3J1-OS7-5

Slice Samplingを用いたSAT技術による確率推論

Probabilistic inference based on Slice Sampling and SAT technologies

山口 雅博 Masahiro Yamaguchi　石畠 正和 Masakazu Ishihata　佐藤 泰介 Taisuke Sato

東京工業大学大学院情報理工学研究科
Graduate School of Information Science and Engineering, Tokyo Institute of Technology

We propose a CS sampling(Constraint-based Slice sampling) with Slice Sampling which is one of MCMC(Markov chain Monte Carlo) methods. A CS sampling can get samples efficiently from conditional distribution under logical constraints. A CS sampling makes it possible to compute expectations approximately in probabilistic models described by Boolian formulas.

1. はじめに

近年，自然言語や関係データベースなどの記号データ中の不確実性を扱う方法として確率モデルが利用されている．確率モデルに基づく予測やモデルパラメータの学習には，観測データが与えられた場合の条件付き確率の計算が必要となる．例えば，文脈自由文法の確率的拡張である PCFG(probabilistic context free grammar, 確率的文脈自由文法) を用いて与えられた文に対する最尤の構文木を予測する場合，与えられた文に対する構文木の条件付き確率分布を計算する必要がある．一方，これらの記号データを扱う確率モデルを統合的に扱う方法として，論理に基づく確率モデリングが提案されている [Sato 10, Richardson 06]．論理に基づく確率モデルでは，確率事象は命題論理式で表現される．その結果条件付き確率分布は，条件を表す命題論理式で制約付けられた確率分布となり，論理式を真にするすべて充足解を求める必要がある．しかしながら，一般的に論理式に対するすべての充足解を求めることは NP-hard であるため，条件部が複雑な事象であるとき，その条件付き確率計算は非常に困難となる．この計算を近似的に行う方法としてサンプリングに基づく方法が考えられる．例えば SampleSAT[Wei 04] は論理式を真にする充足解からの一様サンプリングを実現しており，条件付き確率に関する期待値は SampleSAT を用いた重点サンプリングにより近似的に計算可能となる．しかし重点サンプリングは，真の条件付き分布が一様分布から離れるに連れて近似精度が悪化する．これに対して， MCMC(Markov chain Monte Carlo, マルコフ連鎖モンテカルロ) 法に基づく近似計算は，重点サンプリングよりも少ないサンプル数でよりよい近似精度が得られることが知られている [Bishop 06]．そこで本論文では，論理式に基づく確率モデルに対する MCMC サンプリング法である **CS sampling** (Constraint-based Slice sampling) を提案する．CS sampling は MCMC 法の一種である Slice Sampling[Neal 03] を論理式で条件付けられた (制約付けられた) 分布に適用した手法であり，任意の論理式で制約付けられた分布からサンプリングが可能である．更に CS sampling は既存の SAT 技術を利用することで SampleSAT よりも高速なサンプリングが可能である．本論文では人工データに対して両手法を適用し，それらの速度と得られた近似分布と真の分布の KL ダイバージェンスを比較し，CS sampling が SampleSAT より高速かつ高精度であることを実験的に示す．

連絡先: 山口雅博，東京工業大学大学院情報理工学研究科，152-0033 東京都目黒区大岡山 2-12-1-W8E501, yamaguchi@mi.cs.titech.ac.jp

2. 準備

ここではまず，論理に基づく確率モデルを定式化し，サンプリングの対象となるとなる条件付き確率分布が制約論理式を用いて表現できることを示す．次にその制約論理式からの一様サンプリングを実現する SampleSAT について簡単に説明する．

2.1 論理に基づく確率モデル

本論文では論理に基づく確率モデルを用いる．互いに独立な命題変数集合を $\boldsymbol{X} = \{X_1, X_2, \ldots, X_N\}$ とし，その実現値を $\boldsymbol{x} = \{x_1, x_2, \ldots, x_N\} (\forall i, x_i \in \{0,1\})$ とする．すると $\boldsymbol{X}$ の同時分布 $p(\boldsymbol{X} = \boldsymbol{x})$ は以下の様に定義される．

$$p(\boldsymbol{X} = \boldsymbol{x}) \equiv \prod_{i=1}^{N} \theta_i^{x_i} \, (1-\theta_i)^{(1-x_i)}$$

ここで θ_i は X_i が真をとる確率である．次に $\boldsymbol{X}$ の命題論理式 F の確率を定める．簡単化のため，F が表現する論理関数も F と書く．すると論理関数 $F(\boldsymbol{X})$ は $\boldsymbol{X}$ に対するすべての可能な値の割り当て $\Phi_{\boldsymbol{X}} (\equiv \{0,1\}^N)$ から $\{0,1\}$ への写像であり $F(\boldsymbol{X}) \in \{0,1\}$ である．今，論理式 F の充足解のすべての集合 $\{\boldsymbol{x} \mid \boldsymbol{x} \in \Phi_{\boldsymbol{X}}, F(\boldsymbol{x}) = 1\}$ を Φ_F と書く．すると論理式 F の確率 $p(F)$ は Φ_F を用いて以下の様に定義される．

$$p(F) \equiv \sum_{\boldsymbol{x} \in \Phi_F} p(\boldsymbol{x})$$

ここで $p(\boldsymbol{x})$ は $p(\boldsymbol{X} = \boldsymbol{x})$ の省略形である．

さて，$\Phi_F \subseteq \Phi_{\boldsymbol{X}}$ は確率事象であるため，論理式 F もまた確率事象である．また，任意の確率事象は対応する論理式を持つ．従ってある確率事象で条件付けられた確率分布は，論理式で制約付けられた確率分布として表現できる．論理式 F で条件付けられた (制約付けられた) 確率 $p(\boldsymbol{x} \mid F)$ は以下の様に定義される．

$$p(\boldsymbol{x} \mid F) = \frac{p(\boldsymbol{x}, F)}{p(F)} = \frac{p(\boldsymbol{x}) F(\boldsymbol{x})}{\sum_{\boldsymbol{x}' \in \Phi_F} p(\boldsymbol{x}')}$$

この確率計算を行うには論理式 F の充足解の集合 Φ_F を求める必要があが，一般にこの問題は NP-hard であるので，条

件付き確率 $p(\boldsymbol{x} \mid F)$ の計算も指数的な時間が掛かる事が予想される．本論文で提案する CS sampling$p(\boldsymbol{X} \mid F)$ に関する期待値をサンプリング近似する手法である．

2.2 SampleSAT を用いた近似計算

条件付き確率 $p(\boldsymbol{X} \mid F)$ に関する期待値を素朴に近似する方法として，論理式 F の充足解 Φ_F からの一様サンプリングを利用した重点サンプリングが考えられる．Φ_F から一様サンプリングを実現する方法として SampleSAT[Wei 04] が提案されている．ただし SampleSAT では F の形を CNF(conjunctive normal form, 連言標準型) に限定している．SampleSAT は SA(Simulated annealing, 焼き鈍し法) と確率的 SAT ソルバーである WalkSAT[Selman 96] を組み合わせた手法である．WalkSAT は GSAT[Selman 92] にランダム性を導入したアルゴリズムであり，RandomWalk と GreedyWalk の 2 種類の戦略を確率的に選ぶ．これより SampleSAT は以下の 3 種類の探索戦略を確率的に行う局所探索手法と言える．

- RandomWalk: CNF 中の充足されていない節をランダムに選択し，その中の変数をランダムに 1 つ選んで値を反転する．
- GreedyWalk: 値を反転した時に充足される節数を最大化する変数を選び，値を反転する．
- Simulated annealing: 全変数から変数をランダムに一つ選択し，それを反転したときに充足される節数が増加するならば確率 1 で，減少するならば確率 $e^{-\tau}$ でその変数の値を反転する．

戦略の選択率と温度パラメータ τ はサンプリングの一様性と効率に強く影響する．SA の選択率が高いほどサンプリングは低速になるが，結果は一様分布に近くなる．逆に WalkSAT の選択率が高くなるとサンプリングは高速化するがその結果は一様分布から遠ざかる．本論文では [Selman 96, Wei 04] を参考に SA の選択確率を 0.5，WalkSAT 中での選択率を 0.5，$\tau = 0.1$ とした．

今，サンプル列 $\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(K)}\}$ を SampleSAT により得られた Φ_F からの (近似的な) 一様サンプルとする．すると関数 $f(\boldsymbol{x})$ の条件付き確率 $p(\boldsymbol{X} \mid F)$ による期待値 $E[f] \equiv \sum_{\boldsymbol{x} \in \Phi_{\boldsymbol{X}}} f(\boldsymbol{x}) p(\boldsymbol{x} \mid F)$ は以下の様に近似できる．

$$E[f] \simeq \sum_{k=1}^{K} w_k f(\boldsymbol{x}^{(k)}), \qquad w_k \equiv \frac{p(\boldsymbol{x}^{(k)})}{\sum_{k'=1}^{K} p(\boldsymbol{x}^{(k')})}$$

3. 提案手法

SampleSAT は論理式 F の充足解 Φ_F からの一様サンプリングを実現する．しかしながら一様分布を用いた重点サンプリングは，真の条件付き分布 $p(\boldsymbol{X} \mid F)$ が一様分布から離れるに連れ，近似精度が悪化することが知られている [Bishop 06]．この問題に対して，MCMC(Markov chain Monte Carlo, マルコフ連鎖モンテカルロ) 法に基づく近似計算は，重点サンプリングよりも少ないサンプル数でよりよい近似精度を持つことが知られている [Bishop 06]．これに対し我々は Φ_F から $p(\boldsymbol{X} \mid F)$ に従うサンプリングを実現する **CS sampling** を提案する．CS sampling は MCMC 法の一種である Slice Sampling を論理に基づく確率モデルに適用したものである．Slice Sampling ではサンプリングを行う目的分布 $p(\boldsymbol{X})$ に対し，スライスと呼ばれる確率変数 $\boldsymbol{U}$ を用いて拡張した分布 $p(\boldsymbol{X}, \boldsymbol{U})$ を考える．すると，$p(\boldsymbol{X})$ を定常分布にもつマルコフ連鎖からのサンプルリングは，変数 $\boldsymbol{X}$ を固定した上でのスライス $\boldsymbol{U}$ のサンプリングと， $\boldsymbol{U}$ を固定した上での $\boldsymbol{X}$ のサンプリングを繰り返すことによって得られる．本論文ではこの Slice Sampling を，論理に基づく確率モデルに適用する．我々の目的は条件付き確率分布 $p(\boldsymbol{X} \mid F)$ の計算の近似である．まずこの条件付き確率分布をスライス $\boldsymbol{U} = \{U_1, U_2, \ldots, U_N\}$ を用いて以下のように拡張する．

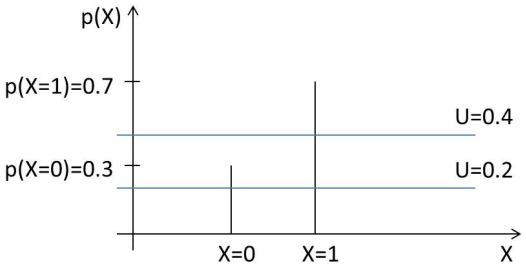

図 1: スライスの例

$$p(\boldsymbol{x}, \boldsymbol{u} \mid F) \equiv \frac{\prod_{i=1}^{N} \delta(0 \le u_i \le p(x_i))}{p(F)} F(\boldsymbol{x})$$

すると，変数 $\boldsymbol{X}$ を固定した上でのスライス $\boldsymbol{U}$ の分布と $\boldsymbol{U}$ を固定した上での $\boldsymbol{X}$ の分布はそれぞれ以下となる．

$$p(\boldsymbol{u} \mid \boldsymbol{x}, F) = \prod_{i=1}^{N} \frac{\delta(0 \le u_i \le p(x_i))}{p(x_i)}$$

$$p(\boldsymbol{x} \mid \boldsymbol{u}, F) = \frac{F(\boldsymbol{x}) \prod_{i=1}^{N} \delta(u_i \le p(x_i))}{\sum_{i=1}^{N} \delta(u_i \le p(x_i))}$$

ただし，$\delta(a)$ は a が真なら 1 を返し，そうでなければ 0 を返す関数である．上式はスライス $\boldsymbol{U}$ の条件付き確率分布は範囲 $0 \le u_i \le p(x_i)$ 上の一様分布であることを示している．更に，変数 $\boldsymbol{X}$ の条件付き確率分布は以下の論理式 $F_{\boldsymbol{U}}$ の充足解 $\Phi_{F_{\boldsymbol{U}}}$ 上の一様分布であることも示している．

$$F_{\boldsymbol{U}} = F \wedge \bigwedge_{x_i \in \{x_i \mid p(1-x_i) \le u_i \le p(x_i)\}} \text{"}X_i = x_i\text{"}$$

よって $\boldsymbol{X}$ のサンプリングに SampleSAT を利用することが可能である．

X をサンプルする際には，与えられた U の値による条件 $u \le p(x)$ を満たす範囲からサンプルする．図 1 の $U = 0.4$ の場合，$p(X = 0) \le 0.4 \le p(X = 1)$ であるため，$X = 0$ は割り当ての候補から外れ，サンプルにおける X の値の候補は $X = 1$ のみである．そこで $X = 1$ を論理式により表現とした "$X = 1$" を F に追加することによりサンプルが充足するための条件とすることができる．逆に図 1 の $U = 0.2$ の場合には X の値を固定することができないので，論理式の追加は起こらない．このようにして確率変数の値が論理式 F の中でいくつか固定される．値が固定された論理式に対しては単位伝搬 (unit propagation) と呼ばれる SAT 問題における高速化手法が知られており，これを用いて節に含まれる固定した値を各節に反映させることにより論理式を簡単化することができる．このように簡単化した論理式を解くことにより充足解を得る時間を短縮できる．以上が提案手法の概略である．表 1 に CS sampling アルゴリズムを示す．

CS sampling(F, T)
F : CNF, T : a number of samples
begin
$\boldsymbol{x}^{(1)} \sim$ SampleSAT(ψ)
for $t = 1$ **to** $T - 1$ **do**
$\forall i, u_i^{(t)} \sim p(\boldsymbol{u} \mid \boldsymbol{x}, F)$
$F_{\boldsymbol{U}} = F \wedge \bigwedge_{x_i \in \{x_i \mid p(1-x_i) \leq u_i \leq p(x_i)\}} "X_i = x_i"$
$\boldsymbol{x}^{(t+1)} \sim p(\boldsymbol{x} \mid \boldsymbol{u}, F)$
return $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(T)}$
end

表 1: CS sampling アルゴリズム

$\forall x, Smokes(x) \Rightarrow Cancer(x)$: 喫煙者は癌になる．
$\forall x, y, Friends(x, y) \Rightarrow (Smokes(x) \Leftrightarrow Smokes(y))$:
友人同士であれば喫煙者同士であるか，非喫煙者同士である．

表 2: 適用したルール

	Total times(secs)	Average steps
SampleSAT	960.00	33.9887
CS sampling	126.05	0.00079

表 3: 実行時間，平均ステップ数の比較

CS Sampling は SampleSAT と比較して速度と精度において優れていると予想される．速度に関しては充足解を得る時間が減少すると期待される．精度に関しては一様分布を用いた重点サンプリングは目的分布と一様分布とがかけ離れている場合には精度が低いことが知られているため，条件付き分布からサンプリングを行う CS sampling はサンプル数が比較的少ない場合でもよい精度が期待される．

4. 実験

ここでは CS sampling の性能を実験的に確かめる．

4.1 問題設定

本論文ではサンプリングの速度を測る実験とサンプリングによる期待値計算の精度を測る実験を行う．まず，サンプリングを行う分布を定義する．本論文において論理式により条件付けられた分布を定める確率モデルとして CBPM(Constraint-based Probabilistic Models)[Sato 10] を用いる．CBPM は論理に基づく確率モデルであり，論理式 F は KB(knowledge base, 知識ベース) で表現する．今，KB を表 2 の喫煙者の友人関係についての規則を複数人の人間に対して適用した論理式とし，その規則を CNF 式で記述する．KB に含まれる論理式の数は人間の数により変化する．この時，互いに独立な確率的命題変数集合 $\boldsymbol{X} = \{X_1, \ldots, X_N\}$ 上の確率分布 $p(\boldsymbol{X})$ を考えると，我々がサンプリングを行う分布は $p(\boldsymbol{X} \mid KB)$ である．この条件付き確率分布 $p(\boldsymbol{X} \mid KB)$ からのサンプリングを CS sampling を用いて行う．一つ目の実験は CS sampling と SampleSAT でそれぞれサンプリングを行い，その時間と充足解を見つけるためのステップ数の平均を比較する．二つ目の実験はサンプリングを行い，そのサンプルを用いて各命題変数の周辺確率 $p(x_i \mid KB)$ を期待値を用いて近似計算する．近似計算については後述する．

4.2 サンプリング時間の比較

CS sampling と SampleSAT に同じ制約を与え，同数のサンプルを得る時間と充足解を求めるための平均ステップ数がどのように変化するかを調べる．KB は表 2 のルールを 14 人の人間に適用して作成した．基底アトムの数は 224，基底節の数は 406 であり，KB はこれらの基底節の連言である．この KB を満たす充足解を 100,000 回サンプリングし，その際の時間と充足解を得るための平均ステップ数を求めた．それぞれの結果は表 3 のようになった．実行時間については提案手法が約 8 倍早くなり，充足解を得る平均ステップ数は提案手法において大幅に減少している．これは CS sampling のスライスによる変数値の固定や単位伝搬の効果によって多くの確率変数の値が決定した効果によるものであり，その結果充足解を得ることが非常に簡単になったと考えられる．

4.3 サンプリング精度の比較

次にサンプリングによる近似精度を比較する．CS sampling と SampleSAT それぞれによって得られたサンプル列 $\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(K)}\}$ を用いて，各命題変数の周辺確率 $p(x_i \mid KB)$ の近似を期待値計算を用いて行う．CS sampling のサンプリング近似は以下の式により行う．

$$p(x_i \mid KB) \simeq \frac{1}{K} \sum_{k=1}^{K} \delta(x_i^{(k)} = x_i)$$

ここで $x_i^{(k)}$ は k 番目のサンプル $\boldsymbol{x}^{(k)}$ の第 i 要素である．また，SampleSAT の重点サンプリングは以下の式により行う．

$$\begin{aligned} p(x_i \mid KB) &\simeq \sum_{k=1}^{K} w_k \delta(x_i^{(k)} = x_i), \\ w_k &= \frac{p(\boldsymbol{x}^{(k)})}{\sum_{k'} p(\boldsymbol{x}^{(k')})} \end{aligned}$$

上式を用いて各命題変数の周辺確率を近似計算し，厳密計算による周辺確率と比較する．比較には KL ダイバージェンスを用いた．各命題変数ごとに真の周辺確率との KL ダイバージェンスを計算し，それらの和を計算することで近似性能を測った．近似精度の比較を行うために厳密計算による周辺確率を計算する必要があるため，表 2 のルールを 3 人の人間に適用して KB を作成した．変数は 15，節の数は 21 である．サンプル数を変化させ，KL ダイバージェンスの総和をプロットした結果を図 2 に示す．この結果から，CS sampling を用いた結果はサンプルが少ない場面でも SampleSAT と比較してよい精度であることを示している．サンプル数が多くなるとどちらのサンプリング手法も同程度の近似精度を示している．

この結果についての理由を考察する．サンプル数が少なくサンプリングを行いたい分布と一様分布が大きく異なる場合には重点サンプリングは精度がよくないことが知られている [Bishop 06]．そこで各モデルに対するサンプル数を数えたところ，図 4，図 5 のようなヒストグラムを得た．図 4 は SampleSAT のヒストグラム，図 5 は CS sampling のヒストグラムである．図 4，図 5 は横軸は充足解の確率が高い順に並べてある．このヒストグラムから SampleSAT は充足解からほぼ一様にサンプリングを行っていることがわかる．図 3 にサンプリングを行った真の分布を示す．図 3, 図 5 を見比べてみると，真の分布と CS sampling のヒストグラムがよく似ており，ので CS sampling は真の分布からのサンプリングを実現してい

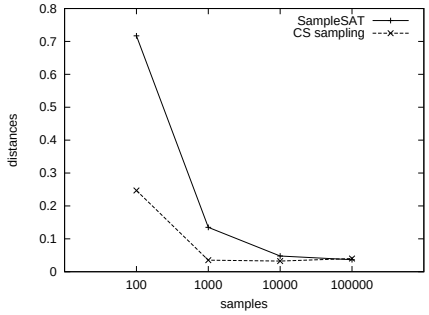

図 2: 周辺確率の KL ダイバージェンスの総和

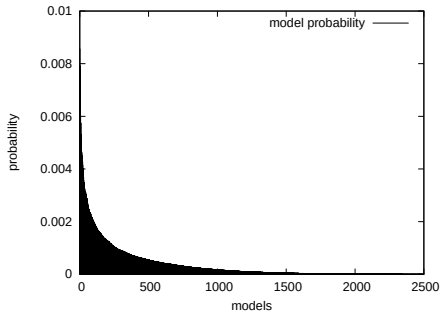

図 3: サンプリングを行った分布

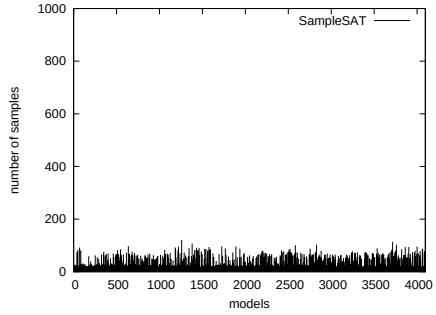

図 4: SampleSAT で得られたサンプルによるヒストグラム

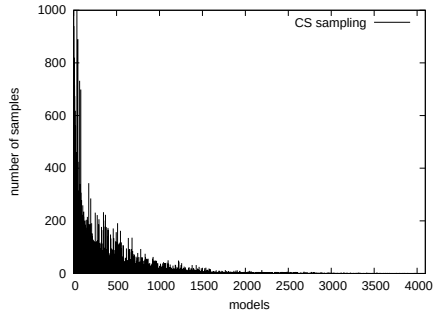

図 5: CS sampling で得られたサンプルによるヒストグラム

る．従って少ないサンプルでもよい近似精度が得られたと考えられる．

5. まとめと今後の課題

本論文では確率分布に論理に基づく制約を与え，その条件付き分布からサンプリングを行う CS sampling を提案した．従来手法よりも充足解の探索が効率的であり，かつ論理式の充足解を条件付き分布からサンプリングできる方法であることを実験的に示した．今後の課題としては確率学習を可能にすることや，条件付き確率の近似による様々な推論への適用などが考えられる．例えば，統計的アブダクション [Sato 10] など論理推論の大規模な問題への適用を試みたい．

参考文献

[Bishop 06] Bishop, C. M.: *Pattern Recognition and Machine Learning*, Springer (2006)

[Neal 03] Neal, R.: Slice Sampling, *The Annals of Statistics*, Vol. 31, No. 3, pp. 705–767 (2003)

[Poon 07] Poon, H. and Domingos, P.: Sound and Efficient Inference with Probabilistic and Deterministic Dependencies, *AAAI-07* (2007)

[Richardson 06] Richardson, M. and Domingos, P.: Markov logic networks, *Machine Learning*, Vol. 62, pp. 107–136 (2006)

[Sato 10] Sato, T., Ishihata, M., and Inoue, K.: *Constraint-based probabilistic modeling for statistical abduction*, shpringer (2010)

[Selman 92] Selman, B., Levesque, H. J., and Mitchell, G., D: A new method for solving hard satisfiability problems, *AAAI-92*, pp. 440–446 (1992)

[Selman 96] Selman, B., Kautz, H., and Cohen, B.: Local Search Strategies for Satisfiability Testing, *Second DIMACS Challenge on Cliques, Coloring and Satisfiability* (1996)

[Wei 04] Wei, W., Erenrich, J., and Selman, B.: Towards Efficient Sampling: Exploiting Random Walk Strategies, *AAAI-04*, pp. 670–676 (2004)