論 文

Technical Papers

報酬の分散を推定するTDアルゴリズムとMean-Variance強化学習法の提案

TD Algorithm for the Variance of Return and Mean-Variance Reinforcement Learning

佐藤 誠 Makoto Sato
東京工業大学 総合理工学研究科
Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology
satom@fe.dis.titech.ac.jp

木村 元 Hajime Kimura
（同　上）
gen@fe.dis.titech.ac.jp

小林 重信 Shibenobu Kobayashi
（同　上）
kobayashi@dis.titech.ac.jp

keywords: reinforcement learning, Markov decision processes, variance penalized criteria, gradient-based learning, machine mainenance problem, TD-method

Summary

Estimating probability distributions on returns provides various sophisticated decision making schemes for control problems in Markov environments, including risk-sensitive control, efficient exploration of environments and so on. Many reinforcement learning algorithms, however, have simply relied on the expected return. This paper provides a scheme of decision making using mean and variance of return-distributions. This paper presents a TD algorithm for estimating the variance of return in MDP(Markov decision processes) environments and a gradient-based reinforcement learning algorithm on the variance penalized criterion, which is a typical criterion in risk-avoiding control. Empirical results demonstrate behaviors of the algorithms and validates of the criterion for risk-avoiding sequential decision tasks.

1. はじめに

近年の強化学習における最も重要な進歩は動的計画法 (DP) の理論と統合されたことである．その結果，政策の最適性，関数近似と組み合わせた場合の収束先，観測の不完全性のモデル化など実問題を扱う上で必要となる問題を理論的に扱うことが可能となった [Sutton 98, Tsitsiklis 96]．DP 理論に基づく強化学習の枠組では，行動決定の主体であるエージェントはマルコフ決定過程 (MDPs) などの確率モデルで表現された環境中を遷移し，その結果得た経験を利用して政策の学習を行う．政策の善し悪しを決める政策評価規範 (policy criteria) は，定義された利得 (return) に基づき与えられる．そして，各状態の各行動について，定義された政策評価規範に関連するなんらかの価値を推定し政策学習に利用する．例えば，代表的な強化学習法である Q-learning [Watkins 89] は，無限期間の総割引報酬を利得として定義し，初期状態からある政策に従い状態遷移した場合の期待利得を政策評価規範としている．そして各状態の各行動について Q-value と呼ばれる行動価値 (action value) を推定し政策学習に利用する．そこでこの枠組において重要となるのは，行動価値を如何に効率的に推定するかという問題である．Temporal Difference(TD) 法 [Sutton 88] は行動価値を推定する代表的な方法である．TD 法は環境モデル用いず，行動価値をその状態からの遷移先の状態の価値と直接報酬のサンプルのみからを推定する (bootstrap) する単純な推定方式であり，1 ステップあたりの時間計算量と環境モデルパラメータを保持する空間計算量に関して効率的であるといえる。そのため，ニューラルネット，線形回帰などさまざまな関数近似と組み合わせることが可能であり，TD 法を用いることにより，強化学習を状態数が非常に多い実問題へ適用することが可能となった．実際，強化学習の応用研究の多くは TD 法を関数近似と組み合わせて利用している [Crites 96, Zhang 96, Neuneier 97, Singh 97, Brown 98].

ところで，これまで提案された多くの強化学習手法は政策評価規範として何らかの利得の期待値を採用しているが，これらは常に有効とは限らない．一般にマルコフ決定過程において得られる利得は確率変数であり，何らかの確率分布に従い発生する．ところが，利得の期待値が同じだが異なる分布をした 2 つの政策を，従来の強化

学習手法は同じ価値の政策と見なしてしまう．例えば，大きな利得と同じ大きさの損失 (負の利得) が 50%ずつの確率で生じる政策と，全く利得が得られない政策を同じ価値と見なす．現実世界では，期待値のみに基づいた評価方法ではなく，それぞれの政策に従った結果生じる利得の確率分布の情報を最大限に利用し，より洗練された行動選択を行う政策が求められている．そこで，金融工学や OR の分野では，利得の期待値だけでなく分布も考慮した政策評価規範をはじめ，さまざまな規範に関する研究が行われている [Fernandez 95, Marcus 97, Kadota 98]. 特に金融工学が扱う問題領域では，利得の期待値は最適値よりも劣るが大きな損失を被る可能性が低い政策が最も高く評価されることが多い．

利得のリスクを考慮した政策という観点から従来研究を概観する．Q-learning や Average Reward Learning (ARL) [Mahadevan 96, Schwartz 93] の政策評価規範は利得の期待値に基づいているため，利得のリスクを考慮した政策の学習を行うことはできない．そこで，最悪ケースを比較したミニマックス規範に基づく強化学習法 [Heger 94] や利得のばらつきが多い行動に負のバイアスをかけて評価する Risk-Sensitive Q-learning [Neuneier 98] が提案されている．これまで，利得の確率分布を明示的に推定し，推定された確率分布に基づきリスクの低い政策を学習する手法は提案されていない．これは，利得の確率分布を推定する効率的な方法が存在しなかったためと考えられる．

本論文では，利得の確率分布の期待値と分散という統計量に注目し，利得の分散を推定する TD アルゴリズム (本論文では TD 法に属するアルゴリズムを TD アルゴリズムと呼ぶ) を提案する．そして，リスクを考慮した規範として良く知られた variance penalized 規範 (mean-variance 規範とも呼ばれる)[Elon 91, White 88] に基づく新しい強化学習アルゴリズムを提案する．2 章において利得の期待値と分散を効率的に推定する方法を提案する．3 章では推定された利得の期待値と分散に基づきリスクを考慮した政策を学習する variance penalized 強化学習法を提案する．4 章では提案した学習アルゴリズムを単純な機械メンテナンス問題に適用し有効性を示す．最後に 5 章において本論文の成果と今後の課題についてまとめる．

2. 利得の分散を推定する TD アルゴリズムの提案

本章では，利得の分布を推定する既存研究を概観した後，利得の分散を推定するための TD アルゴリズムを提案しその収束を証明する．

2·1 利得の確率分布を推定する既存研究について

多くの強化学習手法では，各状態，各行動についてその時点の政策下での利得の期待値のみが推定されるが，利得の確率分布を明示的に推定し利用する手法もいくつか提案されている．

[McCallum 95] は部分観測マルコフ決定過程 (POMDPs) 下での強化学習手法のなかで，利得のサンプルを全て保存しておくことで利得の分布を得ている．そして，部分観測マルコフ決定過程下で部分観測性を解消するために行う状態表現変更の際に，ノンパラメトリック統計を用いてその変更の必要性を判断している．しかし，このアプローチは全てのサンプルを保持する必要がある．

[Dearden 98] は利得の正規分布を求めるために，正規 γ 分布とベイズ推定によってパラメータ推定を行っている．そして，それぞれの行動に関する利得分布を利用して，強化学習でしばしば問題となる exploration と exploitation のトレードオフの問題を解消するための状態空間の効率的な探索方法を提案している．しかし，この方法は各状態・行動について 4 つの変数を保持する必要があり，数値積分など複雑な計算が必要になる．

[Munos 99] はマルコフ連鎖 (Markov Chain) における利得の分散を求めるベルマン方程式を導出し，連続な状態空間の分割の問題を扱っている．しかし，この方法は環境のモデルパラメータを必要とし動的計画法 (Value Iteration) によって値を求めている．

これらの例のように，利得の確率分布を得ることで環境に関して利用できる情報が増え，高度な意志決定を行うことが可能になる。しかし上で述べた方法は，報酬をすべてサンプルする必要があったり，環境に関して状態遷移確率などのモデルパラメータが必要という問題点がある。

2·2 マルコフ決定過程下での利得分散

本論文では，[Munos 99] と同様に利得の分散を明示的に推定する。ただし，1) 環境としてマルコフ連鎖ではなくマルコフ決定過程 (Markov Decision Processes) を対象とする，2) 直接報酬に関しては，通常用いられる決定的なものではなくノイズを含む確率的な関数を扱う，3) 分散値の推定方法として，環境モデルを必要としない TD アルゴリズムを開発する，という 3 つの拡張を行う。

本論文で扱うのは，無限期間の離散時間・離散状態のマルコフ決定過程であり，強化学習の標準的な枠組みに従い，学習を行うエージェントはマルコフ決定過程で表現される環境と相互作用を行うものとする．各時刻 $t \in \{0,1,2,\ldots\}$ において，状態，行動，報酬が決まり，それぞれ，$s_t \in \mathcal{S}$, $a_t \in \mathcal{A}$, $r_t \in \mathcal{R}$ と表す．

環境の状態遷移は状態遷移確率 $\mathcal{P}^a_{ss'} = \Pr\{s_{t+1} = s' | s_t = s, a_t = a\}$ によって決まり，報酬は期待値に関する関数 $\mathcal{R}^a_{ss'} = E\{r_{t+1} | s_t = s, a_t = a, s_{t+1} = s'\}$，および分散に関する関数 $\mathsf{r}^a_{ss'} = Var\{r_{t+1} | s_t = s, a_t = a, s_{t+1} = s'\}$ に

よって決まる．なお，$\mathcal{R}^a_{ss'}$，$\mathsf{r}^a_{ss'}$ および直接報酬に加わるノイズはすべて有界と仮定する．エージェントは l 次元の政策パラメータベクトル $\theta \in \mathcal{R}^l$ を基に各状態，各行動について政策 $\pi(s,a,\theta) = \Pr\{a_t = a | s_t = s, \theta\}$ に従い行動を選択する．

ここで利得を，時刻 t から無限時刻先までの総割引報酬 $R_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}$ と定義する．ただし $\gamma \in [0,1)$ は割引率である．そして，時刻 t において状態 s におり，政策 π に従ったときの利得の期待値と分散をそれぞれ以下のように表す．

$$V^\pi(s) = E\{R_t | s_t = s, \pi\}, \tag{1}$$

$$v^\pi(s) = Var\{R_t | s_t = s, \pi\}. \tag{2}$$

なお今後，式 (1)，式 (2) をそれぞれ状態 s の期待価値，分散価値と呼ぶ．式 (2) の $v^\pi(s)$ に関して，以下のベルマン方程式が得られる (証明：付録 A 参照)．

$$\begin{aligned} v^\pi(s) &= \sum_a \pi(s,a) \sum_{s'} \mathcal{P}^a_{ss'} \{(\mathcal{R}^a_{ss'} + \gamma V^\pi(s') - V^\pi(s))^2 \\ &\quad + \mathsf{r}^a_{ss'} + \gamma^2 v^\pi(s')\} \qquad (3) \\ &= \sum_{s'} \mathcal{P}^\pi_{ss'} \{(\mathcal{R}^\pi_{ss'} + \gamma V^\pi(s') - V^\pi(s))^2 + \mathsf{r}^\pi_{ss'} \\ &\quad + \gamma^2 v^\pi(s')\} \qquad (4) \end{aligned}$$

ただし，

$$\mathcal{P}^\pi_{ss'} = \sum_a \pi(s,a) \mathcal{P}^a_{ss'},$$

$$\mathcal{R}^\pi_{ss'} = \sum_a \pi(s,a) \mathcal{P}^a_{ss'} \mathcal{R}^a_{ss'} / \mathcal{P}^\pi_{ss'},$$

$$\mathsf{r}^\pi_{ss'} = \sum_a \pi(s,a) \mathcal{P}^a_{ss'} \{(\mathcal{R}^a_{ss'} - \mathcal{R}^\pi_{ss'})^2 + \mathsf{r}^a_{ss'}\} / \mathcal{P}^\pi_{ss'}.$$

式 (3), (4) は状態 s から政策 π によって得られる利得の分散が，状態 s と次の状態 s' に関連する利得の期待値と分散および s と s' に関連するマルコフモデルの部分的なモデルパラメータのみから求められることを意味している．この式を利用して $v^\pi(s)$ を求める方法として，すべての s, a, s' に関して $V^\pi(s)$，$\mathcal{P}^\pi_{ss'}$，$\mathcal{R}^\pi_{ss'}$，$\mathsf{r}^\pi_{ss'}$ を推定し行列計算によって求めるアプローチと，確率的近似アルゴリズムと呼ばれる方法によって求めるアプローチがある．我々はモデルを必要としない後者のアプローチを採用する．

2·3 TD アルゴリズムによる推定

確率的近似 (stochastic approximation, stochastic iterative) アルゴリズムの更新式は以下のように表される [Tsitsiklis 96].

$$v \leftarrow (1-\beta)v + \beta(Hv + w) \tag{5}$$

ただし，v は推定対象となっている変数のベクトル，β はステップサイズパラメータ，w はノイズ要素，Hv は推定量 v を用いた推定オペレータである．TD アルゴリズムは，マルコフ決定過程におけるベルマン方程式を利用した確率的近似アルゴリズムである．

さて，前節で導出したベルマン方程式に基づいて，利得の分散を推定するための TD アルゴリズムを提案する．このアルゴリズムは状態 i から状態 j に遷移した場合，以下の更新式に従う．ただし，β_t，ε_t は正のステップサイズパラメータである．

$$V_{t+1}(i) = (1-\beta_t)V_t(i) + \beta_t(r_{t+1} + \gamma V_t(j)), \tag{6}$$

$$\begin{aligned} v_{t+1}(i) = (1-\varepsilon_t)v_t(i) + \varepsilon_t\{&(r_{t+1} + \gamma V_t(j) \\ &- V_t(i))^2 + \gamma^2 v_t(j)\}. \end{aligned} \tag{7}$$

ここで以下の定理が成立する．

［定理 2·1］（TD アルゴリズムの収束） すべての s について，$V_t(s)$ と $v_t(s)$ はそれぞれ式 (6) と式 (7) の TD アルゴリズムによって生じた乱数系列とする．このとき，ステップサイズパラメータ β_t，ε_t が共に条件，

$$\sum_{t=0}^{\infty} \beta_t = \infty, \quad \sum_{t=0}^{\infty} \beta_t^2 < \infty,$$

$$\sum_{t=0}^{\infty} \varepsilon_t = \infty, \quad \sum_{t=0}^{\infty} \varepsilon_t^2 < \infty.$$

を満たすならば，すべての $s \in \mathcal{S}$ について，系列 $V_t(s)$ と $v_t(s)$ はそれぞれ $V^\pi(s)$ と $v^\pi(s)$ に概収束する．

《証明》 付録 B 参照.

この TD アルゴリズムの利点の一つは，様々な線形近似手法やニューラルネットワークなどの関数近似と組み合わせが容易なことである．この特徴は，状態数の多いマルコフ決定過程における効率的・現実的な利得の分散推定を可能にする．

また，TD 法は価値の推定のために遷移先の状態の価値の推定値を使うが，これは，3 章で述べるような評価関数の勾配に従って政策を少しずつ更新する学習アルゴリズムと組み合わせた場合には収束速度の点で効率的となる．なぜなら，少しずつ政策変更を繰り返すアルゴリズムでは，変更前の政策における価値の推定値は，変更後の政策における価値の推定を行う際の有望な初期推定値となるからである．

上で述べた方法によって利得の分散を得ることの利点の一つとして，利得のリスクを考慮した政策の学習が可能になることが挙げられる．そこで，次章において代表的な政策評価規範である variance penalized 規範に注目する．

3. Variance Penalized 強化学習法の提案

特に金融工学の分野では，利得の期待値が最大で分散が大きな政策よりも，利得の期待値がほどほどで分散が小さな政策の方が好まれる場合が多い．無限回の意志決

定を行うことが可能ならば前者の政策も合理的と考えられるが，もし短期間に連続して失敗すると倒産してしまうといった状況下では後者の政策が好まれる．良く知られる規範としては，最悪ケースのみを考慮したミニマックス規範が存在するが，多くの確率的なシステムにおいては悲観的な政策しか学習できない．

本章では，与えられたリスクレベルに応じた政策の価値を定義可能な variance penalized 政策評価規範 [Elon 91, White 88] に着目し，新しい強化学習アルゴリズムである VPD-GPI(Gradient-based Policy Iteration for Variance Penalized Discounted MDPs) アルゴリズムを提案する．

3·1 政策評価規範に基づくアプローチ

variance penalized 政策評価規範には，どの程度の期間 (horizon) の報酬を考慮するか，報酬の割り引きを考慮するかによるバリエーションがあるが，我々は最も一般的な無限期間で割引のある場合を扱う．標準的なマルコフ決定過程の政策規範と同様に初期状態分布ベクトル $\mathbf{s}_0$ を用い，学習の目的は以下の π を得ることとする．

$$\max_{\pi} f^{\pi}, \quad f^{\pi} = \mathbf{s}'_0(\mathbf{V}^{\pi} - \alpha \mathbf{v}^{\pi}). \tag{8}$$

ここで，$\mathbf{V}^{\pi}$，$\mathbf{v}^{\pi}$ はそれぞれ，前節で定義された利得の期待値，分散を表す $V^{\pi}(s)$，$v^{\pi}(s)$ のベクトルを表す．また，$\mathbf{x}'$ はベクトル $\mathbf{x}$ の転置ベクトルを表す．α は正の実数パラメータであり，リスクと利益水準の好みを調節する．

この規範の下では定常な最適政策は一般に存在しないことが知られている [White 88]．すなわち，この評価関数は非マルコフ決定問題を生じさせるので，最適な政策を求めようとすると非定常な政策空間を探索せねばならず，非常に小さな問題を除いて実現は不可能である．

そこで，我々は評価式 (8) の準最適解の一つを効率的に求めることを目指す．このような非マルコフ問題を解くための強化学習法として，部分観測マルコフ決定過程 (POMDPs) における強化学習のアプローチがある [Jaakkola 95, Baird 98, Sutton 99]．それらの手法では，まず政策評価関数の勾配を求め，その勾配方向に確率的政策を少しずつ変更する．環境の部分観測性に因する非マルコフ性と政策評価規範に因する非マルコフ性の違いはあるものの，本章が対象とする問題においてもこのアプローチは有効である．この考えに基づき，我々は勾配法ベースの variance penalized 強化学習アルゴリズムを提案する．

3·2 Variance Penalized 勾配式

まず，状態 s で行動 a を行い，その後政策 π に従って行動する場合の利得の期待値と分散を以下のように定義する．

$$Q^{\pi}(s,a) = E\{R_t | s_t = s, a_t = a, \pi\}, \tag{9}$$

$$q^{\pi}(s,a) = E\{(R_t - V^{\pi}(s))^2 | s_t = s, a_t = a, \pi\} \tag{10}$$

ここで，$\pi(s,a)$ は政策であり，状態 s において行動 a を選択する確率を表す。$V^{\pi}(s) = \sum_a \pi(s,a) Q^{\pi}(s,a)$，および，$v^{\pi}(s) = \sum_a \pi(s,a) q^{\pi}(s,a)$ という関係が成り立つ．そして，政策 π の式 (8) における評価値を f^{π} と表したとき，政策を保持するための近似パラメータ θ に対する f^{π} の勾配に関して以下の定理が成立する．

［定理 3·1］（variance penalized 勾配式） 任意のマルコフ決定過程について，

$$\frac{\partial f^{\pi}}{\partial \theta} = \sum_{t=0}^{\infty} \sum_{s} \Pr(s_t = s | \mathbf{s}_0, \pi) \{ \gamma^t \partial Q^{\pi}(s,\theta) - \alpha(\gamma^{2t} \partial q^{\pi}(s,\theta) + 2e(s,t) \partial Q^{\pi}(s,\theta)) \}, \tag{11}$$

ただし，

$$\partial Q^{\pi}(s,\theta) = \sum_a Q^{\pi}(s,a) \frac{\partial \pi(s,a)}{\partial \theta},$$

$$\partial q^{\pi}(s,\theta) = \sum_a q^{\pi}(s,a) \frac{\partial \pi(s,a)}{\partial \theta},$$

$$e(s,t) = \begin{cases} 0 & (t=0) \\ E\{\sum_{\tau=1}^{t} \gamma^{t+\tau}(r_\tau + \gamma V^{\pi}(s_\tau) \\ -V^{\pi}(s_{\tau-1}))|s_t = s\} & (t \neq 0). \end{cases}$$

《証明》 付録 C 参照．

ここで $e(s,t)$ はスタート時刻 0 から時刻 t までの TD 誤差を割り引いて足し合わせたものであり，$e(s,t) = E\{\gamma^{2t}(r_t + \gamma V(s_t) - V(s_{t-1})) + \gamma e(s_{t-1}, t-1) | s_t = s\}$ という関係が成立する．

3·3 Policy Iteration 型学習アルゴリズム

式 (11) の勾配式により推定された $Q^{\pi}(s,a)$ と $q^{\pi}(s,a)$，各時刻での状態占有確率 $\Pr(s_t = s|\ldots)$ のサンプルおよび累積 TD 誤差のサンプルを用いて，政策評価関数式 (8) の政策保持パラメータ θ に対する勾配を近似的に求めることができる．ここではパラメータベクトルの更新量 $\Delta\theta$ を以下のように求める．

$$\Delta\theta = \kappa_k \frac{\partial f^{\pi}}{\partial \theta}, \tag{12}$$

ただし，κ_k は k 回目の政策更新のためのステップサイズパラメータであり，k の増加と共に減少するという性質を持つ．

式 (6) と式 (7) の TD アルゴリズムによる推定，式 (11) に基づく勾配推定および式 (12) の更新式から，図 1 の Policy Iteration アルゴリズムを構成することができる．このアルゴリズムを VPD-GPI と呼ぶ．VPD-GPI はモデルの推定を行わない非逐次型の学習アルゴリズムである．政策評価ステップにおいて様々な関数近似システムと組み合わせることにより，大規模問題への適用も可能

(1) 初期化: すべての s,a について $Q(s,a)=q(s,a)=0$ とし，$\theta=\theta_0$, $\kappa=\kappa_0$, $k=0$ とする．
(2) 政策評価: 全ての s,a について，政策 $\pi(\theta)$ 下での状態遷移に基づいて，式 (6) と (7) の TD アルゴリズムを用いて $Q(s,a)$ と $q(s,a)$ を推定する．
(3) 政策改善:
- 勾配定理を用いてすべての l に関して $\frac{\partial f}{\partial \theta_l}$ を推定する．具体的には，政策 $\pi(\theta)$ 下での状態遷移に基づいて，$Q(s,a)$ と $q(s,a)$，占有確率 $\Pr(s_t=s|\ldots)$ のサンプル，割り引き累積 TD 誤差 $(e(s,t))$ のサンプルを用いる．
- $\theta \leftarrow \theta+\kappa_k \frac{\partial f}{\partial \theta}$ という式に従い政策保持パラメータベクトル θ を更新する．
- もし k が十分大きい，あるいは，更新ベクトル $\Delta\theta$ のノルムが十分に小さいならば，ループを出る．そうでないならばステップサイズ κ をスケジュールに基づき減少させ，k を 1 増やし，(2) へ．

図 1 VPD-GPI アルゴリズム

と考えられる．VPD-GPI は variance penalized 規範下で効率的に準最適政策を探索する初めてのアルゴリズムである．

3·4 発 展 的 考 察

本節では，3.1. 節で提示した variance penalized 規範と提案した VPD-GPI に関して，工学的応用という観点から重要ないくつかのアルゴリズムについて考察する．

- 決定的政策: variance penalized 規範を用いた場合でも，前節のように確率的政策を用いず，単に以下の決定的政策 $\mu(s)$ を選択することもできる．

$$\mu(s)=\arg\max_a (Q^\pi(s,a)-\alpha q^\pi(s,a)) \quad \forall s\in\mathcal{S} \tag{13}$$

関数近似を用いるような大規模な問題では，政策を少しずつ変更せず最善の評価の行動のみを選ぶ政策は，学習速度の面で有効と考えられる．ただし，variance penalized 規範下ではマルコフ的政策は最適という保証はないことおよび学習により政策の質が劇的に変化するため，場合によっては推定値が不安定になることに留意する必要がある．

- 勾配式の近似：前節で導出した勾配定理には，$e(s,t)$ という TD エラーを割り引いて足し合わせた項が含まれていた．TD エラーの和は現在の状態 s に遷移する以前のエピソードにおいて，平均よりも良い状態遷移が連続して起これば大きな正の値になり，逆に平均よりも悪い状態遷移が連続して起これば大きな負の値になる．もし多くの状態において平均よりも良い状態遷移と悪い状態遷移がエピソード中に同程度に起こると仮定すると，$e(s,t)$ の値は $q^\pi(s,a)$ と比較して十分小さいと期待できる．すると近似勾配式は以下の式になる．$e(s,t)$ の推定には多くのサンプルが必要となるのでこのアルゴリズムは効率的である．

$$\frac{\partial f}{\partial \theta}\approx \sum_{t=0}^{\infty}\sum_s \Pr(s_t=s)\{\gamma^t \partial Q^\pi(s)-\alpha\gamma^{2t}\partial q^\pi(s)\} \tag{14}$$

ただし $\partial Q^\pi(s)$, $\partial q^\pi(s)$ はそれぞれ $\partial Q^\pi(s,\theta)$, $\partial q^\pi(s,\theta)$ を表す．分散が大きな環境においてはこのような近似を用いることも可能と考えられる．

- Lookup tables：今まで政策がパラメータベクトル θ によって保持される一般的な場合を扱ってきたが，ここでは特殊ケースとして，すべての状態 $s\in\mathcal{S}$, 行動 $a\in\mathcal{A}$ について独立のパラメータ $p(s,a)$ (即ち lookup table) を保持する場合を考える．これらは，$\sum_a p(s,a)=1$ という制約条件を持つパラメータ変数と考えればよい．このとき，

$$\sum_a Q(s,a)\frac{\partial \pi(s,a)}{\partial p(s,a)}\propto Q(s,a)-V(s),$$
$$\sum_a q(s,a)\frac{\partial \pi(s,a)}{\partial p(s,a)}\propto q(s,a)-v(s)$$

という関係がそれぞれ成立するので，例えば上の近似勾配式は以下の式になる．

$$\frac{\partial f}{\partial p(s,a)}\approx \sum_{t=0}^{\infty}\sum_s \Pr(s_t=s)\{\gamma^t(Q(s,a)-V(s)) -\alpha\gamma^{2t}(q(s,a)-v(s))\} \tag{15}$$

この式から，ある状態 s の行動の中で $V^\pi(s)$ よりも大きな $Q^\pi(s,a)$ をもつ行動の確率を増やす要求と $v^\pi(s)$ よりも大きな $q^\pi(s,a)$ をもつ行動の確率を減らす要求が α によって足し合わされていることが分かる．

4. 実　　験

3 章において利得のリスクを考慮した variance penalized 政策評価規範を示し，その評価規範に基づいた効率的な学習アルゴリズムを提案した．本章では，提案した学習アルゴリズムを単純な機械メンテナンス問題 [Mahadevan 97] に適用し，理論値と比較することにより提案したアルゴリズムの正しさを示す．また，variance penalized 規範の妥当性と提案アルゴリズムの有効性を示す．

4·1 機 械 整 備 問 題

政策学習の対象となるシステムは図 2-(a) の機械整備問題である．このシステムは，3 状態 (0:正常, 1:不調, 2:故障)，2 行動 (0:使用, 1:整備)，3 種類の報酬 (機械を整備するコスト，機械を使用する利益，故障した機械を修理するコスト) のマルコフ決定過程としてモデル化できる．ただし，故障状態 (状態 2) では整備 (修理) 以外の行動選択の余地はないため，実際には最適な 2 次元政策

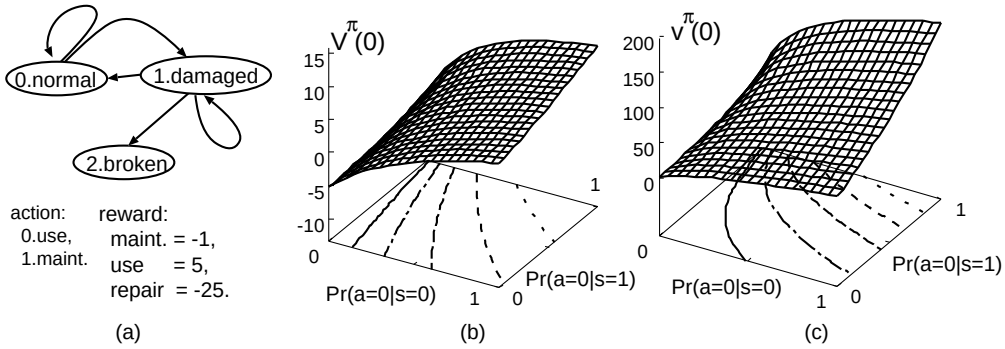

図 2 (a) 機械整備問題の状態遷移図, 初期状態における利得の期待値 (b) と分散 (c)

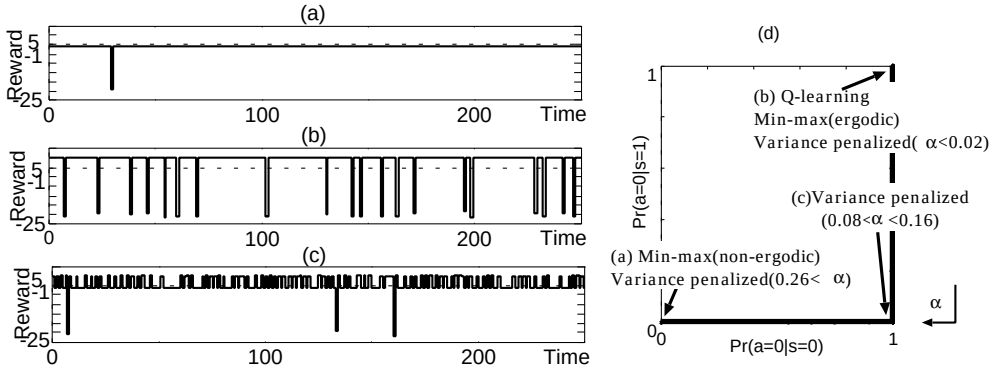

図 3 代表的政策下 (a:整整-政策，b:使使-政策, c:使整-政策) の報酬のスナップショット と様々な α における最適政策

パラメータ $p(0) = \Pr(\text{使用} \mid \text{正常})$(正常状態における使用行動の選択確率)，および，$p(1) = \Pr(\text{使用} \mid \text{不調})$(不調状態における使用行動の選択確率) を求める問題となる．このシステムの状態遷移関数と報酬関数を付録.D に示す値に設定し，割り引き係数 γ を 0.8 に，正常状態を唯一の初期状態とすると，各政策に対する初期状態の期待価値と分散価値 ($V(0)$,$v(0)$) の値は，図 2-(b),(c) の 3 次元グラフで表示される．図の x-y 平面は政策を表し，z 軸はそれぞれ状態 0 における利得の期待値と分散を，x-y 平面上の点線はそれぞれ状態 0 における利得の期待値と分散の等高線を表す。

図 2 によると，例えば $p(0) = p(1) = 1.0$ (以下，使使-政策と呼ぶ）などの利得の期待値が高い政策は利得の分散も高く，例えば $p(0) = p(1) = 0.0$ (以下，整整-政策と呼ぶ）などの利得の分散が低い政策は利得の期待値も低くなっている．これは利得の期待値を大きくしたいという要請と利得の分散を小さくしたいという要請の間のトレードオフが存在することを意味する．この問題は単純な問題であるが，利得のリスクを考慮した学習法について調べる目的には適している．

4·2 各種規範下での政策の挙動

図 3-(a),(b),(c) は 3 種類の代表的な決定的政策における報酬の 250 時刻分のスナップショットである．(a) の整整-政策は報酬の分散は非常に低く故障の確率が小さいが，全く正の報酬は得られない．(b) の使使-政策は報酬分散が非常に高いが故障の確率も大きい．一方，(c) の使整-政策は中間的な報酬水準と，故障の確率をとる．

報酬の最悪値のみ状態価値として保持し，最悪値を最大化する政策を学習するミニマックス強化学習 [Heger 94] では，整備行動を選ぶと必ず修理が可能な環境 (即ち非エルゴード的環境) では常に修理をする (a) を学習する。しかし，少しでも失敗の可能性がある環境 (エルゴード的環境) では逆に全く修理をしない政策 (b) を学習する。代表的な Q-learning は任意の $\gamma \in [0,1)$ において政策政策を学習する．割り引き係数 γ によって，時間軸方向への利得のリスクに関する嗜好を反映させることができることは知られているが，各時刻での利得のばらつきは考慮されない．一方，variance penalized 規範を用いると，適切なリスク嗜好パラメータ α のもとで，(a) から (c) の政策を学習する。図 3-(d) に α を変化させた時の最適政策の変化を示す．この図から，α の値によっては確率的政策が最善となる場合もあり得る (例えば，$\alpha = 0.20$ の場合 $p(0) = 0.75, p(1) = 0.0$ 付近が最善となる)．よって，この問題においては確率的政策空間を対象とする VPD-GPI のアプローチは有効と考えられる．

4·3 利得の分散推定

2.3 節で提案した TD アルゴリズムを用いて，4.1 節で述べた機械整備システムにおける利得の期待値と分散を推定する．ここで政策を固定しなければならないが，最も分散が大きくなる使使-政策を選択した．ステップサイ

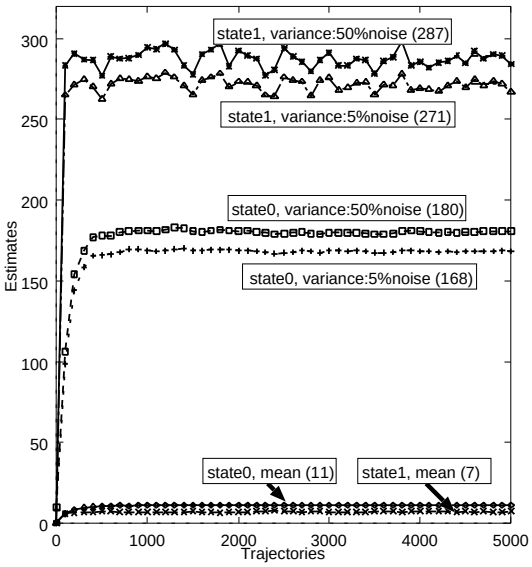

図 4 利得の統計量の TD アルゴリズムによる推定結果

ズパラメータは全ての状態について，期待値，分散ともに以下の式 [Boyan 99] で計算した．

$$\beta_n = \beta_0 \frac{n_0 + 1}{n_0 + n} \qquad n = 1, 2, \ldots \tag{16}$$

ここで，β_n は n エピソード (trajectory) 目のステップサイズを，β_0 は初期ステップサイズ (実際には 0.01 に設定) を，n_0 はステップサイズ減少用パラメータ (実際には 1000 に設定) を表しており，収束定理の条件を満たしている．直接報酬にかかるノイズに関しては，各報酬の絶対値の 5%と 50%の 2 種類について調べた．図 4 はその結果であり，提案した TD アルゴリズムによって理論値 (() の中が理論値) とほぼ同じ推定値が得られていることが確認された．なお，数値は 10 試行の平均値である．

4·4 VPD-GPI の学習結果

図 1 に示した VPD-GPI をこの機械整備問題に適用した．各状態 s，各行動 a について利得の期待値 $Q(s,a)$，利得の分散 $q(s,a)$ および政策 $\pi(s,a)$ を保持しなければならないが，この問題は状態数が少ないことと，関数近似による準最適政策への収束を避ける目的で，独立なパラメータ (lookup table) を用いた．そして，k 回目の政策更新時における政策パラメータ更新用ステップサイズパラメータは $\kappa_k = 0.02(0.99)^k$ という値を用いた．また，政策更新回数の最大値を 100 回とした．図 5-上は学習結果である．$\alpha = 0.0, 0.1, 0.2, 0.3$ のそれぞれの場合について，様々な初期政策から学習を行った場合の政策の変化である．図の点線の違いは初期政策の違いを表している。この結果から，α を変化させることにより，使使-政策 ($\alpha = 0.0$, Q-learning の最適政策)，使整-政策 ($\alpha = 0.1$)，確率的政策 ($\alpha = 0.2$)，整整-政策 ($\alpha = 0.3$，ミニマックス学習の最適政策) が最終的に得られていることが確認できる．また，図 5-下はそれぞれ，上の図と対応した α における各政策の評価理論値である．図の x-y 平面は政策を表し，z 軸は式 ((8)) の政策評価関数 f^π の値を表している。また，x-y 平面上の点線は f^π の等高線を表す。

上下の図を比較することにより，政策の学習が勾配方向に適切に行われていることが確認された．また，評価値の勾配が急でない高台部分では政策が安定しないことも観察されている．これを避けるためには，慣性項を導入するなどより高度なパラメータ変更が必要と考えられる．

これらの結果から，提案した利得の期待値と分散を推定する TD アルゴリズムおよび VPD-GPI の妥当性と有効性が示されたといえる．

5. お わ り に

本論文では，マルコフ決定過程における利得の確率分布を得ることは有用であるという考えのもとに，利得の分散を効率的に得る方法を提案した．利得の分散に関するベルマン方程式を拡張し，分散のための TD アルゴリズムの収束性を証明した．本論文で扱った利得は無限期間の総割り引き報酬という最も一般的な設定である．また，利得の平均と分散を明示的に利用する方法として，リスクを考慮した政策に関する評価規範である variance penalized (variance penalised) 規範に基づく新しい強化学習アルゴリズム VPD-GPI を提案した．VPD-GPI は政策の評価と政策の改善のフェーズを繰り返すことにより，確率的政策空間の準最適解を学習する，この学習アルゴリズムは variance penalized 規範に基づく初めての効率的なアルゴリズムである．機械整備問題において，提案した TD アルゴリズムと VPD-GPI の学習結果と理論値と比較することで有効性を示した．

このアルゴリズムを関数近似と組み合わせることにより大規模問題において利得の平均と分散を推定し，リスクを考慮した政策を学習することが可能になると期待できる．推定された利得の分散は，本論文で扱ったリスクを考慮した政策の学習の他にも強化学習でしばしば課題となる exploration と exploitation のトレードオフの問題や，思考ゲームなど様々な場面で利用できると考えられる．

上で述べた新しい意志決定の枠組を提供するのに加え，本論文の理論を部分観測マルコフ決定過程 (POMDPs) やセミマルコフ決定過程 (SMDPs) に拡張すること，関数近似と組み合わせた際の性質について調べることが今後の課題である．

◇ 参 考 文 献 ◇

[Baird 98] Baird, L., Moore, A.: *Gradient descent for general reinforcement learning.*, Advances in Neural Information Processing Systems 11, (1998).

[Boyan 99] Boyan, J. A.: *Least-Squares Temporal Difference Learning.*, Proceedings of the 16th International Conference on Machine Learning, pp.49-56, (1999).

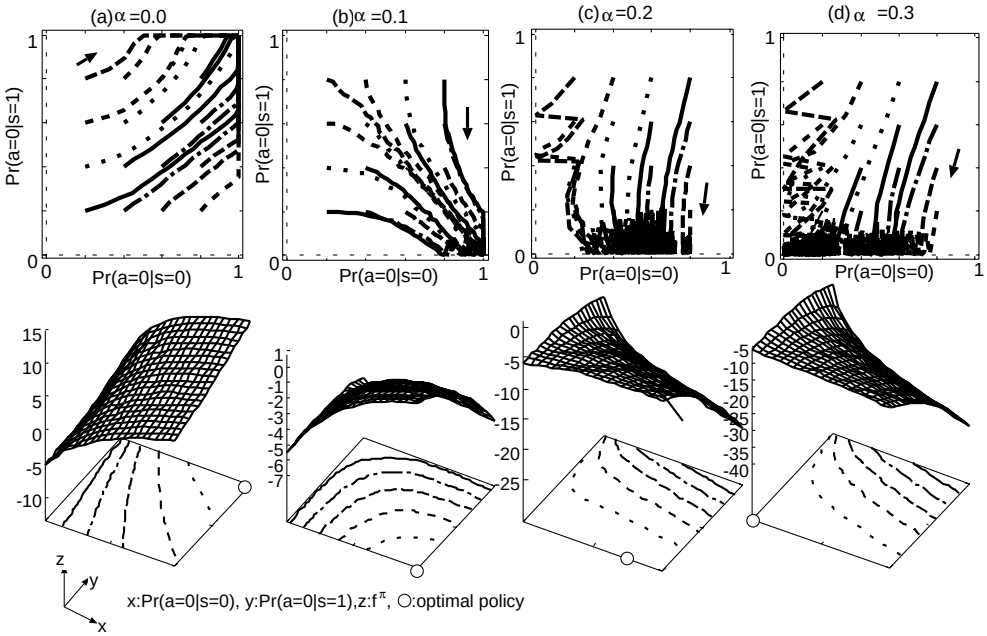

図 5 様々な α における政策の変化 (上) と理論値 (下)

[Brown 98] Brown, T. X., Tong, H., Singh, S.: *Optimizing admission control while ensuring quality of service in multimedia networks via reinforcement learning.*, Advances in Neural Information Processing Systems 11, pp.982-988, (1998).

[Crites 96] Crites, R. H., Barto, A. G.: *Improving elevator performance using reinforcement learning*, Advances in Neural Information Processing Systems 9, pp.1017-1024, (1996).

[Dearden 98] Dearden, R., Friedman, N., Russell, S.: *Bayesian q-learning*, In Proceedings of the 15th National Conference on Artificial Intelligence, pp.761-768, (1998).

[Elon 91] Elon, E., Gruber, M.: *Modern Portfolio Theory and Investment Analysis.*, John Wiley and Sons, Inc, (1991).

[Fernandez 95] Fernandez, E., Marcus, M. I.: *Non-standard optimality criteria for stochastic control problems*, Technical Reports on ISR-TR, pp.95-101, (1995).

[Heger 94] Heger, M.: *Consideration of risk and reinforcement learning.* Proceedings of the 11th International Conference on Machine Learning, pp.105-111, (1994).

[Jaakkola 95] Jaakkola, T., Singh, S. P., Jordan, M.I.: *Reinforcement learning algorithm for partially observable Markov decision problems.*, NIPS '95, pp.345-352, (1995).

[Kadota 98] Kadota, Y., Kurano, M., Yasuda, M.: *On the general utility of diccounted Markov decision processes*, IFORS96 Pepar No.115, vol.22, (1998).

[McCallum 95] McCallum, R.A.: *Instance-based utile distinctions for reinforcement learning with hidden state*, Proceedings of the 12th International Conference on Machine Learning, pp.387-395, (1995).

[Mahadevan 96] Mahadevan, S.: *Average reward reinforcement learning: Fundations, algorithms, and empirical results.* Machine Learning, vol.22, pp.159-196, (1996).

[Mahadevan 97] Mahadevan, S., Marchalleck, N., Das, T.K., Gosavi, A.: *Self-Improving Factory Simulation using Continuous-time Average-Reward Reinforcement Learning.* Proceedings of the 14th International Conference on Machine Learning, pp.202-210, (1997).

[Marcus 97] Marcus, S. I., Fernandez, E., Hernandez, D., Coraluppi, S., Fard, P.: *Risk Sensitive Markov Decision Processes.* Systems and Control in the Twenty-First Century, Boston Birkhauser, pp.263-279, (1997).

[Munos 99] Munos, R., Moore, A.: *Variable resolution discretization for high-accuracy solutions of optimal control problems.* 16th International Joint Conference on Artificial Intelligence, pp.1348-1355, (1999).

[Neuneier 97] Neuneier, R.: *Enhancing Q-Learning for Optimal Asset Allocation.*, Advances in Neural Information Processing Systems 10, pp.936-942, (1997).

[Neuneier 98] Neuneier, R., Mihatsch, O.: *Risk Sensitive Reinforcement Learning.*, Advances in Neural Information Processing Systems 11, pp.1031-1037, (1998).

[Schwartz 93] Schwartz, A.: *A reinforcement learning method for maximizing undiscounted rewards .*, Proceedings of the 10th International Conference on Machine Learning, pp.298-305, (1993).

[Singh 97] Singh, S., Bertsekas, D.: *Reinforcement learning for dynamic channel allocation in cellular telephon systems*, Advances in Neural Information Processing Systems 10, pp.974-980.

[Sutton 88] Sutton, R. S.: *Learning to predict by the methods of temporal differences.*, Machine Learning 3, pp.9-44 (1998).

[Sutton 98] Sutton, R. S., Barto, A. G.: *Reinforcement learning: An Introduction.*, MIT Press (1998).

[Sutton 99] Sutton, R. S., McAllester, D., Singh, S., Mansour, Y.: *Policy gradient methods for reinforcement learning with function approximation.*, Advances in Neural Information Processing Systems 12,(1999).

[Tsitsiklis 96] Tsitsiklis, J. N.: *Neuro-Dynamic Programming.*, Athena Scientific (1996).

[Watkins 89] Watkins, C.: *Learning from Delayed Rewards.*, PhD thesis, King's College, England (1989).

[White 88] White, D. J.: *Mean, variance, and probabilistic criteria in finite Markov decision processes: A review.* Journal of Optimization Theory and Applicatoins, vol.56(1), pp.1-29, (1988).

[Zhang 96] Zhang, W., Dietterich, T. G.: *High performance job-shop scheduling with a time-delay TD(λ) network.*, Advances in Neural Information Processing Systems 9, pp.1024-1030.

〔担当委員：櫻井彰人〕

2000 年 6 月 28 日　受理

◇ 付　　録 ◇

A. 利得の分散に関する Bellman 方程式

状態 s において行動 a を選択し状態 s' に遷移したのちに，政策 π に従い行動選択する時の利得 R_t の分散は以下のように求まる．

$$
\begin{aligned}
&E\{(R_t - V^\pi(s))^2 | s_t = s, a_t = a, s_{t+1} = s', \pi\} \\
&= E\{(r_{t+1} + \gamma \sum_{k=1}^{\infty} \gamma^k r_{t+k+1} - V^\pi(s))^2 | \ldots\} \\
&= E\{(r_{t+1} + \gamma \sum_{k=0}^{\infty} \gamma^k r_{t+k+2} - V^\pi(s) + \gamma V^\pi(s') - \gamma V^\pi(s'))^2 | \ldots\} \\
&= E\{(r_{t+1} + \gamma V^\pi(s') - V^\pi(s))^2 | \ldots\} \\
&\quad + 2\gamma E\{(\sum_{k=0}^{\infty} \gamma^k r_{t+k+2} - V^\pi(s'))(r_{t+1} + \gamma V^\pi(s') - V^\pi(s)) | \ldots\} \\
&\quad + \gamma^2 E\{(\sum_{k=0}^{\infty} \gamma^k r_{t+k+2} - V^\pi(s'))^2 | \ldots\} \\
&= E\{(R^a_{ss'} + w + \gamma V^\pi(s') - V^\pi(s))^2 | \ldots\} \\
&\quad + 2\gamma E\{(\sum_{k=0}^{\infty} \gamma^k r_{t+k+2} - V^\pi(s')) | \} E\{(r_{t+1} + \gamma V^\pi(s') - V^\pi(s)) | \} \\
&\quad + \gamma^2 E\{(\sum_{k=0}^{\infty} \gamma^k r_{t+k+2} - V^\pi(s'))^2 | \ldots\} \\
&= E\{(R^a_{ss'} + w + \gamma V^\pi(s') - V^\pi(s))^2 | \ldots\} + \gamma^2 v^\pi(s') \\
&= (R^a_{ss'} + \gamma V^\pi(s') - V^\pi(s))^2 \\
&\quad + 2(R^a_{ss'} + \gamma V^\pi(s') - V^\pi(s)) E\{w | \ldots\} + E\{w^2 | \ldots\} + \gamma^2 v^\pi(s') \\
&= (R^a_{ss'} + \gamma V^\pi(s') - V^\pi(s))^2 + \mathsf{r}^a_{ss'} + \gamma^2 v^\pi(s')
\end{aligned}
$$

ここで，w は直接報酬に加わるノイズであり，環境モデルの定義より，平均 0，分散 $\mathsf{r}^a_{ss'}$ である．このような関係がすべての $s, s' \in \mathcal{S}, a \in \mathcal{A}$ について成立するので，以下の関係が成り立つ．

$$
\begin{aligned}
&v^\pi(s) = E\{(R_t - V^\pi(s))^2 | s_t = s, \pi\} \\
&\sum_a \pi(s,a) \sum_{s'} \mathcal{P}^a_{ss'} E\{(R_t - V^\pi(s))^2 | s_t = s, a_t = a, s_{t+1} = s', \pi\} \\
&\sum_a \pi(s,a) \sum_{s'} \mathcal{P}^a_{ss'} \{(R^a_{ss'} + \gamma V^\pi(s') - V^\pi(s))^2 + \mathsf{r}^a_{ss'} + \gamma^2 v^\pi(s')\}.
\end{aligned}
$$

Q.E.D

B. TD アルゴリズムの収束

全ての $i \in \mathcal{S}$ について，$V(i)$ が式 (6) の TD アルゴリズによって推定されると仮定する．このとき，式 (7) の TD アルゴリズムは以下のように変換できる．

$$v_{t+1}(i) = (1 - \varepsilon_t) v_t(i) + \varepsilon_t ((H_t v_t)(i) + w_t(i) + u_t(i)), \quad \forall i \in S,$$

ただし，

$$
\begin{aligned}
(H_t v_t)(i) &= \sum_{j \in S} \mathcal{P}^\pi_{ij} \{(\mathcal{R}^\pi_{ij} + \gamma V_\infty(j) - V_\infty(i))^2 + \mathsf{r}^\pi_{ij} + \gamma^2 v_t(j)\}, \\
w_t(i) &= (R^\pi_{ij} + n + \gamma V_\infty(j) - V_\infty(i))^2 + \gamma^2 v_t(j) - (H_t v_t)(i), \\
u_t(i) &= \{\gamma(V_t(j) - V_\infty(j)) - (V_t(i) - V_\infty(i))\}\{\gamma(V_t(j) - V_\infty(j)) \\
&\quad - (V_t(i) - V_\infty(i)) + 2(\mathcal{R}^\pi_{ij} + \gamma V_\infty(j) - V_\infty(s))\}.
\end{aligned}
$$

であり，n は直接報酬にかかるノイズを表す．また，$w_t(i)$ は推定ノイズ項を，$u_t(i)$ はバイアス項を表す．ここで，もしこのアルゴリズムが以下の 4 つの性質を満たしステップサイズパラメータが定理のなかで述べた条件を満たすならば，[Tsitsiklis 96] における (Prop.4.5 の) 収束定理を適用することができ，推定値は真の値に概収束する．

(1) H は weighted maximum ノルムに関して縮小写像である．
(2) 全ての i,t について $E\{w_t(i) | \mathcal{F}_t\} = 0$ である (推定ノイズ項の平均が 0).
(3) 全ての i,t, 任意のノルム $\|\cdot\|$ について $E\{w_t^2(i) | \mathcal{F}_t\} \leq A + B \| v_t \|^2$ である (推定ノイズ項の分散が有界).
(4) 全ての i,t についてバイアス項の絶対値が 0 に概収束する正の乱数系列 θ_t よりも大きくならない (バイアス項が 0 に収束).

ただし，$\mathcal{F}_t$ は時刻 t までの状態遷移の履歴を表す．

まず，詳細は省略するが，(1) に関しては利得の期待値を推定する通常の TD アルゴリズムと同様の方法で示すことができる ([Tsitsiklis 96] の p.249 参照)．(2) に関しては $w_t(i)$ の定義式と，$E\{n\} = 0$ から明らかである．

(3) に関しては，

$$
\begin{aligned}
E[w_t^2 | \mathcal{F}_t] &\leq E[\{(R^\pi_{ss'} + n + \gamma V_\infty(s') - V_\infty(s))^2 + \gamma^2 v_t(s')\}^2] \\
&\leq (((1+\gamma)K_1 + K_2 + K_3)^2 + \gamma^2 \max_s |v_t|)^2 \\
&\leq 2\gamma^4 \| v_t \|^2 + 2((1+\gamma)K_1 + K_2 + K_3)^4
\end{aligned}
$$

ただし $\|\cdot\|$ は maximum ノルム，$K_1 = \max_s |V_\infty(s)|$, $K_2 = \max_{ss'} |\mathcal{R}^\pi_{ss'}|$, K_3 は直接ノイズの上限である．よって (3) は成立する．

さらに，$G_t = \max_i |V_t - V_\infty|$ とすると，全ての i について $V_t(i)$ は $V_\infty(i)$ に概収束するので，G_t は 0 に概収束する正の乱数系列である．ここで，

$$|u_t(i)| \leq (1+\gamma) G_t \{(1+\gamma) G_t + 4K_1 + 2K_2\}$$

という関係が成立するので，(4) が成立する．

以上から，もしステップサイズが定理中で述べた条件を満たすならば，全ての i について $v_t(i)$ は $v^\pi(i)$ に概収束する．

Q.E.D.

C. Variance Penalized 政策勾配定理

全ての $s \in \mathcal{S}, a \in \mathcal{A}$ について政策 $\pi(s,a)$ が $\Delta\pi(s,a)$ だけ変化したとする．この時，平均価値に関するベルマン方程式を用いると，状態 s の平均価値の変化量 $\Delta V(s)$ は以下の条件を満たす．

$$
\begin{aligned}
\Delta V(s) &= \sum_a \Delta\pi(s,a) Q^\pi(s,a) + \gamma \sum_a \pi(s,a) \sum_{s'} \mathcal{P}^a_{ss'} \Delta V(s') \\
&= \Delta Q(s) + \gamma \sum_{s'} \mathcal{P}^\pi_{ss'} \Delta V(s') \qquad \text{(C.1)}
\end{aligned}
$$

ここで，$\Delta Q(s) = \sum_a \Delta\pi(s,a) Q^\pi(s,a)$ である．同様に分散価値に関するベルマン方程式を用いると，状態 s の分散価値の変化量 $\Delta v(s)$ は以下の条件を満たす．

$$
\begin{aligned}
\Delta v(s) &= \sum_a \Delta\pi(s,a) q^\pi(s,a) + \gamma^2 \sum_a \pi(s,a) \sum_{s'} \mathcal{P}^a_{ss'} \Delta v(s') \\
&\quad + 2\gamma^2 \sum_a \pi(s,a) \sum_{s'} \mathcal{P}^a_{ss'} (\mathcal{R}^a_{ss'} + \gamma V^\pi(s') - V^\pi(s)) \Delta V(s') \\
&= \Delta q(s) + \gamma^2 \sum_{s'} \mathcal{P}^\pi_{ss'} \Delta v(s') \\
&\quad + 2\gamma^2 \sum_{s'} \mathcal{P}^\pi_{ss'} (\mathcal{R}^\pi_{ss'} + \gamma V^\pi(s') - V^\pi(s)) \Delta V(s') \qquad \text{(C.2)}
\end{aligned}
$$

ただし，$\Delta q(s) = \sum_a q^\pi(s,a) \Delta\pi(s,a)$ である．ここで，式 (C.1),(C.2) を全ての s について集め，行列表現を用いると，

$$
\begin{aligned}
\mathbf{\Delta V} &= \mathbf{\Delta Q} + \gamma \mathbf{P}^\pi \mathbf{\Delta V} \\
\mathbf{\Delta v} &= \mathbf{\Delta q} + \gamma^2 \mathbf{P}^\pi \mathbf{\Delta v} + 2\gamma^2 \mathbf{E} \mathbf{\Delta V}
\end{aligned}
$$

ここで，$\mathbf{\Delta Q}$, $\mathbf{\Delta q}$ は i 番目の要素がそれぞれ $\Delta Q(i)$, $\Delta q(i)$ に相当する $|\mathcal{S}|$ 次元ベクトルであり，$\mathbf{E}$ は (i,j) 要素が，$\mathcal{P}^\pi_{ij}(\mathcal{R}^\pi_{ij} + \gamma V^\pi(j) - V^\pi(i))$ に相当する $|S| \times |S|$ 行列である．

このとき，評価関数 f の変化量 Δf を $\Delta f = \Delta f_V - \alpha \Delta f_v$ と分解すると，以下の関係が成立する．

$$\begin{aligned}\Delta f_V &= \mathbf{s}'_0(\mathbf{I}-\gamma\mathbf{P}^\pi)^{-1}\mathbf{\Delta Q},\\ \Delta f_v &= \mathbf{s}'_0(\mathbf{I}-\gamma^2\mathbf{P}^\pi)^{-1}(\mathbf{\Delta q}+2\gamma^2\mathbf{E}(\mathbf{I}-\gamma\mathbf{P}^\pi)^{-1}\mathbf{\Delta Q})\\ &= \mathbf{s}'_0(\mathbf{I}-\gamma^2\mathbf{P}^\pi)^{-1}\mathbf{\Delta q} \qquad \text{(C.3)}\\ &\quad + 2\gamma^2\mathbf{s}'_0(\mathbf{I}-\gamma^2\mathbf{P}^\pi)^{-1}\mathbf{E}(\mathbf{I}-\gamma\mathbf{P}^\pi)^{-1}\mathbf{\Delta Q} \text{ (C.4)}\end{aligned}$$

ここで更に Δf_v を式 (C.3) と式 (C.4) に分解し，$\Delta f_v = \Delta f_{v1} + \Delta f_{v2}$ と表す．そして逆行列を $(\mathbf{I}-\alpha\mathbf{P})^{-1} = (\mathbf{I}+\alpha\mathbf{P}+\alpha^2\mathbf{P}+\cdots)$ という関係を使って展開すると以下の関係を得る．

$$\begin{aligned}\Delta f_V &= \mathbf{s}'_0(\mathbf{I}+\gamma\mathbf{P}+\gamma^2\mathbf{P}^2+\cdots)\mathbf{\Delta Q},\\ \Delta f_{v1} &= \mathbf{s}'_0(\mathbf{I}+\gamma^2\mathbf{P}+\gamma^4\mathbf{P}^2+\cdots)\mathbf{\Delta q}\\ \Delta f_{v2} &= 2\gamma^2\mathbf{s}'_0(\mathbf{I}+\gamma^2\mathbf{P}+\gamma^4\mathbf{P}^2+\cdots)\mathbf{E}\\ &\qquad (\mathbf{I}+\gamma\mathbf{P}+\gamma^2\mathbf{P}^2+\cdots)\mathbf{\Delta Q}\end{aligned}$$

ここで $\mathbf{s}'_0\mathbf{P}^{n-1}$ は，i 番目の要素が $\Pr(s_{t+n}=i|\mathbf{s}'_0,\pi)$ を意味する $|\mathcal{S}|$ 次元ベクトルの転置ベクトルなので，Δf_V と Δf_{v1} に関しては以下の関係が成立する．

$$\begin{aligned}\Delta f_V &= \sum_{t=0}^{\infty}\sum_s \Pr(s_t=s|\mathbf{s}_0,\pi)\gamma^t\Delta Q(s)\\ \Delta f_{v1} &= \sum_{t=0}^{\infty}\sum_s \Pr(s_t=s|\mathbf{s}_0,\pi)\gamma^{2t}\Delta q(s)\end{aligned}$$

また Δf_{v2} に関しては，以下のように展開できる．

$$\begin{aligned}&\frac{\Delta f_{v2}}{2\gamma^2}\\ &= \mathbf{s}'_0\mathbf{E}\mathbf{\Delta Q}+\gamma\mathbf{s}'_0\mathbf{EP}\mathbf{\Delta Q}+\gamma^2\mathbf{s}'_0\mathbf{EP}^2\mathbf{\Delta Q}+\cdots\\ &\qquad +\gamma^2\mathbf{s}'_0\mathbf{PE}\mathbf{\Delta Q}+\gamma^3\mathbf{s}'_0\mathbf{PEP}\mathbf{\Delta Q}+\cdots\\ &\qquad\qquad +\gamma^4\mathbf{s}'_0\mathbf{P}^2\mathbf{E}\mathbf{\Delta Q}+\cdots\end{aligned}$$

これをまとめると，

$$\Delta f_{v2} = 2\sum_{t=1}^{\infty}\sum_s \Pr(s_t=s|\mathbf{s}_0,\pi)e(s,t)\Delta Q(s)$$

となる．ただし，

$$e(s,t) = E\{\sum_{\tau=1}^{t}\gamma^{\tau+t}(r_t+\gamma V^\pi(s_\tau)-V^\pi(s_{\tau-1}))|s_t=s,\mathbf{s}_0,\pi\}.$$

最後に，$\Delta\pi(s,a) = \sum_l \frac{\partial\pi(s,a)}{\partial\theta_l}\Delta\theta_l$ であるので，以下の勾配定理が導かれる．

$$\begin{aligned}\frac{\partial f_V}{\partial\theta} &= \sum_{t=0}^{\infty}\sum_s \Pr(s_t=s|\mathbf{s}_0,\pi)\gamma^t\sum_a Q^\pi(s,a)\frac{\partial\pi(s,a)}{\partial\theta},\\ \frac{\partial f_{v1}}{\partial\theta} &= \sum_{t=0}^{\infty}\sum_s \Pr(s_t=s|\mathbf{s}_0,\pi)\gamma^{2t}\sum_a q^\pi(s,a)\frac{\partial\pi(s,a)}{\partial\theta},\\ \frac{\partial f_{v2}}{\partial\theta} &= 2\sum_{t=1}^{\infty}\sum_s \Pr(s_t=s|\mathbf{s}_0,\pi)e(s,t)\sum_a Q^\pi(s,a)\frac{\partial\pi(s,a)}{\partial\theta}.\end{aligned}$$

Q.E.D.

D. 機械整備問題のパラメータ

我々が用いた状態遷移確率は以下の通りである．ただし，T は状態遷移関数，$RMean$ は報酬の期待値関数，$RVar$ は報酬の分散関数である．

ここで，例えば状態 1 (不調) において整備行動を選択した場合に状態 0 (正常) に遷移する確率は 50%となる．本論文の設定では利得の期待値と分散のトレードオフを強調するため，通常の直観よりも整備行動の魅力を下げた設定にしている．

また，下の a は 0 に近い小さな正の実数を，b は 1 に近い 1 未満の実数を表している．これは，状態遷移のエルゴード性を維持しつつ，故障状態を終端状態ととらえることを可能にするためである．即ち，状態 2 に遷移してしまうと，状態 2 へのセルフループを長い期間繰り返す．その結果，何の報酬も入らない状態遷移が長く続くことになるので，状態 2 の報酬の期待値と分散がともに 0 に近くなる．よって，状態 2 を近似的に終端状態と見なすことができる．

直接報酬のノイズに関しては一様分布を仮定し，パラメータ k によって特徴づけた．5%ノイズの場合は $k=0.05$ を，50%ノイズの場合は $k=0.5$ を代入すればよい．

$$T(\text{使}) = \begin{pmatrix}0.1 & 0.0 & a\\ 0.9 & 0.9 & 0.0\\ 0.0 & 0.1 & b\end{pmatrix}$$

$$T(\text{整}) = \begin{pmatrix}0.95 & 0.50 & a\\ 0.05 & 0.45 & 0.0\\ 0.00 & 0.05 & b\end{pmatrix}$$

$$RMean(\text{使}) = \begin{pmatrix}5.0 & 0.0 & 0.0\\ 5.0 & 5.0 & 0.0\\ 0.0 & -25.0 & 0.0\end{pmatrix}$$

$$RMean(\text{整}) = \begin{pmatrix}-1.0 & -1.0 & 0.0\\ -1.0 & -1.0 & 0.0\\ 0.0 & -25.0 & 0.0\end{pmatrix}$$

$$RVar(\text{使}) = \begin{pmatrix}8.33k^2 & 8.33k^2 & 0.0\\ 8.33k^2 & 8.33k^2 & 0.0\\ 0.0 & 208.33k^2 & 0.0\end{pmatrix}$$

$$RVar(\text{整}) = \begin{pmatrix}0.33k^2 & 0.33k^2 & 0.0\\ 0.33k^2 & 0.33k^2 & 0.0\\ 0.0 & 208.33k^2 & 0.0\end{pmatrix}$$

著 者 紹 介

佐藤 誠(正会員)

1996 年 3 月 東京工業大学工学部制御システム工学科卒業．1998 年 3 月 同大学総合理工学研究科知能システム科学専攻修士課程修了．同年 4 月株式会社東芝入社．主として，強化学習，ニューラルネットワーク，データマイニングに関する研究に従事．情報処理学会会員．

木村 元(正会員)

1992 年東京工業大学工学部制御工学科卒業．1994 年同大学大学院総合理工学研究科知能科学専攻修士課程修了．1997 年同大学大学院博士課程修了，同年 4 月日本学術振興会 PD 研究員，1998 年 4 月，東京工業大学大学院総合理工学研究科助手，現在に至る．人工知能，特に強化学習に関する研究に従事．計測自動制御学会，日本ロボット学会各会員．

小林 重信(正会員)

1974 年東京工業大学大学院博士課程経営工学専攻修了．同年 4 月，同大学工学部制御工学科助手．1981 年 8 月，同大学大学院総合理工学研究科助教授．1990 年 8 月，教授．現在に至る．問題解決と推論制御，知識獲得と学習などの研究に従事．計測自動制御学会，情報処理学会各会員．