

Title	Annealed stein variational policy gradient
Sub Title	
Author	高木, 智也(Kobayashi, Kei) 小林, 景
Publisher	慶應義塾大学AI・ 高度プログラミングコンソーシアム
Publication year	2024
Jtitle	AICカンファレンス予稿集 (2024.) ,p.15- 16
JaLC DOI	
Abstract	Stein Variational Policy Gradient (SVPG) は、 ノンパラメトリックな変分推論アルゴリズムであるStein VariationalGradient Descentを方策勾配法に応用したものである. SVPGでは温度パラメータが定数として扱われ、 学習段階における探索と活用のバランス調整は行われていなかった. 本研究ではAnnealed Stein Variational Policy Gradient (ASVPG) を提案する. ASVPGでは温度パラメータに焼きなまし法を適用することにより、 学習効率と収束性の向上が実験的に示された.
Notes	会議名：AICカンファレンス2024 開催地：慶應義塾大学協生館AICラウンジ、および多目的教室 日時：2024年3月7日 ポスター発表要旨
Genre	Conference Paper
URL	https://koara.lib.keio.ac.jp/xoonips/modules/xoonips/detail.php?koara_id=KO11003001-20240307-0015

慶應義塾大学学術情報リポジトリ(KOARA)に掲載されているコンテンツの著作権は、それぞれの著作者、学会または出版社/発行者に帰属し、その権利は著作権法によって保護されています。引用にあたっては、著作権法を遵守してご利用ください。

The copyrights of content available on the KeiO Associated Repository of Academic resources (KOARA) belong to the respective authors, academic societies, or publishers/issuers, and these rights are protected by the Japanese Copyright Act. When quoting the content, please follow the Japanese copyright act.

Annealed Stein Variational Policy Gradient

高木智也¹, 小林景²

¹ 慶應義塾大学理工学部数理科学科, ² 慶應義塾大学大学院理工学研究科基礎理工学専攻

Abstract: Stein Variational Policy Gradient (SVPG) は、ノンパラメトリックな変分推論アルゴリズムである Stein Variational Gradient Descent を方策勾配法に応用したものである。SVPG では温度パラメータが定数として扱われ、学習段階における探索と活用のバランス調整は行われていなかった。本研究では Annealed Stein Variational Policy Gradient (ASVPG) を提案する。ASVPG では温度パラメータに焼きなまし法を適用することにより、学習効率と収束性の向上が実験的に示された。

Keywords: Variational Inference, Stein Variational Gradient Descent, Reinforcement Learning, Policy Gradient Method, Annealing Method

1. 研究背景・目的

方策勾配法は強化学習における主要な手法であり、多くの応用例が存在する。Liu et al. (2017) は方策勾配法をベイズ推論の枠組みで再解釈し、変分推論の一手法である Stein Variational Gradient Descent (SVGD)[1] を方策勾配法に適用した Stein Variational Policy Gradient (SVPG)[2] を提案した。SVPG は方策勾配法における平均収益と効率性を向上させたが、提案論文では温度パラメータを定数として扱い、学習過程での探索と活用のバランス調整を行っていなかった。本研究では、温度パラメータに焼きなまし法を適用することで、学習の効率性と収束性が向上する可能性があることを示す。

2. 方法

SVPG ではニューラルネットワークによる方策のパラメータ θ がある分布 q に従うと仮定し、以下の目的関数 $\hat{J}(\theta)$ を最大化するようにパラメータを更新していた:

$$\hat{J}(\theta) = \mathbb{E}_{\theta \sim q}[J(\theta)] - \alpha D_{\text{KL}}(q||q_0).$$

ただし、 q_0 は事前分布、 $\alpha \in \mathbf{R}_{>0}$ である。

本研究の新規性は、温度パラメータ α に焼きなまし法を適用し、学習の効率性と収束性が向上する可能性があることを示した点にある。

なお、方策勾配法のアルゴリズムとして REINFORCE を、環境として OpenAI Gym が提供する Cartpole-v1 を用いた。

3. 結果

Table 1 より、REINFORCE-ASVPG が他のアルゴリズムに比べて最大報酬を獲得するまでに必要な平均エピソード数が少ないことがわかり、Fig. 1 より、獲得した報酬の平均が大きいことがわかる。これらより、ASVPG が学習効率性において優れていることが示唆される。

さらに、Table 2 より、最大報酬を連続して達成するまでの平均エピソード数も、ASVPG が他のアルゴリズムより少ないことから、ASVPG が収束性においても優れていることが示唆される。

Table 1 α の値と最大報酬を得るまでに必要な平均エピソード数と標準誤差。

アルゴリズム	α	平均エピソード数
REINFORCE-ASVPG	$1/\sqrt{l}$	120.0 (3.66)
	$1/l$	110.2 (4.10)
	$1/l^2$	114.8 (2.86)
	$1/l^3$	118.5 (2.93)
REINFORCE-SVPG	$10^{-8} \simeq 0$	130.4 (3.59)
	10^{-4}	138.4 (3.66)
	1.0	139.9 (3.75)
	10.0	143.7 (5.05)
	100.0	182.6 (4.02)
REINFORCE	None	190.7 (11.09)

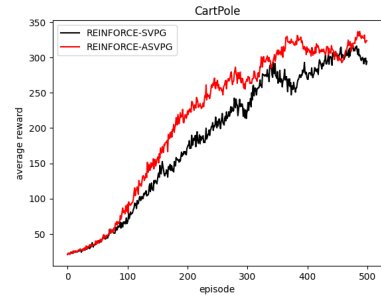


Fig. 1 x 軸はエピソード番号、 y 軸は獲得した報酬の平均を表す。REINFORCE-ASVPG では $\alpha = 1/l$ 、REINFORCE-SVPG では $\alpha = 10.0$ を用いた。

Table 2 最大報酬が 475 以上を 10 連続達成するまでに必要な平均エピソード数と標準誤差。

アルゴリズム	α	平均エピソード数
REINFORCE-ASVPG	$1/l$	196.9 (4.94)
REINFORCE-SVPG	10.0	258.2 (6.82)
REINFORCE	None	310.6 (18.08)

4. 考察

REINFORCE-ASVPG が他のアルゴリズムに比べて最大報酬を獲得するまでに必要な平均エピソード数が少ないことと、獲得した報酬の平均が大きいことから、ASVPG が学習効率性において優れていることが示された。これは ASVPG における温度パラメータへの焼きなまし法の適用が、効果的な探索と活用のバランスを実現していることを示唆している。

さらに、最大報酬を連続して達成するまでの平均エピソード数も、ASVPG が他のアルゴリズムより少ないことから、

ASVPG が収束性においても優れていることが明らかになった。これは、ASVPG が学習中のパラメータの動きに柔軟に適応でき、最適な方策に対してより迅速に収束する能力を持っていることを示している。

これらの結果は、ASVPG の理論と矛盾しない。温度パラメータの適応的な調整が、学習過程における探索と活用との適切なトレードオフを可能にし、結果として効率的かつ収束性の高い学習を実現している。

5. 結論と今後の展望

REINFORCE 以外の方策勾配法に ASVPG を適用することは有用である。ASVPG をさまざまな強化学習の課題に適用し、その汎用性を確認することも重要である。本研究では比較的単純な焼きなまし法を採用したが、より複雑なアプローチを採用することで、さらなる改良ができる可能性がある。また、SGVD におけるカーネルの選択は性能に大きな影響を与えることが知られている [3]。ASVPG においても異なるカーネルを試すことで、さらなる性能改善が期待できる。

参考文献

- [1] Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. *Advances in neural information processing systems*, Vol. 29, , 2016.
- [2] Yang Liu, Prajit Ramachandran, Qiang Liu, and Jian Peng. Stein variational policy gradient. *arXiv preprint arXiv:1704.02399*, 2017.
- [3] Dilin Wang, Ziyang Tang, Chandrajit Bajaj, and Qiang Liu. Stein variational gradient descent with matrix-valued kernels. *Advances in neural information processing systems*, Vol. 32, , 2019.