expectations. However, these are only applicable to graphical models which have discrete variables. By following the procedure of HG and DHG, we propose a new deterministic sampling algorithm, continuous DHG (CDHG), which can generate samples from a graphical model which has continuous variables. We also investigate the performance of CDHG numerically and show that it can estimate expectations more efficiently than random Gibbs sampling.

1. 導入

統計や機械学習などの分野ではモンテカルロ法 (MC) による期待値の計算がよく行われる．これは，確率分布 $p(x)$ の上で関数 $f(x)$ の値の期待値 $E_{x\sim p}[f(x)] = \int f(x)p(x)dx$ を求めるという問題である．モンテカルロ法では，確率分布 p に従うサンプル列 $x^{(1)}, x^{(2)}, \ldots, x^{(T)}$ を生成し，それによって $E_{x\sim p}[f(x)] \approx (1/T)\sum_{t=1}^{T} f(x^{(t)})$ と近似する．

一般の確率分布 p からの独立同分布サンプリングは困難であることが多く，この時にはマルコフ連鎖モンテカルロ法 (MCMC) という手法が用いられる．MCMC では，1 つ前の時刻のサンプル $x^{(t-1)}$ をもとにして次の時刻のサンプル $x^{(t)}$ に遷移するということを繰り返す．この遷移をうまく設計することにより，t を十分大きくとったときに $x^{(t)}$ の分布が p に収束するようにできる．

Gibbs sampling は マルコフ連鎖モンテカルロ法の一種である．これは，多変数からなる確率分布からのサンプリングに対して用いられ， 1 つの変数の更新を繰り返すことで遷移を行う．

Herded Gibbs (HG) は，Bornn ら [Bornn et al., 2013] によって提案された， Gibbs sampling を決定的に行うアルゴリズムである．これは，Gibbs sampling における条件付き分布からのサンプリングを herding [Welling, 2009] と呼ばれる手法に置き換えることによって得られる．Herding は，決定的なサンプリングアルゴリズムであり，サンプルの特徴量の期待値を指定した値に合わせながらサンプリングを行う．

Herded Gibbs での期待値の推定の収束は，限られた条件の下では理論的に保証されており，さらに Gibbs sampling と比べて速いことが示されている．それ以外の一般の状況では，Gibbs sampling を上回る性能を持つものの，推定値がバイアスを含む場合があることが報告されている．

Herded Gibbs では，更新に用いる条件付き分布のそれぞれに対応した重み変数が用いられる．大規模な問題ではこの重み変数が大量に必要になり，空間計算量が増大する．この問題に対処するため，重み変数の数を削減した discretized herded Gibbs (DHG) が提案された [Eskelinen, 2013].

Herded Gibbs は離散的な分布に対するアルゴリズムであって，連続分布に対して適用することはできない．本稿では，HG と DHG のアイデアをもとに，変数が連続値をとるようなグラフィカルモデルに対する効率的な決定論的サンプリング法を提案する．

連絡先: 山下 洋史，東京大学大学院情報理工学系研究科，
hiroshi_yamashita@mist.i.u-tokyo.ac.jp

2. 背景

2.1 Herding

標本空間を有限集合 $\mathcal{X}$ とし，そこにいくつか特徴量 $\phi_\alpha(x) : \mathcal{X} \to \mathbb{R}$ が決められているとする．それぞれの特徴量にはモーメント $\mu_\alpha \in \mathbb{R}$ が与えられており，特徴量の期待値 $E[\phi_\alpha(x)]$ が与えられたモーメント μ_α に等しいような分布を求めたいという状況を考える．Herding [Welling, 2009] はこの条件を満たす分布からサンプル列 $x^{(1)}, x^{(2)}, \ldots$ を発生させる．Herding によって得られるサンプルの分布は，モーメントの条件 $E[\phi_\alpha(x)] = \mu_\alpha$ をみたす分布のなかで比較的エントロピーが高いという特徴を持つ．表記を簡単にするため，特徴量とモーメントについて，それぞれを縦に並べてベクトルにしたものを $\boldsymbol{\phi}(x), \boldsymbol{\mu}$ と書くことにする．

Algorithm 1 Herding

for $t = 1, 2, \ldots, T$ **do**
$$x^{(t)} = \operatorname*{argmax}_{x\in\mathcal{X}} \langle \boldsymbol{w}^{(t-1)}, \boldsymbol{\phi}(x)\rangle$$
$$\boldsymbol{w}^{(t)} = \boldsymbol{w}^{(t-1)} + \boldsymbol{\mu} - \boldsymbol{\phi}(x^{(t)})$$
end for

Herding のアルゴリズムを Algorithm 1 に示す．$\boldsymbol{w}^{(t)}$ は重みと呼ばれる．重みの更新は決定的に行われる．これを力学系として考えると，重みは弱カオス的な挙動を示す．この弱カオス的な挙動により，モーメントの条件を満たしながらも擬似的にランダムなサンプル列を発生させることができる．

また，このアルゴリズムは，各時刻毎に

$$\left\| \boldsymbol{\mu} - \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{\phi}(x^{(t)}) \right\|$$

を最小にする $x^{(T)}$ を求めているのと等価である．

2.2 Herded Gibbs

Herding は決定的なサンプリングアルゴリズムであるが，グラフィカルモデルにはそのまま適用することはできない．Bornn ら [Bornn et al., 2013] はグラフィカルモデルに対して決定的にサンプリングを行うアルゴリズムである herded Gibbs (HG) を提案した．

$\{0,1\}$ の 2 値をとる N 個の変数からなるグラフィカルモデルを考える．変数 x_i が依存するすべての変数を集めたものを近傍 $\mathcal{N}(i)$ と呼ぶことにする．すなわち，$p(x_i = 1|\boldsymbol{x}_{-i}) = p(x_i = 1|\boldsymbol{x}_{\mathcal{N}(i)})$ である．Herded Gibbs は，Gibbs sampling における変数のサンプリングを herding を用いて行う．各変数 x_i に対し，その近傍 $\boldsymbol{x}_{\mathcal{N}(i)}$ の状態数と同じ $2^{|\mathcal{N}(i)|}$ 個の重み変数 $w_{i,\boldsymbol{\sigma}}$ を用意する．変数 x_i の更新は，近傍の状態 $\boldsymbol{\sigma}$ に対応した $w_{i,\boldsymbol{\sigma}}$ を用いて herding を行う．具体的なアルゴリズムは次の通りである．

Algorithm 2 Herded Gibbs

for $t := 1, 2, \ldots, T$ **do**
  for $i := 1, 2, \ldots, N$ **do**
    $\boldsymbol{\sigma} = \boldsymbol{x}_{\mathcal{N}(i)}^{(t-1)}$ ▷ x_i の近傍の状態
    $p_{i,\boldsymbol{\sigma}} = P(x_i = 1 \mid \boldsymbol{\sigma})$
    $x_i^{(t)} = \mathbb{I}(w_{i,\boldsymbol{\sigma}}^{(t-1)} > 0)$
    $w_{i,\boldsymbol{\sigma}}^{(t)} = w_{i,\boldsymbol{\sigma}}^{(t-1)} + p_{i,\boldsymbol{\sigma}} - x_i^{(t)}$
  end for
end for

5, 6 行目が herding の $\mathcal{X} = \{0,1\}$, $\boldsymbol{\phi}(\boldsymbol{x}) = x_i$, $\boldsymbol{\mu} = p_{i,\boldsymbol{\sigma}}$ とした場合に対応している．Herded Gibbs は，単に Gibbs sampling を決定的なものに書き換えただけではなく，Gibbs sampling を上回る性能をもつことが示されている．理論的には，完全グラフ上のグラフィカルモデルに対して，Gibbs sampling の推定値の分散は $O(1/\sqrt{T})$ で減少するのに対して，herded Gibbs では $O(1/T)$ の速さで減少することが示されている [Bornn et al., 2013]．ただし，それ以外の一般のグラフの上では，推定値が正しい値に収束するとは限らず，そのときの herded Gibbs の近似精度についてはくわしい解析はなされていない．実験的には，画像処理や自然言語処理のタスクにおいて herded Gibbs が Gibbs sampling と同程度もしくはそれを上回る性能をもつという結果が報告されている [Bornn et al., 2013]．

2.3 Discretized Herded Gibbs

HG は各変数に対して重み変数が $2^{|\mathcal{N}(i)|}$ 個必要であり，大規模なグラフィカルモデルに対しては，空間計算量の観点からみて実行不可能である．

この問題を解決するため，discretized herded Gibbs (DHG) というアルゴリズムが提案されている [Eskelinen, 2013]．DHG のアルゴリズムを Algorithm 3 に示す．DHG では，各変数に対して，$[0,1]$ を B 個の区間に分割する．更新の際に用いる重み変数を，条件付き確率 $P(x_i = 1|\boldsymbol{x}_{\mathcal{N}(i)})$ が入る区間によって定める．p が同じ区間に入る近傍の状態の間で更新に用いる重み変数は共有される．特定の $w_{i,b}$ に対して，更新に用いる確率の値は毎回異なりうる．そのため，DHG での重み変数の更新は必ずしも herding の力学系とは一致せずに近似したものとなり，推定値にバイアスが生じる．分割の細かさを表すパラメータ B が大きいほどこの近似が良くなるため，バイアスは小さくなる．つまり，B に比例する空間計算量とバイアスの間にはトレードオフの関係が生じている．

Algorithm 3 Discretized herded Gibbs

for $t := 1, 2, \ldots, T$ **do**
  for $i := 1, 2, \ldots, N$ **do**
    $p = P(x_i = 1 \mid \boldsymbol{x}_{\mathcal{N}(i)}^{(t-1)})$
    b を $p \in [\frac{b}{B}, \frac{b+1}{B})$ を満たす整数とする．
    $x_i^{(t)} = \mathbb{I}(w_{i,b}^{(t-1)} > 0)$
    $w_{i,b}^{(t)} = w_{i,b}^{(t-1)} + p - x_i^{(t)}$
  end for
end for

3. 連続分布上での Herded Gibbs

連続的なグラフィカルモデル上に拡張された herded Gibbs を提案する．以下，このアルゴリズムを continous DHG (CDHG) と呼ぶ．

3.1 アルゴリズム

DHG と同様に，1 つの変数に対する周辺分布をいくつかの集合に分け，その中で重み変数を共有して変数の更新を行う．(i) 隣接変数のどのような配置の間で「重みの共有」を行うか，(ii) 共有された重みをどのように更新をするか，(iii) 重みからどのように新しい変数の値を決めるか，の問題があるが，CDHG では次のようにする．

(i) DHG では条件付き確率の値を区間に分割し，同じ区間のなかで重み変数を共有していた．これに倣って，次のように共有を行う．条件付き周辺分布が低次元のパラメータベクトル $\boldsymbol{\theta}(\boldsymbol{x}_{\mathcal{N}(i)})$ によって $p(x_i \mid \boldsymbol{x}_{\mathcal{N}(i)}) = p_i(x_i; \boldsymbol{\theta}(\boldsymbol{x}_{\mathcal{N}(i)}))$ と表されているとする．パラメータベクトルのなす空間を適当な大きさの領域 $\{R_b\}$ に分割し，各領域に 1 つずつ重み変数 $w_{i,b}$ を対応させる．更新は $\boldsymbol{\theta}(\boldsymbol{x}_{\boldsymbol{\mathcal{N}(i)}})$ が入る領域 R_b に対応する重み変数 $w_{i,b}$ を用いて行う．

(ii) Algorithm 2 の 5, 6 行目と同様の状況として，$\mathcal{X} = \{0,1\}$, $\boldsymbol{\phi}(x) = x$, $\boldsymbol{\mu} = \pi$ とした場合の herding を考える．これは，重み変数 w を $w \leftarrow w \bmod 1$ とおき直すことで Algorithm 4 のように等価なアルゴリズムに書き換えることができる．

Algorithm 4 Herding におけるコイントス

for $t = 1, 2, \ldots, T$ **do**
  $x^{(t)} = \mathbb{I}(w^{(t-1)} < \pi)$
  $w^{(t)} = w^{(t-1)} + \pi \pmod 1$
end for

このアルゴリズムでは，内部状態 w は回転写像によって更新され，π が無理数のとき，w は $[0,1]$ 区間内に一様に分布する．

CDHG でも同様に，w は $[0,1]$ 区間内に一様に分布するように更新していく．また，推定の効率を上げるために，乱数列と比較してより均一に分布することが知られている次のような数列を用いる．

- 黄金比回転列 $w^{(t)} \leftarrow w^{(t-1)} + (\sqrt{5}-1)/2 \pmod 1$

- (2 を底とした) van der Corput 列 [van der Corput, 1935] $x^{(t)}$ を用いて，$w^{(t)} = w^{(0)} + x^{(t)} \pmod 1$

の 2 つの列を用いることを検討する．Van der Corput 列 $x^{(t)}$ は，t が 2 進数で $t = b_K \cdots b_2 b_1$ と表されるとして，

$$x^{(t)} = \sum_{k=1}^{K} b_k 2^{-k}$$

である．初期値 $w^{(0)}$ は $[0,1]$ から一様ランダムにとる．$(x^{(t)}) = 0, \frac{1}{2}, \frac{1}{4}, \frac{3}{4}, \frac{1}{8}, \frac{5}{8} \cdots$ となり，$[0,1]$ 区間内に一様に分布する．

(iii) 乱数の発生に逆関数法を用いる．逆関数法は，1 変数確率分布からのサンプリングにおいて，一様乱数 $u \in [0,1]$ と累積分布関数の逆関数 $F^{-1}(x)$ を用いてて $x = F^{-1}(u)$ とサンプルする方法である．重み変数 w は $[0,1]$ 区間内に均一に分布するため，これを u の代わりに用いることにする．

4. 数値実験

提案した CDHG の性能を調べるため，2 変数正規分布

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right)$$

からのサンプリングを行った．条件付き分布 $p(x_i \mid x_j)$ $(j \neq i)$ は正規分布 $\mathcal{N}(\mu, \sigma^2)$ となるが，$\mu = \rho x_j$，$\sigma^2 = 1 - \rho^2$ であって，σ は x_j によらず一定である．よって，パラメータを μ のみによって $\ldots, [-0.5/B, 0.5/B), [0.5/B, 1.5/B), \ldots$ と幅 $1/B$ ごとに分割した．B は分割のパラメータである．

通常の Gibbs sampling と，重みの更新に使う列を黄金角回転列 (G) あるいは van der Corput 列 (V) とした 2 種類の CDHG を比較する．以下ではカッコ内の数字は CDHG における B の値を表す．

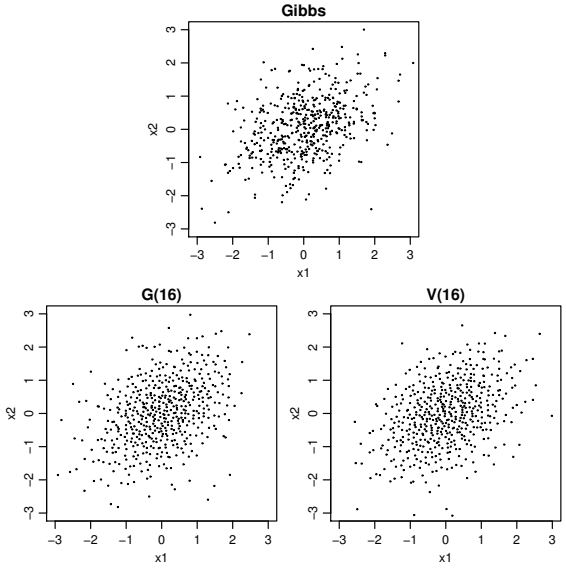

図 1: 各アルゴリズムの初めの 500 点の分布

Gibbs sampling と $B = 16$ とした CDHG での初めの 500 点をプロットしたものを図 1 に示す．Gibbs sampling と比べ，CDHG ではどちらの数列を使った場合もサンプル点がより均等に散らばっていることが視覚的に分かる．

また，各アルゴリズムを用いて $E[x_1], E[x_1 x_2]$ の推定を 100 回ずつ行った．CDHG では分割のパラメータを $B = 4, 16, 64$ の 3 通りに変えて実験を行った．

$E[x_1], E[x_1 x_2]$ の推定誤差の箱ひげ図を図 2 に，平均と分散を表 1 に示す．箱ひげ図では，太線は中央値を，四角は 25 パーセンタイル点から 75 パーセンタイル点までの範囲を表す．また，点線は最小値から最大値までの範囲を表す．

どの T においても CDHG の分散が減少しており，大きい T においてその度合が大きい．分散の小ささを見ると，小さい $T = 10^2$ では分割の粗い $G(4), V(4)$ が最良であるが，T が大きくなるにつれて分割の細かいものが同程度あるいは上回るようになる傾向がみられる．また，分割の粗い G(4),V(4) では $T = 10^6$ において推定のバイアスがばらつきを上回ってしまっている．

5. まとめ

本稿では，herded Gibbs の連続分布への拡張を行い，2 変数正規分布のサンプリングにおいて，Gibbs sampling と性能を比較した．

モンテカルロ積分によって $\int_{u \in [0,1]^s} f(u)du$ を求める際に，$[0,1]^s$ 上の i.i.d. 乱数列を用いるのではなく，$[0,1]^s$ 上に均等に散らばる低食い違い列 (low-discrepancy sequence) と呼ばれる決定的な点列を用いることによって推定値の分散を削減する手法があり，準モンテカルロ法 (quasi-Monte Carlo, QMC) と呼ばれている．Owen ら [Owen & Tribble, 2005] はその低食い違い列を Gibbs sampling に用いることで準モンテカルロ法のような推定値の分散の削減ができることを示した．この手法は MCQMC (Markov Chain quasi-Monte Carlo) と呼ばれる．

MCQMC と本稿で提案された連続分布上での herded Gibbs は決定的なサンプリングアルゴリズムであり，類似点も多い．その詳細な考察と，比較実験を今後の課題としたい．

6. 謝辞

本研究の一部は，総合科学技術・イノベーション会議により制度設計された革新的研究開発推進プログラム (ImPACT) により，科学技術振興機構を通して委託されたものである．

参考文献

Bornn, L., Chen, Y., de Freitas, N., Eskelin, M., Fang, J., & Welling, M. (2013). Herded Gibbs sampling. In *International Conference on Learning Representations (ICLR)*.

Eskelinen, M. (2013). Herded Gibbs and discretized herded Gibbs sampling. Master's thesis, The University of British Columbia.

Owen, A. B. & Tribble, S. D. (2005). A quasi-monte carlo metropolis algorithm. In *Proceedings of the National Academy of Sciences of the United States of America*, volume 102 (pp. 8844–8849).

van der Corput, J. (1935). Verteilungsfunktionen. I. Mitt. *Proc. Akad. Wet. Amsterdam*, 38, 813–821.

Welling, M. (2009). Herding dynamical weights to learn. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML)* (pp. 1121–1128).

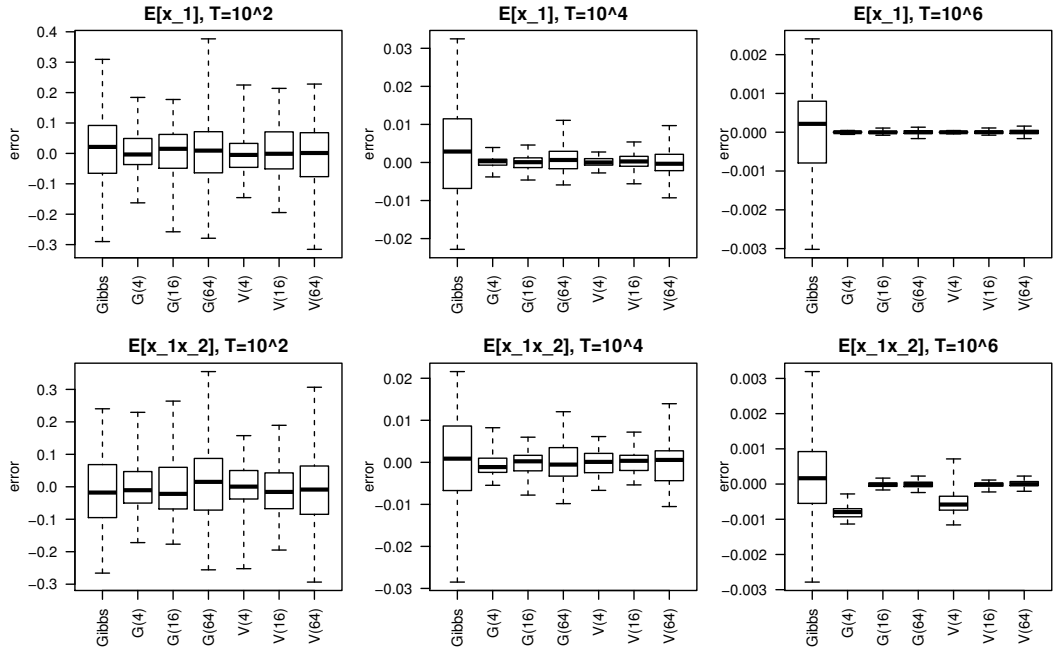

図 2: 推定誤差の箱ひげ図

表 1: $E[x_1]$(上段), $E[x_1x_2]$(下段) の推定誤差

	$T=10^2$		$T=10^4$		$T=10^6$	
	mean	variance	mean	variance	mean	variance
Gibbs	1.10e-02	1.35e-02	2.15e-03	1.56e-04	-4.05e-06	1.42e-06
G(4)	4.14e-03	4.10e-03	1.25e-04	1.83e-06	1.39e-06	3.36e-10
G(16)	3.01e-03	7.27e-03	5.58e-05	3.99e-06	2.12e-06	1.04e-09
G(64)	1.08e-02	1.16e-02	7.09e-04	1.09e-05	1.00e-06	3.24e-09
V(4)	-7.23e-03	3.78e-03	1.31e-04	1.26e-06	5.00e-08	3.38e-10
V(16)	2.42e-03	6.92e-03	1.58e-04	4.02e-06	9.60e-07	1.41e-09
V(64)	1.66e-03	1.08e-02	-3.88e-04	1.23e-05	7.16e-06	3.61e-09

	$T=10^2$		$T=10^4$		$T=10^6$	
	mean	variance	mean	variance	mean	variance
Gibbs	-9.20e-03	1.19e-02	5.68e-04	1.09e-04	1.59e-04	1.18e-06
G(4)	1.35e-03	5.48e-03	-7.67e-04	7.34e-06	-7.96e-04	6.67e-07
G(16)	-2.36e-03	8.20e-03	-2.66e-04	7.89e-06	-1.47e-05	5.16e-09
G(64)	8.61e-03	1.24e-02	-4.39e-05	2.07e-05	-5.70e-06	9.04e-09
V(4)	-4.55e-04	4.41e-03	-2.10e-04	7.85e-06	-5.53e-04	4.05e-07
V(16)	-1.51e-02	6.64e-03	3.18e-05	6.52e-06	-1.51e-05	4.12e-09
V(64)	-1.86e-02	1.34e-02	-3.76e-04	2.18e-05	1.04e-05	7.90e-09