マルチレベルモンテカルロ法を用いた変分推論における分散抑制法についての研究

47-186121：藤澤 将広
杉山・佐藤・本多研究室
修士論文発表会

2020 年 2 月 5 日

1 背景・概要

モンテカルロ変分推論 [1] は，大規模なデータ・複雑な確率モデルに対する推論方法として，近年重要視されている．しかし，モンテカルロ変分推論の性能は，確率的に推定した勾配の分散に大きく依存することが知られており，より複雑な確率モデルの社会応用における大きな障害となっている．この問題に対し，制御変量法や重点サンプリングなどといった様々な分散抑制方法が提案されてきた．これらの手法の欠点は，分散を抑制するための新しい関数などを人為的に付与しなくてはならず，さらには「推論の収束性」,「分散抑制の程度」,「分散抑制された確率的勾配の質」についての理論保証がない，あるいは享受できないことである．そこで，本研究では，マルチレベルモンテカルロ法を用い，最適化の行程で自然に入手できる過去のパラメータ情報を再利用する，新しい分散抑制法を提案する．提案法は，モンテカルロ変分推論において標準的に使用されつつある再パラメータ化勾配のもとで自然に導出される．また，確率的勾配の分散と計算コストの比によって推定に用いる確率変数の数を動的に推定する，新しい確率的勾配推定法および最適化アルゴリズムを提案する．さらに，確率的勾配降下法における提案法の勾配の収束性，勾配の分散の大きさ，および SN 比に基づく勾配推定量の質について，理論解析を行う．最後に，いくつかの数値実験を行い，基準となる既存法と比較し，提案法の性能を評価する．

2 モンテカルロ変分推論

変分推論では，変分分布 $q(\mathbf{z}|\lambda)$ ($\lambda \in \mathbb{R}^d$は変分パラメータ) とモデル $p(\mathbf{x},\mathbf{z})$ との自由エネルギー：

$$\mathcal{L}(\lambda) = \mathbb{E}_{q(\mathbf{z}|\lambda)}[\log q(\mathbf{z}|\lambda) - \log p(\mathbf{x},\mathbf{z})] \tag{1}$$

の最小化問題を考える．これは，変分分布と求めたい事後分布 $p(\mathbf{z}|\mathbf{x})$ との間のカルバック・ライブラー距離を最小にすることと等価である．推論の際は，式 (1) の勾配および最適化の枠組みを用いる．

しかし，式 (1) の勾配はスコア関数 $\nabla_\lambda \log q(\mathbf{z}|\lambda)$ を含み，かつ $q(\mathbf{z}|\lambda)$ についての期待値で表されることから，変分パラメータ λ の挙動に直接依存し，推論がしばしば不安定になる．この問題の主要な対処法の一つに再パラメータ化勾配を用いる方法がある．

2.1 再パラメータ化勾配

再パラメータ化 [2] とは，変分分布で表される確率変数 $\mathbf{z}$ を，パラメータを持たない簡易な分布から得られる確率変数 $\epsilon \sim p(\epsilon)$ を用いて表すことをいう．すなわち，$\mathbf{z}$ は，ある変換 $\mathcal{T}$ を用いて，$\mathbf{z} = \mathcal{T}(\epsilon;\lambda)$ と表す．これにより，自由エネルギーの再パラメータ化勾配は，変分分布 $q(\mathbf{z}|\lambda)$ に依らず，$p(\epsilon)$ についての期待値で表せる：

$$\nabla_\lambda \mathcal{L}(\lambda) = \mathbb{E}_{p(\epsilon)}[\nabla_\lambda \log q(\mathcal{T}(\epsilon;\lambda)|\lambda) - \nabla_\lambda \log p(\mathbf{x},\mathcal{T}(\epsilon;\lambda))] \tag{2}$$

よって，変分パラメータ λ の挙動の影響を抑えることができる．

2.2 モンテカルロ変分推論

大規模なデータや複雑なモデルのもとで推論を行う場合，式 (1) および式 (2) の勾配の解析的な計算は難しい，または不可能になる．そこで，勾配を確率的に推定することを考える．式 (2) は，

$$g_\lambda(\epsilon) = \nabla_\lambda \log q(\mathcal{T}(\epsilon;\lambda)|\lambda) - \nabla_\lambda \log p(\mathbf{x},\mathcal{T}(\epsilon;\lambda)), \tag{3}$$

とすると，$\nabla_\lambda \mathcal{L}(\lambda) = \mathbb{E}_{p(\epsilon)}[g_\lambda(\epsilon)]$ と表せ，これをモンテカルロ法によってサンプリングした有限個の確率変数 $\{\epsilon_1, \epsilon_2, \ldots, \epsilon_N\}$ を用いて近似する．このとき，ある $t \in \mathbb{N}$ 回目の最適化における $\mathbb{E}_{p(\epsilon)}[g_\lambda(\epsilon)]$ の不偏推定量は，

$$\widehat{\nabla}_{\lambda_t} \mathcal{L}(\lambda) = \frac{1}{N}\sum_{n=1}^{N} g_{\lambda_t}(\epsilon_n), \tag{4}$$

と表せる．式 (4) を用いて，自由エネルギーは，確率的勾配降下法 [3] などの確率的最適化の枠組みで最適化できる．

しかし，モンテカルロ法による勾配の推定は，分散が大きくなり，確率的最適化の収束が遅くなることが問題となっている．

3 マルチレベルモンテカルロ変分推論

3.1 マルチレベルモンテカルロ法

関数 P_L ($L \in \mathbb{N}$) を近似する関数列 $P_0, P_1, \ldots, P_{L-1}$ が得られたとき，P_L の期待値 $\mathbb{E}[P_L]$ の不偏推定量は，期待値の線形性より，

$$\mathbb{E}[P_L] \approx N_0^{-1}\sum_{n=1}^{N_0} P_0^{(0,n)} + \sum_{l=1}^{L}\left\{ N_l^{-1}\sum_{n=1}^{N_l}(P_l^{(l,n)} - P_{l-1}^{(l,n)})\right\}, \tag{5}$$

と表せる．ここで，(l,n) は各レベルで独立な N_l 個の確率変数を用いることを示す．マルチレベルモンテカルロ法の推定精度は，レベル数 L が増えるほど良くなる [4].

3.2 マルチレベルモンテカルロ変分推論

マルチレベルモンテカルロ法を，式 (2) に適用することを考える．ある $T \in \mathbb{N}$ ($t = 1, \ldots, T$) 回目の最適化において，再パラメータ化勾配は，期待値の線形性を用いて以下のように表せる：

$$\begin{aligned}&\nabla_{\lambda_T}^{\mathrm{MRG}} \mathcal{L}(\lambda)\\ &= \mathbb{E}_{p(\epsilon)}[g_{\lambda_0}(\epsilon)] + \sum_{t=1}^{T}\left(\mathbb{E}_{p(\epsilon)}[g_{\lambda_t}(\epsilon) - g_{\lambda_{t-1}}(\epsilon)]\right).\end{aligned} \tag{6}$$

以後，これをマルチレベル再パラメータ化勾配 (Multi-level reparameterized gradient : MRG) と呼ぶ．これに不偏推定量を構成すると，

$$\begin{aligned}\widehat{\nabla}_{\lambda_T}^{\mathrm{MRG}} \mathcal{L}(\lambda) &= N_0^{-1}\sum_{n=1}^{N_0} g_{\lambda_0}(\epsilon_{(n,0)})\\ &\quad + \sum_{t=1}^{T}\left(N_t^{-1}\sum_{n=1}^{N_t}[g_{\lambda_t}(\epsilon_{(n,t)}) - g_{\lambda_{t-1}}(\epsilon_{(n,t)})]\right),\end{aligned} \tag{7}$$

と表せる．よって，各レベルの 1 サンプルあたりの分散と計算コストを $\mathbb{V}_t, C_t$ ($t = 1, 2, \ldots, T$) とすると，MRG 推定量の分散と計算コストは，$\sum_{t=1}^{T} N_t^{-1}\mathbb{V}_t$, $\sum_{t=1}^{T} N_t C_t$ となる．

3.3 アルゴリズムの導出

本節では，MRG 推定量の分散を最小にするような最適なサンプル数 N_t およびそれを基にしたアルゴリズム導出を行う．

まず，最適なサンプル数 N_t について，以下の定理が成り立つ．

定理 1 (分散を最小にする最適な N_t)**.** $\mathbb{V}_0, C_0 = c$ を 1 サンプルあたりの $g_{\lambda_0}(\epsilon)$ の分散と計算コスト，$\mathbb{V}_t, C_t = 2c$ を 1 サンプルあたりの $g_{\lambda_t}(\epsilon) - g_{\lambda_{t-1}}(\epsilon)$ の分散と計算コストとする．ここで，c は正の定数である．このとき，$\widehat{\nabla}_{\lambda_T}\mathcal{L}_{\mathrm{MRG}}(\lambda)$ の分散 $\sum_{t=1}^{T} N_t^{-1}\mathbb{V}_t$ を最小にするサンプル数 N_t は，

$$N_t = \begin{cases} \sqrt{\frac{\mathbb{V}_1}{2\mathbb{V}_0}} N_0 & (t=1), \\ \sqrt{\frac{\mathbb{V}_t}{\mathbb{V}_{t-1}}} N_{t-1} & (t=2,3,\ldots,T), \end{cases} \tag{8}$$

と表せる．

以上の結果より，最適な N_t は現在・過去の勾配の分散の比によって求めることができる．最適化の結果分散が小さくなれば，サンプル数は減少し，大きくなれば増加することがわかる．

また，MRG 推定量は，この定式化のまま使用すると，最適化が進行するにつれて計算コストが著しく大きくなってしまう．これを避けるために，以下の補題を示した．

補題 1 (パラメータ更新則)．*MRG* 推定量は，$t \leq 1$ のとき，

$$\begin{aligned} &\widehat{\nabla}_{\lambda_t} \mathcal{L}_{\mathrm{MRG}}(\lambda) \\ &= \widehat{\nabla}_{\lambda_{t-1}} \mathcal{L}_{\mathrm{MRG}}(\lambda) + N_t^{-1} \sum_{n=1}^{N_t} [g_{\lambda_t}(\epsilon_{(n,t)}) - g_{\lambda_{t-1}}(\epsilon_{(n,t)})], \end{aligned} \tag{9}$$

と表せ，学習率減少付き確率的勾配降下法におけるパラメータ更新則は，

$$\begin{aligned} \lambda_{t+1} = \lambda_t &+ \frac{\eta_t}{\eta_{t-1}} (\lambda_t - \lambda_{t-1}) \\ &- \alpha_t N_t^{-1} \sum_{n=1}^{N_t} \left[g_{\lambda_t}(\epsilon_{(n,t)}) - g_{\lambda_{t-1}}(\epsilon_{(n,t)}) \right], \end{aligned} \tag{10}$$

と表せる．

これらの定理，補題を基に提案法の推論アルゴリズムを構成する．アルゴリズムの概要は口頭発表にて紹介する．

4 理論解析・理論保証

まず，提案した勾配の分散抑制性能を確認するために，いくつかの一般的な仮定のもとで，1 サンプルあたりの分散 $\mathbb{V}_t$ の上界

$$\mathbb{E}_{p(\epsilon)} \left[\| g_{\lambda_t}(\epsilon) - g_{\lambda_{t-1}}(\epsilon) \|_2^2 \right] \leq \alpha_{t-1}^2 N^{-1} K_2 (C_1 + d\delta C_2),$$

を示し，その大きさが $\mathcal{O}(\alpha_{t-1}^2)$ で表されることを示した．ここで，C_1, C_2, δ は正の定数，d はパラメータ空間の次元数である．この結果より，$\mathbb{V}_t$ は $\mathbb{V}_t \overset{t\to\infty}{\to} 0$ となることがわかり，サンプル数 N_t は最適化が進むにつれて減少することが示される．

さらに，最適化の収束性の解析のために，提案法における勾配の l_2 ノルムの学習率重み付き平均が，以下の上界

$$\begin{aligned} &\frac{1}{A_T} \sum_{t=1}^{T} \alpha_t \mathbb{E}[||\nabla_{\lambda_t} \mathcal{L}(\lambda_t)||_2^2] \\ &\leq G_T + \frac{\alpha_0^2 K_1}{2A_T} \sum_{t=1}^{T} \eta_t^2 \left(\mathbb{E}\left[||\nabla_{\lambda_t} \mathcal{L}(\lambda_t)||_2^2 \right] + \frac{\kappa}{N_t} \eta_{t-1}^2 \right). \end{aligned}$$

で抑えられることを示した．ここで，$A_T = \sum_{t=1}^{T} \alpha_t$，$G_T = \frac{1}{A_T}[\mathcal{L}(\lambda_1) - \mathcal{L}(\lambda^*)]$ であり，α_t がロビンス-モンロー条件：$\sum_{t=1}^{\infty} \alpha_t = \infty$，$\sum_{t=1}^{\infty} \alpha_t^2 < \infty$ を満たすとき，$T \to \infty$ で，この重み付き平均は 0 に収束する．すなわち，(局所) 最適解 λ^* に収束する．既存手法もこれと同様にして解析を行った．結果，提案法は，サンプル数を増やす以外に，学習率減衰関数が収束性を向上させることが明らかとなった．

また，提案法の勾配推定量の質は，SN 比

$$\mathrm{SNR}(\lambda) = \frac{\| \mathbb{E}_{p(\epsilon_{1:N})}[\hat{g}_\lambda(\epsilon_{1:N})] \|_2^2}{\sqrt{\mathbb{V}[\hat{g}_\lambda(\epsilon_{1:N})]}},$$

を基準にして，

$$\mathrm{SNR}(\lambda_t) \geq \frac{\|\nabla_{\lambda_t} \mathcal{L}(\lambda_t)\|_2^2}{\sqrt{\kappa}} \cdot \frac{\sqrt{N_t}}{\eta_{t-1}},$$

と評価できることを示した．これは，提案法が，勾配推定量の質を学習率の減衰関数によって制御することができることを表す．

5 実験

提案法の理論的性質，および性能の確認のために，実データを用いた複数の計算機実験を行った．ここでは，ベイジアンロジスティック回帰による 2 値分類実験 (図 1，1 行目) と，50 の隠れ層・活性化関数に ReLU 関数を設定したベイジアンニューラルネットワークによる回帰実験 (図 1，2 行目) の結果を報告する．比較手法は，モンテカルロ法による手法とランダム化した準モンテカルロ法による手法 [5] を用いた．また，学習率の減衰関数は，ステップ減衰関数：$\eta_t = \alpha_0 \cdot \beta^{\lfloor t/r \rfloor}$ を用い，$\beta = 0.5$，$r = 100$ とした．本実験における，テストデータに対する尤度と更新項の分散についての結果を以下に示す．この実験を通して，提案法が，既存法に比べて，推論の収束が高速で，モデルによっては最適な解を得，勾配の分散を抑制できていることがわかった．

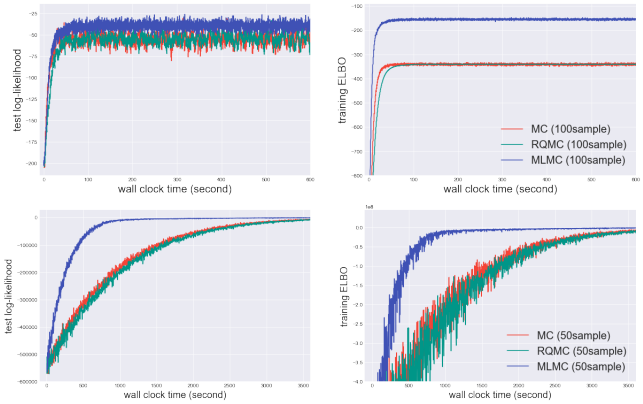

図 1　テストデータに対する尤度・学習曲線 (自由エネルギー)．

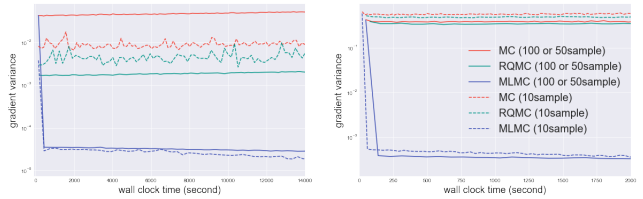

図 2　提案法の勾配推定量の分散

6 結論

本研究では，マルチレベルモンテカルロ法を，モンテカルロ変分推論に適用可能な形に拡張し，過去の勾配情報とパラメータ情報を用いて推定量の分散を抑制する方法を提案した．さらに，分散の大きさ，最適化における収束性，および SN 比を基準にした勾配推定量の質についての理論解析を行った．これにより，提案法が，最適化の収束性および勾配推定量の SN 比を学習率の減衰関数を用いて制御可能であることを示した．最後に，実データにおける計算機実験を通して，提案法の実用性を確認した．提案法は純粋なモンテカルロサンプリングを基に構築されているため，近年提案されてきた様々な分散抑制法と組み合わせて使用できる可能性がある．理論解析では，変分分布の種類を仮定して議論を進めた．今後は，より一般的な変分分布に提案法を拡張し，その基で理論解析を行う予定である．

参考文献

[1] Rajesh Ranganath, Sean Gerrish, and David Blei. Black box variational inference. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 814–822, 2014.

[2] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014.

[3] H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407, 1951.

[4] Michael B. Giles. Multi-level monte carlo path simulation. *Operations Research*, 56(3):607–617, 2008.

[5] Alexander Buchholz, Florian Wenzel, and Stephan Mandt. Quasi-Monte Carlo variational inference. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 668–677, 2018.