# 強化学習における多次元入出力の扱いとロボットへの適用

九州大学　大学院工学研究院
木村　元

RACOT研究会
2007. 03. 16
講演資料

# 強化学習における多次元入出力の扱いとロボットへの適用

## 発表の流れ

（1） 強化学習とは？

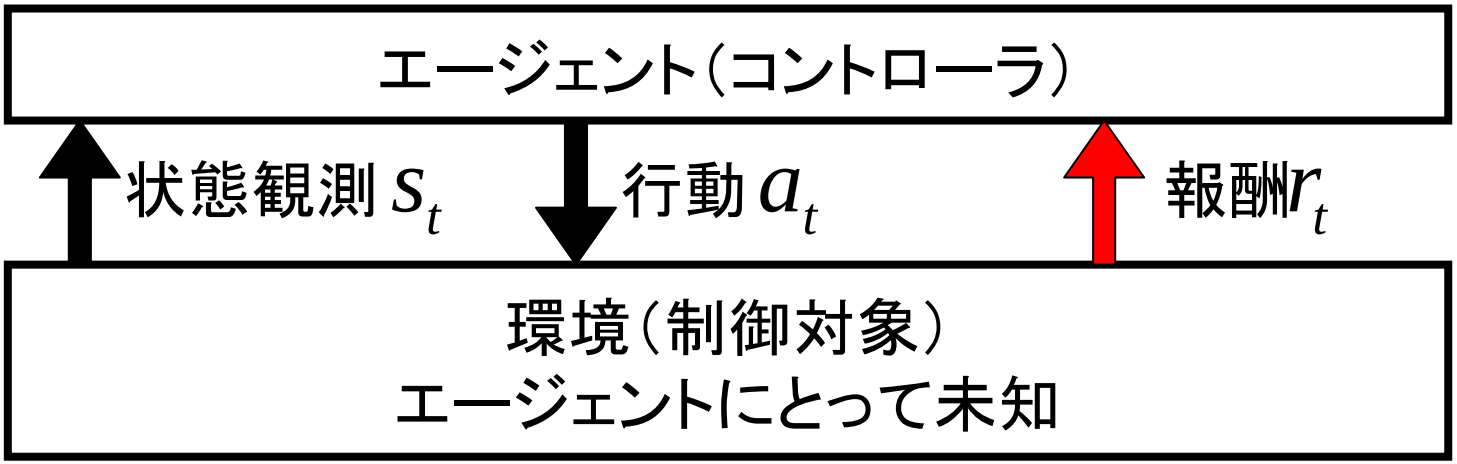
（2） 確率的政策を**確率的勾配法**によって改善していく強化学習法
（自然勾配Actor-Critic法）

（3） **Q-learning アルゴリズム**を多次元状態-行動空間へ拡張
（ランダムタイリングによる多次元空間の関数近似と
Gibbsサンプリングによる行動選択を組合わせた強化学習）

# 強化学習とは？

試行錯誤を通じて環境に適応する学習制御の枠組み

生体の「脳」のシステムを模倣




- 状態観測 → 行動選択 →（状態遷移）→ 報酬　　繰り返し
- 何回か状態遷移した後，やっと報酬を得る
  → 多段決定過程 （報酬に遅れ）
- 報酬合計が最大になる行動を探索する

# 最適化問題としての強化学習

- 試行錯誤を繰り返し，より多くの報酬を得る振舞いを獲得
- 振舞い＝「状態入力→行動出力」の写像パターン
  → 報酬合計の最大化を目指す最適化問題

- 「試行錯誤による学習」 ＝ 最適化のための探索過程
- 「学習による適応」 ＝ 最適化の結果，報酬合計が最大

強化学習アルゴリズム ＝ 報酬合計の最適化手法

# 強化学習の理論的特徴

- ●状態遷移に不確実性を伴う制御問題を理論的に扱う
- ●離散的な状態遷移も含んだ段取り的な制御も理論的に扱う

→　環境を確率過程（マルコフ決定過程）でモデル化

---

# 応用上の特徴

「何をすべきか」を「報酬」によって簡単に指示するだけで
「どのように実現するか」という制御規則を学習により自動的に獲得

1）制御プログラミングの自動化・省力化
2）ハンドコーディングよりも優れた解：
　特に不確実な要素（摩擦やガタ，振動，誤差など）や計測困難な未知パラメータが多い場合に有利
3）自律性と想定外の環境変化への対応：
　通信が物理的に困難だったり現象のダイナミクスが人間にとって早過ぎる場合や，機械故障など急激な変化やプラント経年変化など予め想定しておくことが困難な環境の変化に対し自動的に追従

# 強化学習における多次元入出力の扱いとロボットへの適用

## 発表の流れ

（1） 強化学習とは？

（2） 確率的政策を**確率的勾配法**によって改善していく強化学習法
（自然勾配Actor-Critic法）

（3） **Q-learning アルゴリズム**を多次元状態-行動空間へ拡張
（ランダムタイリングによる多次元空間の関数近似と
Gibbsサンプリングによる行動選択を組合わせた強化学習）

Off-policy 学習法：

政策に依存せずに

**最適な状態-行動評価値（Q値）を推定**

Q-learning

**最適性：** ○

ただし、あらゆる状態-行動の経験を
十分に行う必要あり

On-policy 学習法：

政策に依存した状態評価値や
行動評価値を推定して**政策を改善する**

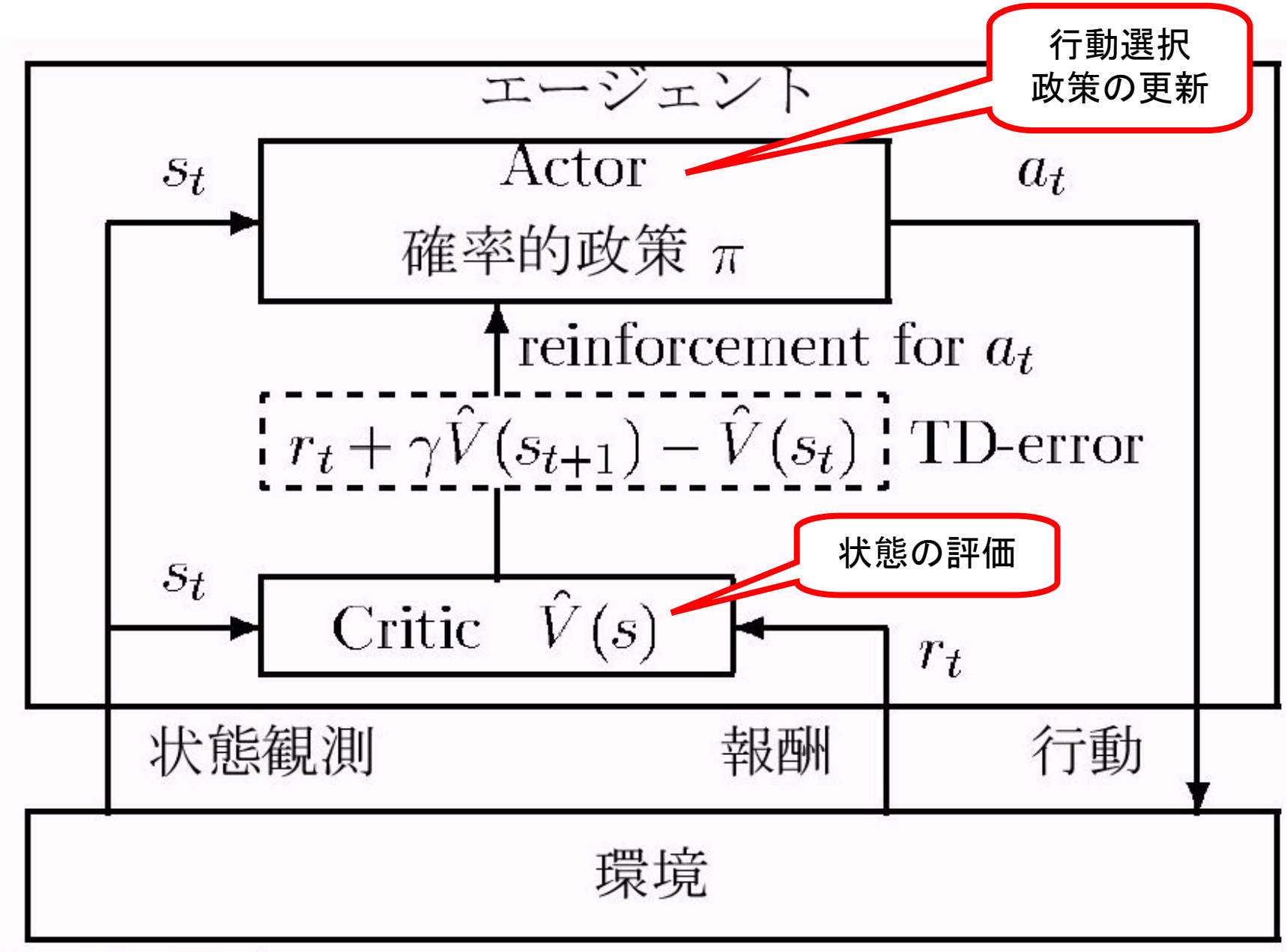
SARSA

Actor-Critic

**最適性：** ×

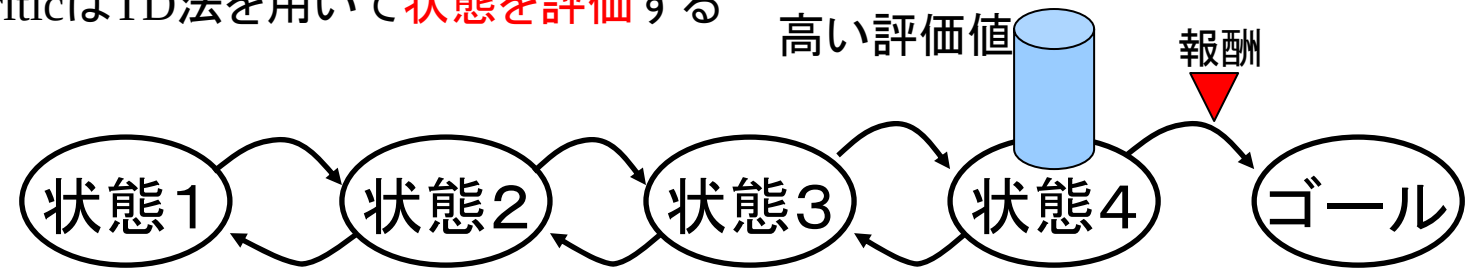
局所解へ陥る危険はあるが、
膨大な状態-行動空間では
学習速度の点で有望

# Actor-Critic法の構成

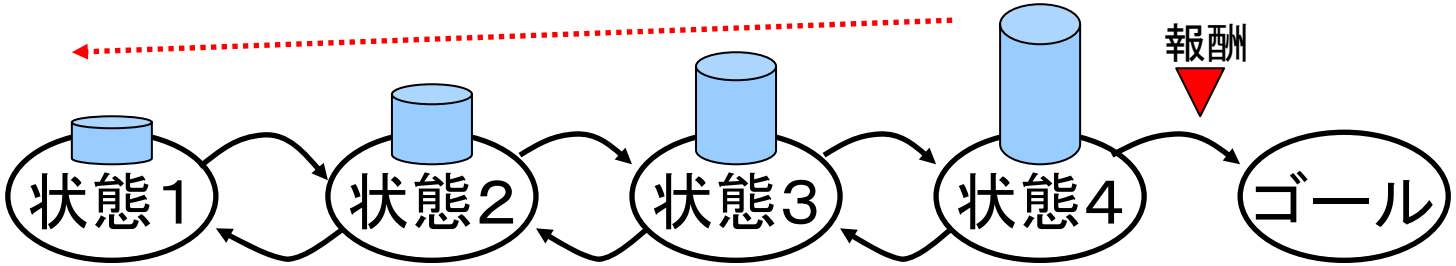


# Actor-Critic法による学習の仕組み

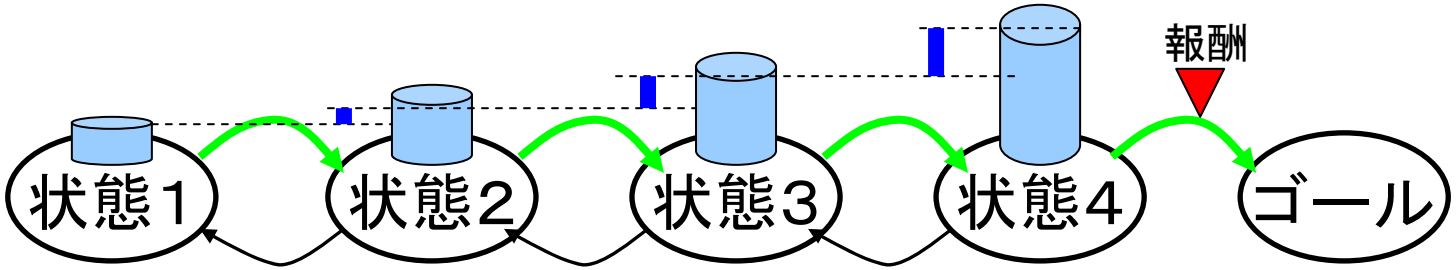
1）CriticはTD法を用いて状態を評価する




状態遷移を繰り返すうちにcriticは状態の評価値を伝播して学習する




2）ActorはTD-errorを手がかりに行動を学習
TD-errorが大きな正の値になる状態へ遷移する行動は良い行動＝選択確率を高く




**割引報酬合計を最大化する方向へ政策パラメータを更新する勾配法**

# キーワード1：Actor の適正度（eligibility）

Actor-Criticアルゴリズムでは、政策パラメータ $\theta$ を以下のように更新：

$$\Delta\theta = (\text{TD-error})\frac{\partial}{\partial\theta}\ln\pi(a|\theta,s)$$

状態

適正度

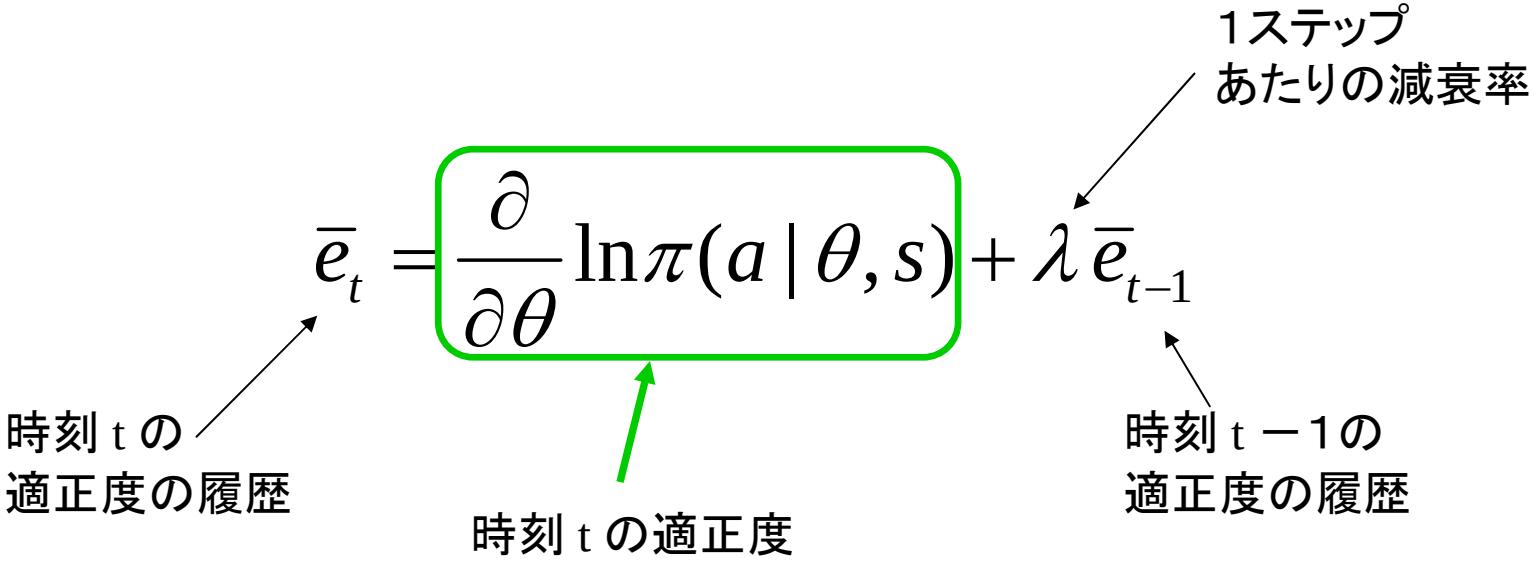
選択した行動 $a$ の確率（密度）
すなわち確率的政策

$$\frac{\partial}{\partial\theta}\ln\pi(a|\theta,s) = \frac{1}{\pi(a|\theta,s)}\frac{\partial}{\partial\theta}\pi(a|\theta,s)$$

行動選択確率の違いで行動の強化に偏りがおきないように補正する効果

# キーワード2：適正度の履歴（eligibility trace）

適正度を1ステップあたり減衰率 λ で割引いて合計したもの




**適正度の履歴による効果：**

1） 非マルコフ環境において学習可能

2） Critic の関数近似がいいかげんでもActorの学習が可能

# 政策勾配定理（1）（Sutton 2000）

$$d^{\pi}(s) = \sum_{t=0}^{\infty} \gamma^{t} \Pr(s_t = s \mid s_0, \pi)$$

割引報酬を改善する
政策パラメータの方向

状態訪問頻度

$$\frac{\partial}{\partial \theta} V^{\pi}(s_0) = \sum_{s \in S} d^{\pi}(s) \sum_{a \in A} \frac{\partial \pi(s,a)}{\partial \theta} \mathrm{Q}^{\pi}(s,a)$$

政策（行動選択確率関数）
の政策パラメータによる微分

**真のQ関数**
（状態-行動評価値）

つまり、政策を改善するには、

**（1） 政策関数を政策パラメータで微分した値**

**（2） 真のQ関数**

があれば良い

連続な状態-行動空間では、真のQ関数を表現できない
関数近似で表現せざるをえない
近似表現に最も適した基底関数は？

# 政策勾配定理（2）（Sutton 2000）

$$d^{\pi}(s)=\sum_{t=0}^{\infty}\gamma^{t}\Pr(s_t=s\mid s_0,\pi)$$

状態訪問頻度

政策（行動選択確率関数）の政策パラメータによる微分

割引報酬を改善する政策パラメータの方向

$$\frac{\partial}{\partial\theta}V^{\pi}(s_0)=\sum_{s\in S}d^{\pi}(s)\sum_{a\in A}\frac{\partial\pi(s,a)}{\partial\theta}\hat{\mathrm{Q}}^{\pi}(s,a)$$

近似Q関数

状態評価値

$$\hat{\mathrm{Q}}^{\pi}(s,a)=\left(\nabla_{\theta}\ln\pi(s,a)\right)^{T}\mathbf{w}+V^{\pi}(s)$$

適正度ベクトル

重み変数

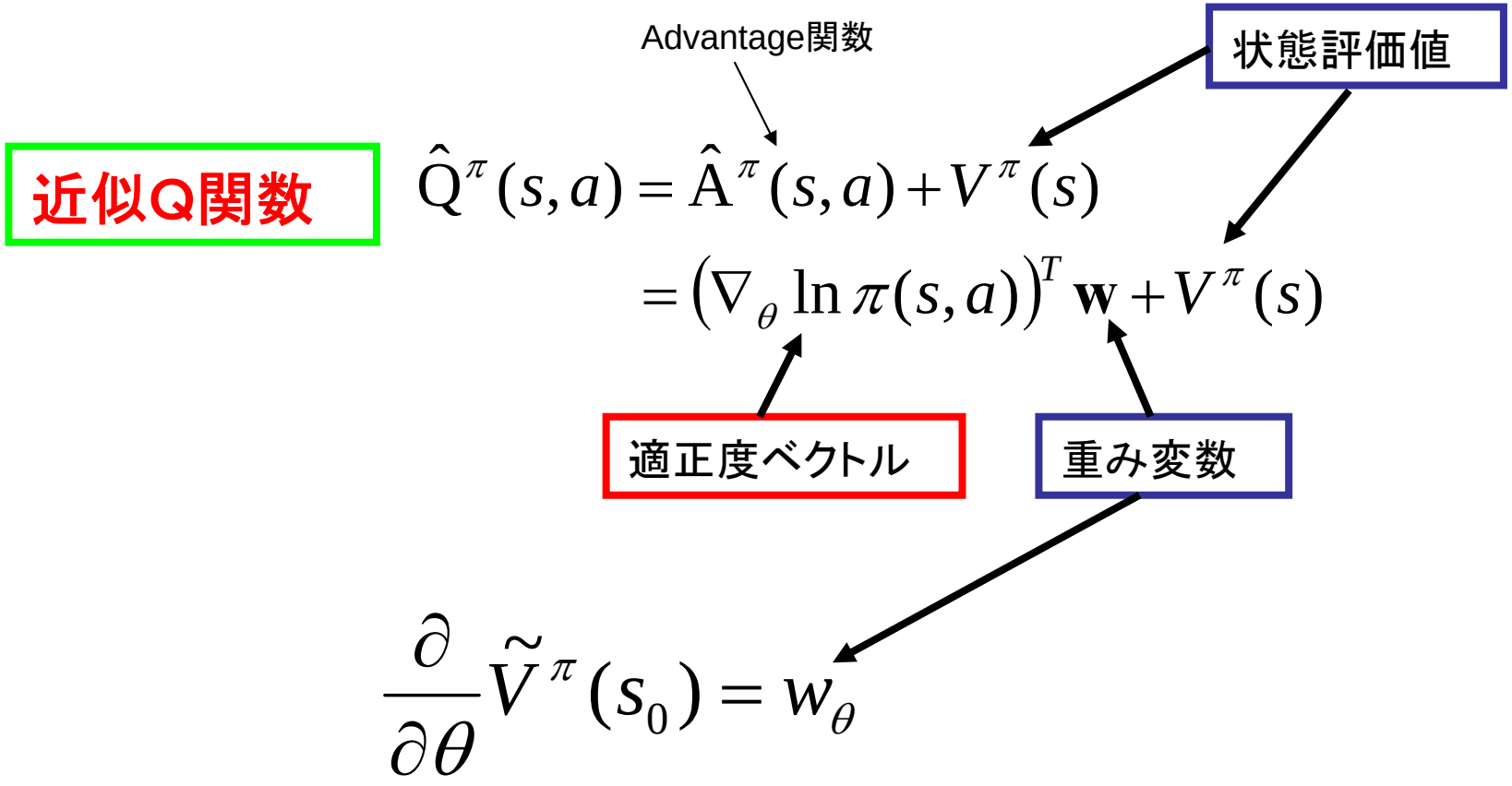
真のQ関数を用いなくても、
**適正度ベクトル**を基底とした
線形**近似関数**を用いれば
真の勾配方向へ政策を改善できる

ただし、真のQ関数に対して2乗誤差が最小の重み変数になっていると仮定

ただし、あくまでも**一次勾配**の方向のみ

# 自然勾配理論 （Kakade2002, Peters2003）

- 一次偏導関数だけでなく、2次偏導関数も考慮に入れた勾配法
- プラトー（平原）における探索停滞を回避
- 一般に2次偏導関数行列（ヘッセ行列）の逆行列計算を要するが、
  強化学習における政策パラメータに関する自然勾配では計算不要




**自然勾配 ＝ Advantage関数の重み変数ベクトル**

# 自然勾配Actor-Critic法

（1） 状態 $s_t$ を観測し、行動選択確率（密度）関数 $\pi(s_t, a)$ に従って行動 $a_t$ を実行
報酬 $r_t$ と遷移先の状態 $s_{t+1}$ を観測

（2） 適正度よりAdvantage関数を計算：

$$\hat{\mathrm{A}}^{\pi}(s_t, a_t) = \left(\nabla_{\theta} \ln \pi(s_t, a_t)\right)^T \mathbf{w}$$

政策パラメータ $\boldsymbol{\theta}$ の関数

（3） 報酬と状態評価値を用いてTD-errorを計算：

$$\left(\mathrm{TD\text{-}error}\right) = \left(r_t + \gamma \hat{V}^{\pi}(s_{t+1})\right) - \hat{V}^{\pi}(s_t)$$

（4） TD-errorを用いて状態評価値を更新：

$$\hat{V}^{\pi}(s_t) \leftarrow \hat{V}^{\pi}(s_t) + \alpha\left(\mathrm{TD\text{-}error}\right)$$

（5） TD-errorと適正度を用いてAdvantage関数のパラメータwを更新：

$$\mathbf{w} \leftarrow \mathbf{w} + \alpha\left(\mathrm{TD\text{-}error} - \hat{\mathrm{A}}^{\pi}(s_t, a_t)\right)\nabla_{\theta} \ln \pi(s_t, a_t)$$

（6） 自然勾配を用いて政策パラメータを更新： $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha_p \mathbf{w}$

# 適正度の履歴を用いる自然勾配Actor-Critic法の提案

（1） 状態 $s_t$ を観測し、行動選択確率（密度）関数 $\pi(s_t, a)$ に従って行動 $a_t$ を実行
報酬 $r_t$ と遷移先の状態 $s_{t+1}$ を観測

（2） 適正度よりAdvantage関数・Q関数を計算：

$$\hat{\mathrm{A}}^{\pi}(s,a) = \left(\nabla_{\theta} \ln \pi(s,a)\right)^{T} \mathbf{w}$$

$$\hat{Q}^{\pi}(s,a) = \hat{V}(s) + \hat{A}(s,a)$$

政策パラメータ $\boldsymbol{\theta}$ の関数

（3） 適正度の履歴を更新：

$$\bar{\mathrm{e}}_{\mathrm{t}} = \nabla_{\theta} \ln \pi(s_t, a_t) + \lambda \gamma \, \bar{e}_{t-1}$$

（4） 適正度の履歴とQ関数を用いてAdvantage関数のパラメータwを更新：

$$\delta_t = r_t + \gamma \, \hat{Q}^{\pi}(s_{t+1}, a_{t+1}) - \hat{Q}^{\pi}(s_t, a_t)$$

$$\mathbf{w} \leftarrow \mathbf{w} + \beta \delta_t \bar{e}_t$$

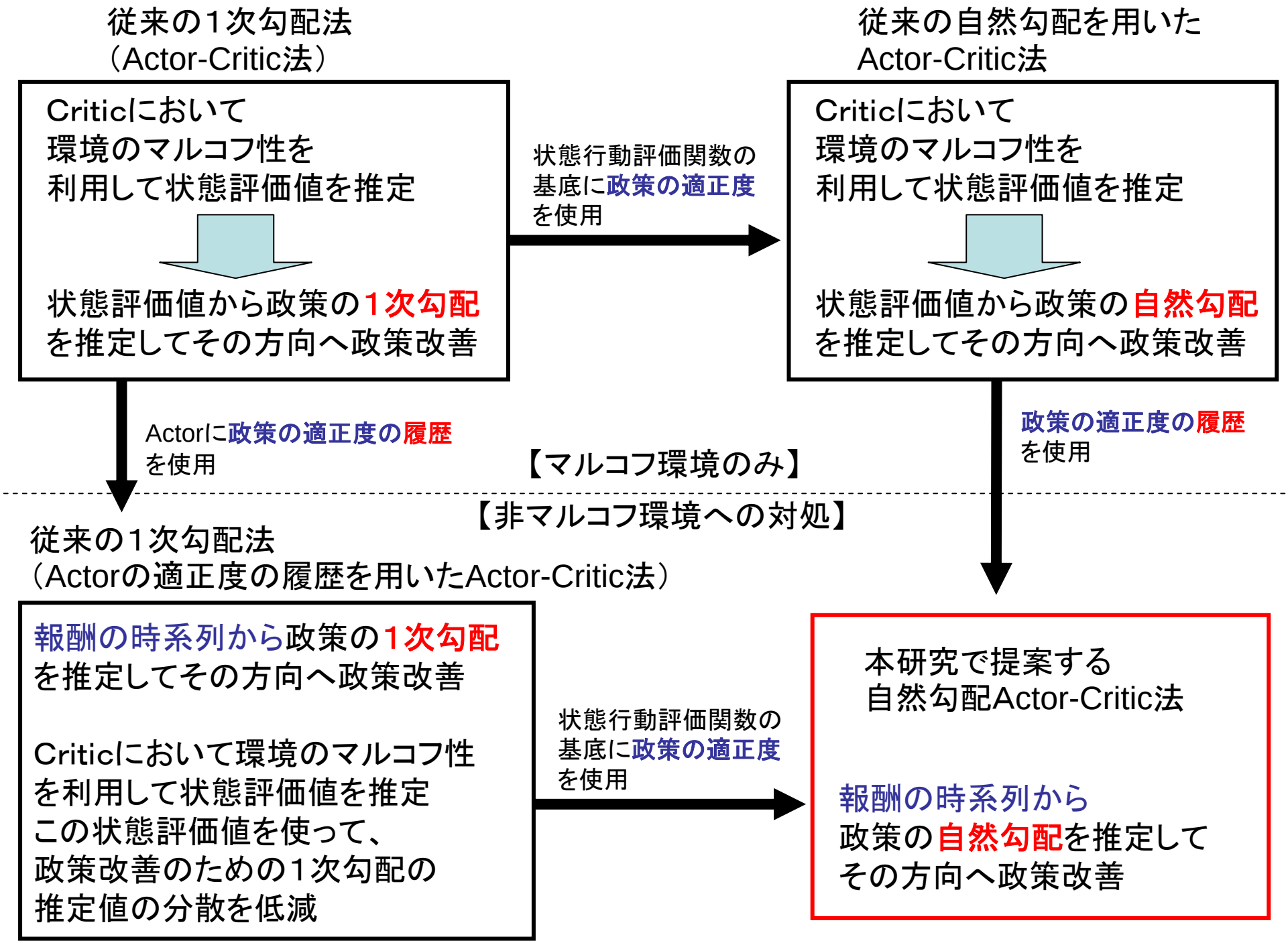
（5） 状態評価値を報酬とTD-errorによって更新：

$$\left(\mathrm{TD\text{-}error}\right) = \left(r_t + \gamma \hat{V}^{\pi}(s_{t+1})\right) - \hat{V}^{\pi}(s_t)$$

$$\hat{V}^{\pi}(s_t) \leftarrow \hat{V}^{\pi}(s_t) + \alpha\left(\mathrm{TD\text{-}error}\right)$$

SARSA（λ）に極めて類似

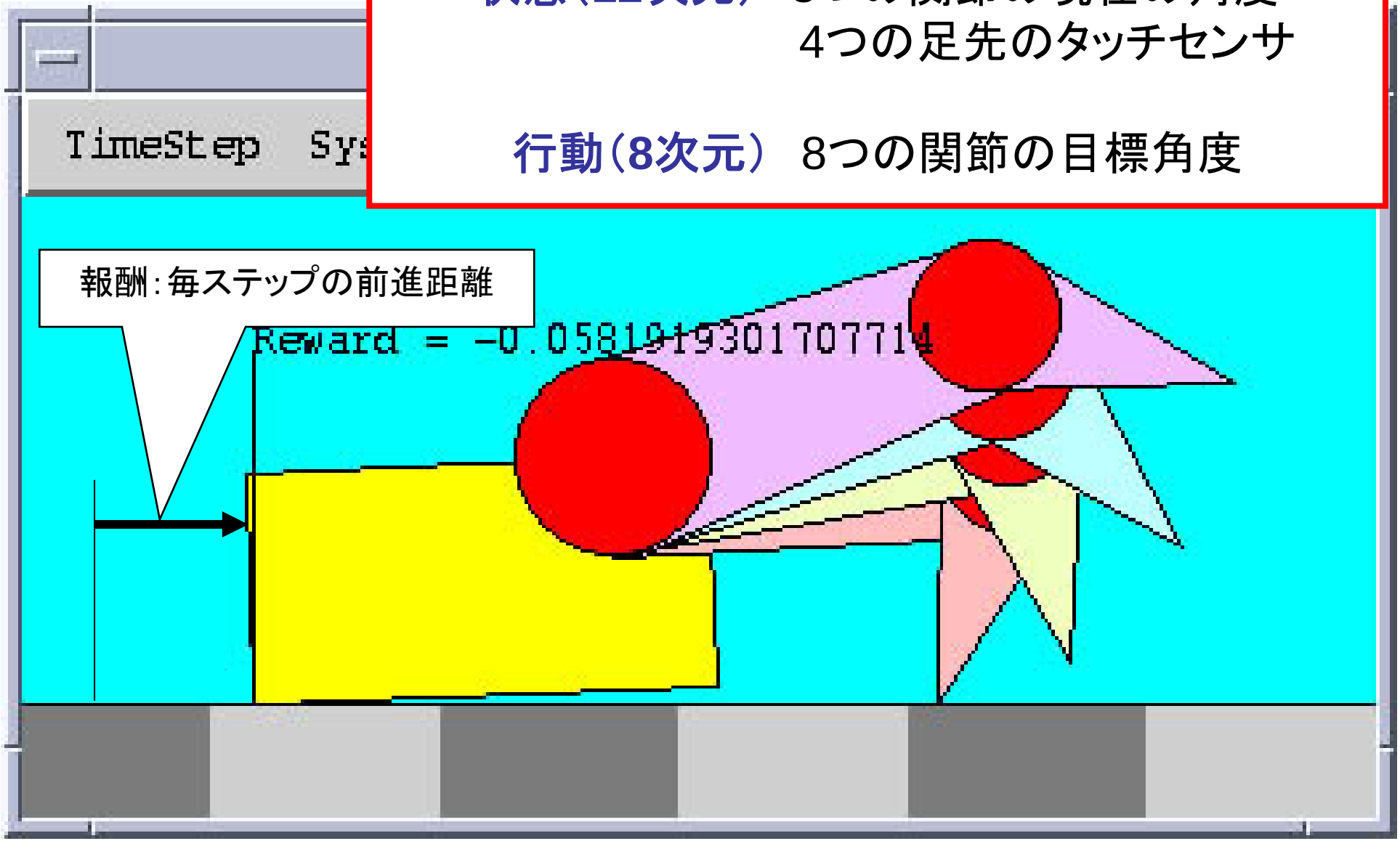
（6） 自然勾配を用いて政策パラメータを更新： $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha_p \mathbf{w}$


従来の１次勾配法
（Actor-Critic法）
Criticにおいて
環境のマルコフ性を
利用して状態評価値を推定
状態評価値から政策の１次勾配
を推定してその方向へ政策改善
状態行動評価関数の
基底に政策の適正度
を使用
従来の自然勾配を用いた
Actor-Critic法
Criticにおいて
環境のマルコフ性を
利用して状態評価値を推定
状態評価値から政策の自然勾配
を推定してその方向へ政策改善
Actorに政策の適正度の履歴
を使用
政策の適正度の履歴
を使用
【マルコフ環境のみ】
【非マルコフ環境への対処】
従来の１次勾配法
（Actorの適正度の履歴を用いたActor-Critic法）
報酬の時系列から政策の１次勾配
を推定してその方向へ政策改善
Criticにおいて環境のマルコフ性
を利用して状態評価値を推定
この状態評価値を使って、
政策改善のための１次勾配の
推定値の分散を低減
状態行動評価関数の
基底に政策の適正度
を使用
本研究で提案する
自然勾配Actor-Critic法
報酬の時系列から
政策の自然勾配を推定して
その方向へ政策改善

# シミュレーション実験： 仮想ほふくロボット

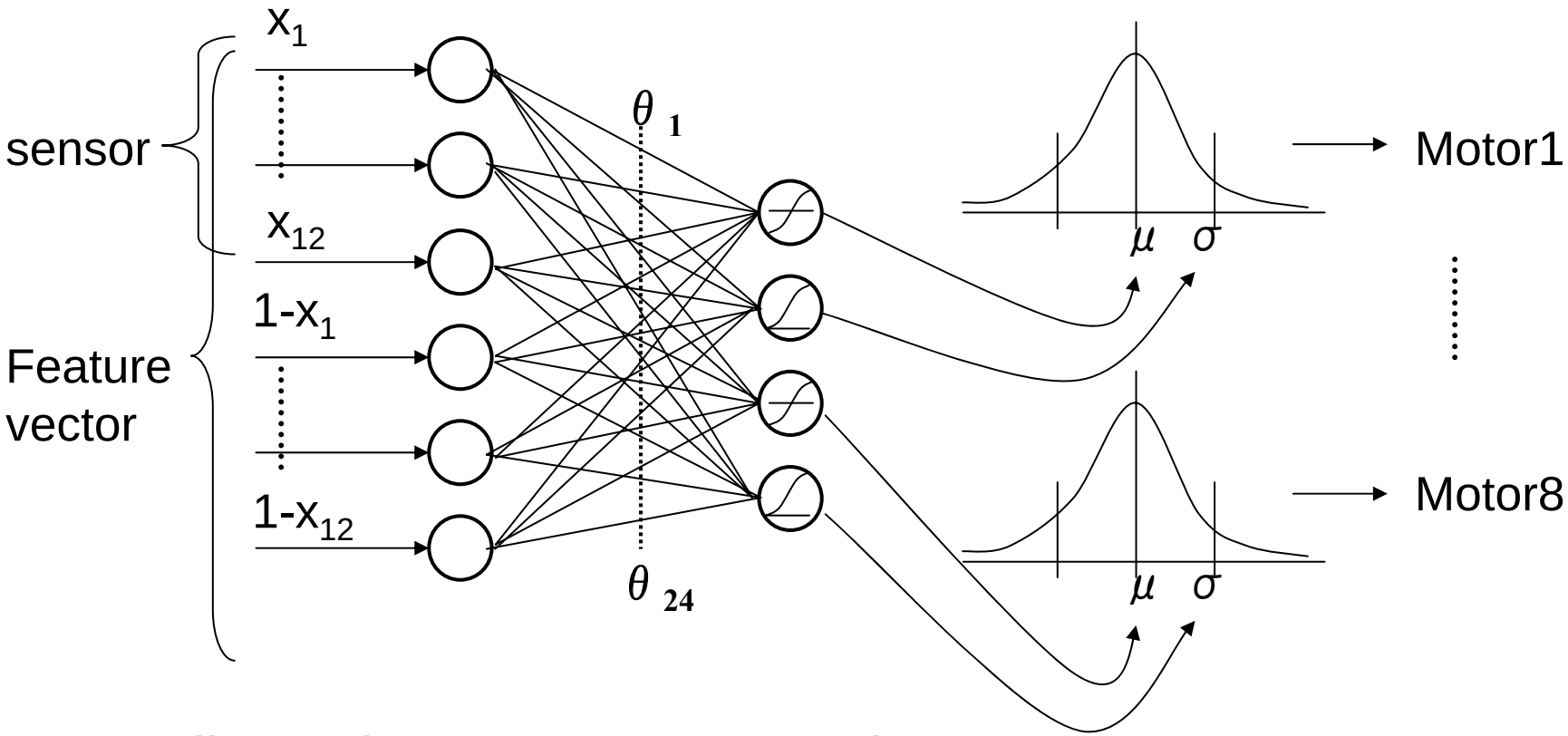
**状態（12次元）** 8つの関節の現在の角度
4つの足先のタッチセンサ

**行動（8次元）** 8つの関節の目標角度



# 【政策表現: Linear Coding】

状態価値関数の基底は
各軸3分割タイル分割



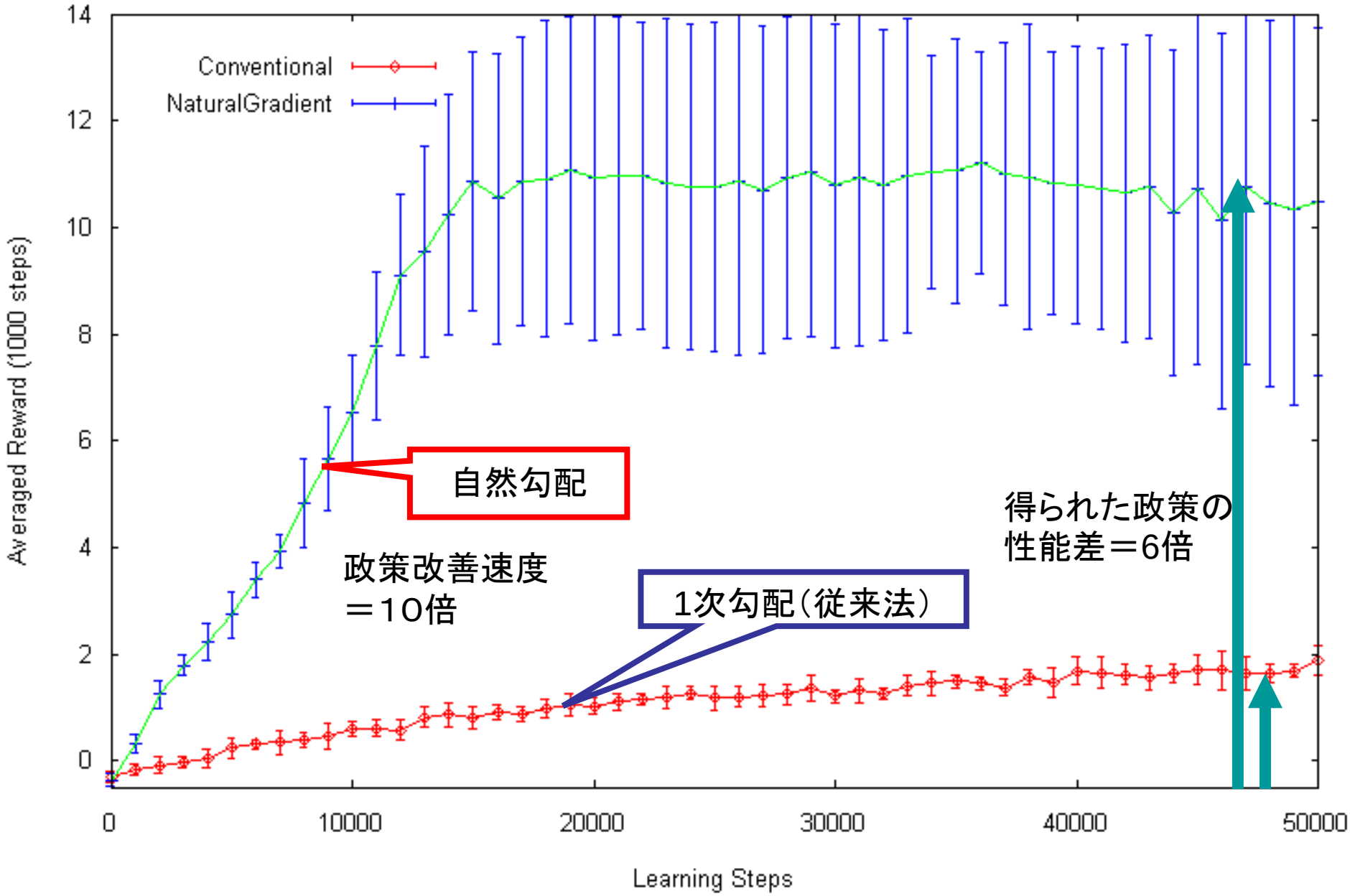

- 24-dimensional feature vector is constructed from 12-dimensional sensor's information.
- Action is sampled from Cauchy distribution N( $\mu$ , $\sigma$ ).

$$\mu_i = 1 \Big/ \left( 1 + \exp\left( -\sum_{k=1}^{6} \theta_{k,i} x_k \right) \right)$$

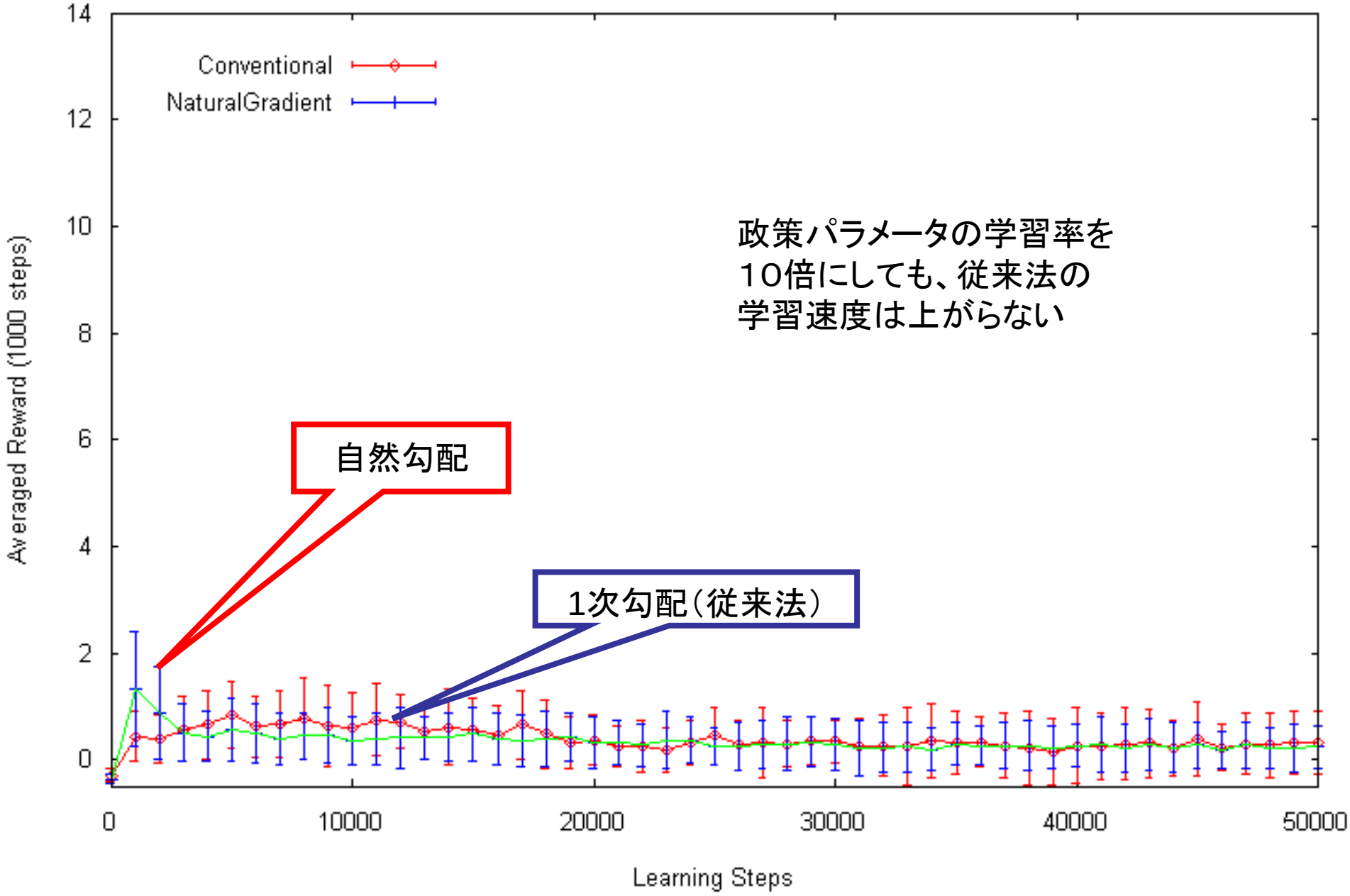
$$\sigma_i = 1 \Big/ \left( 1 + \exp\left( -\theta_{7,i} x_7 \right) \right) + 0.1$$

# シミュレーション結果

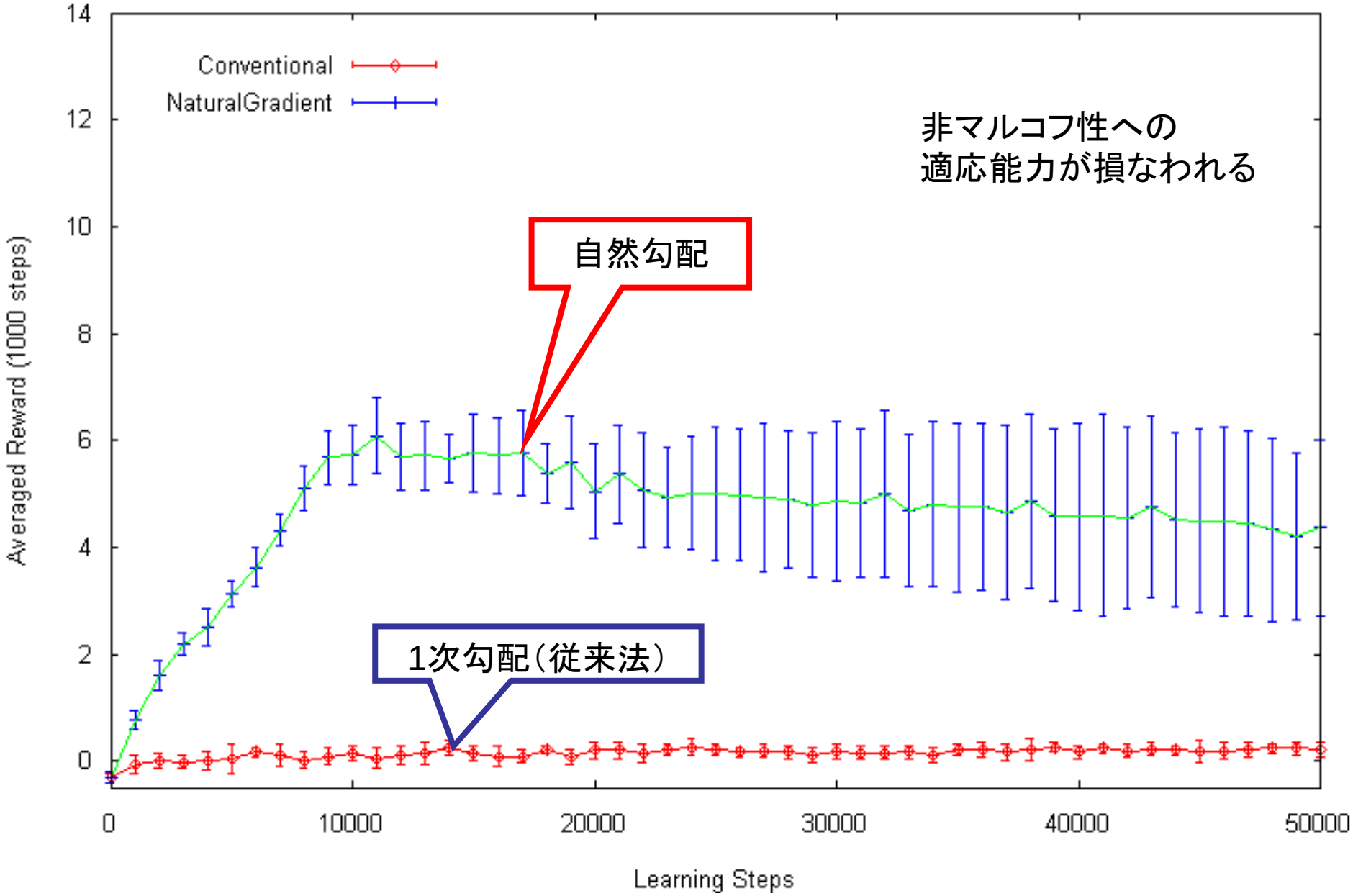
両手法とも、政策勾配を推定し、
同一の学習率にて政策を更新



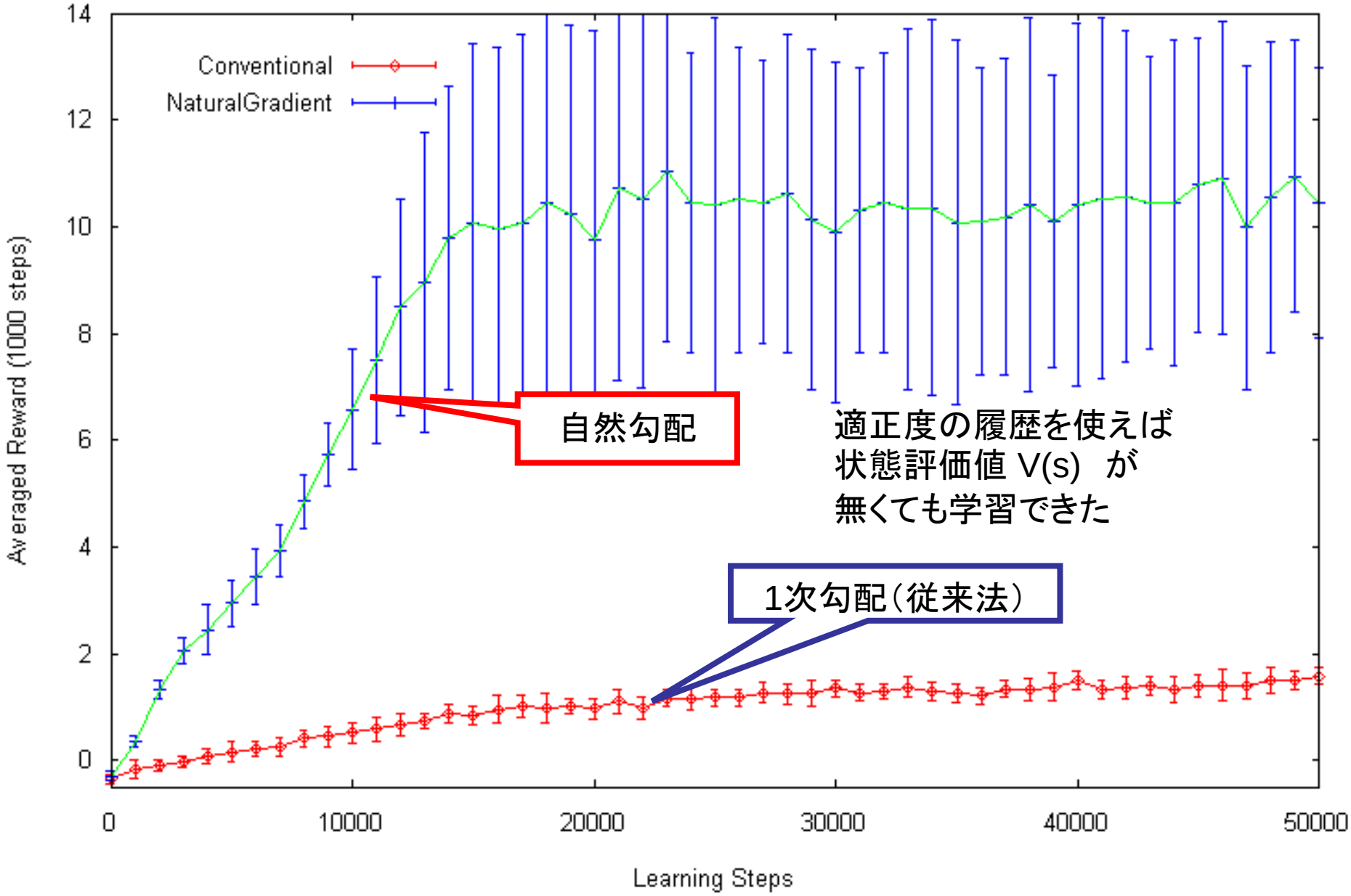
# シミュレーション結果　（学習率10倍）



# 適正度の履歴を用いない場合

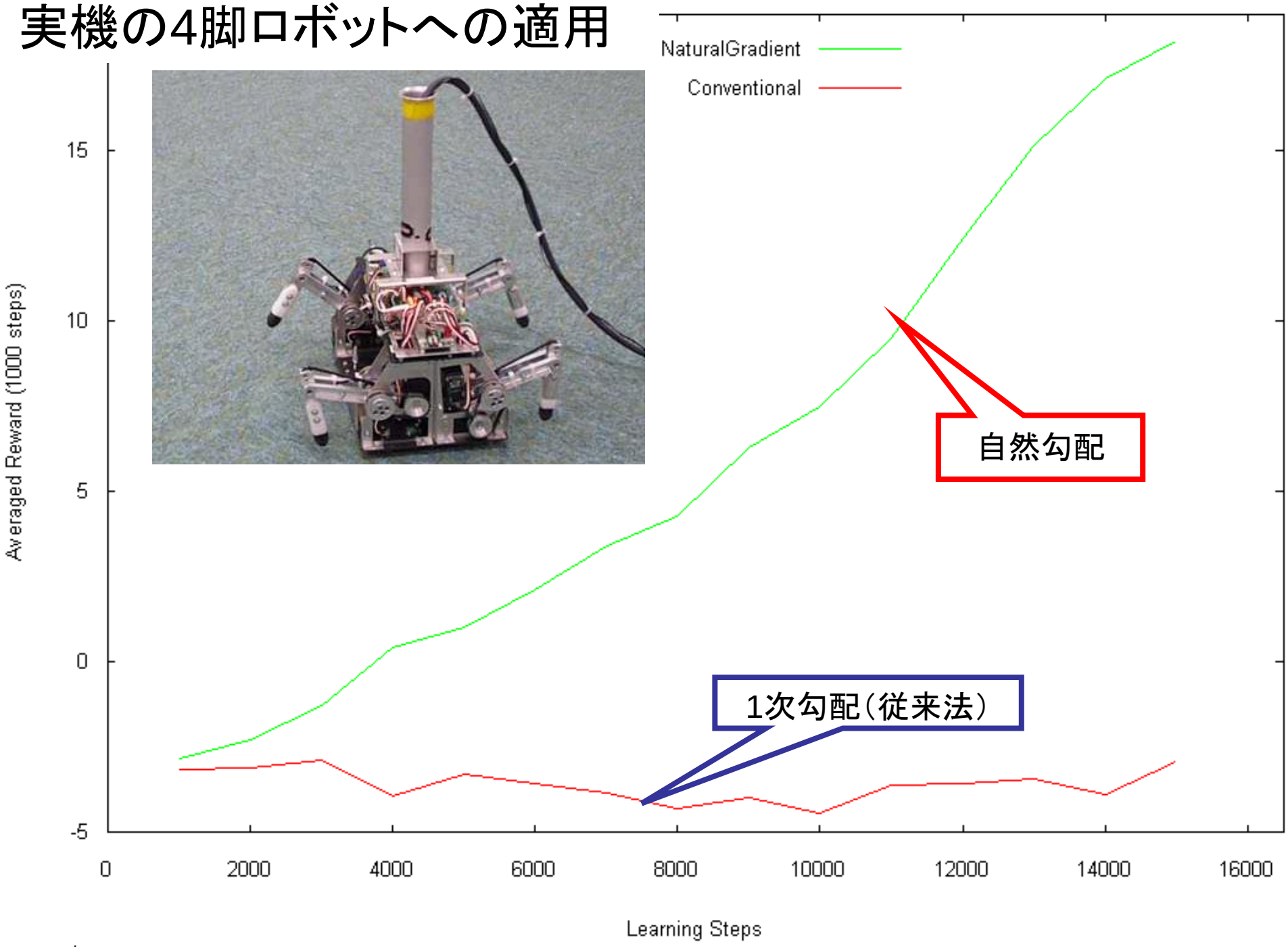


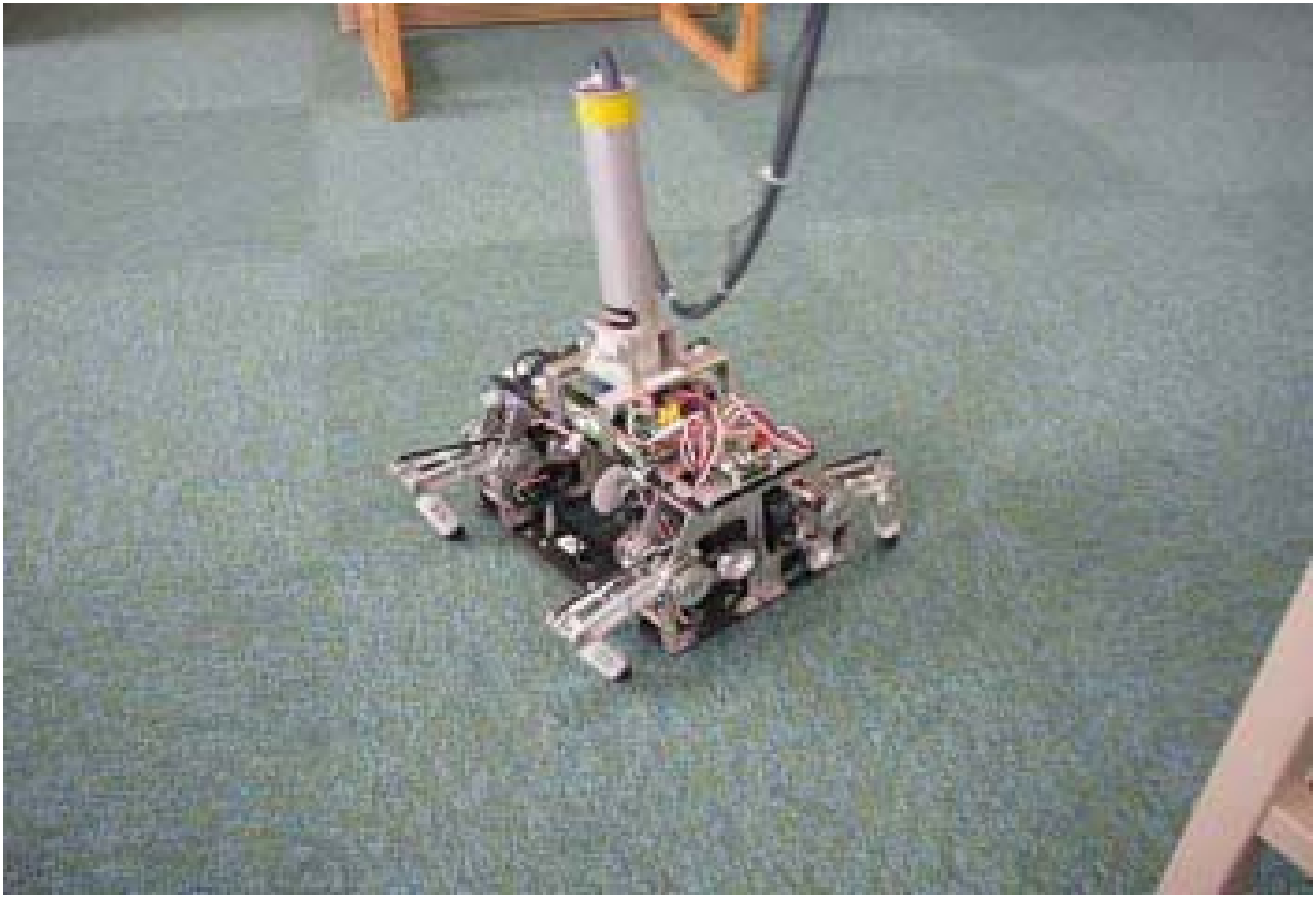
# Criticを用いない場合



# 実機の4脚ロボットへの適用



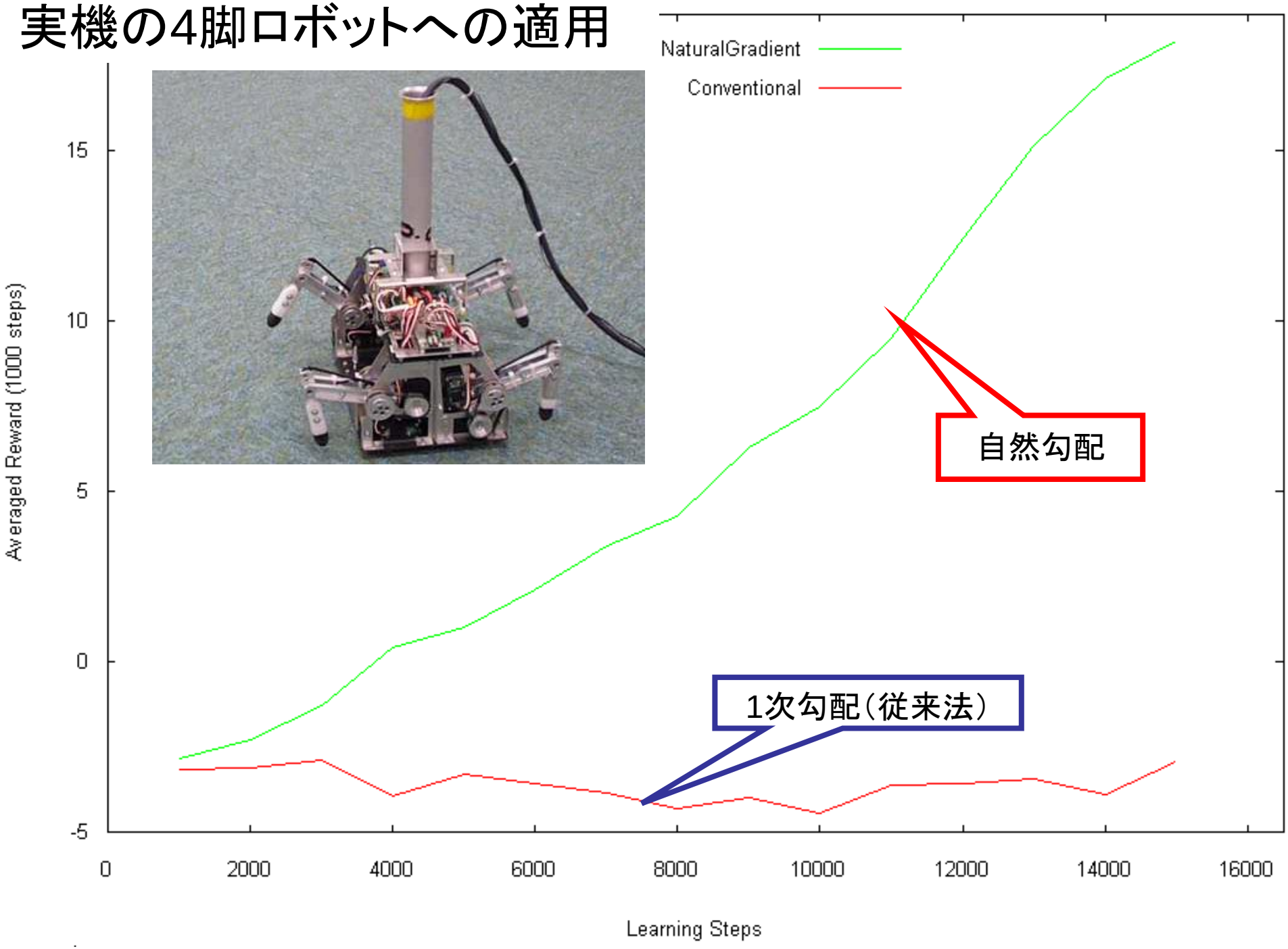


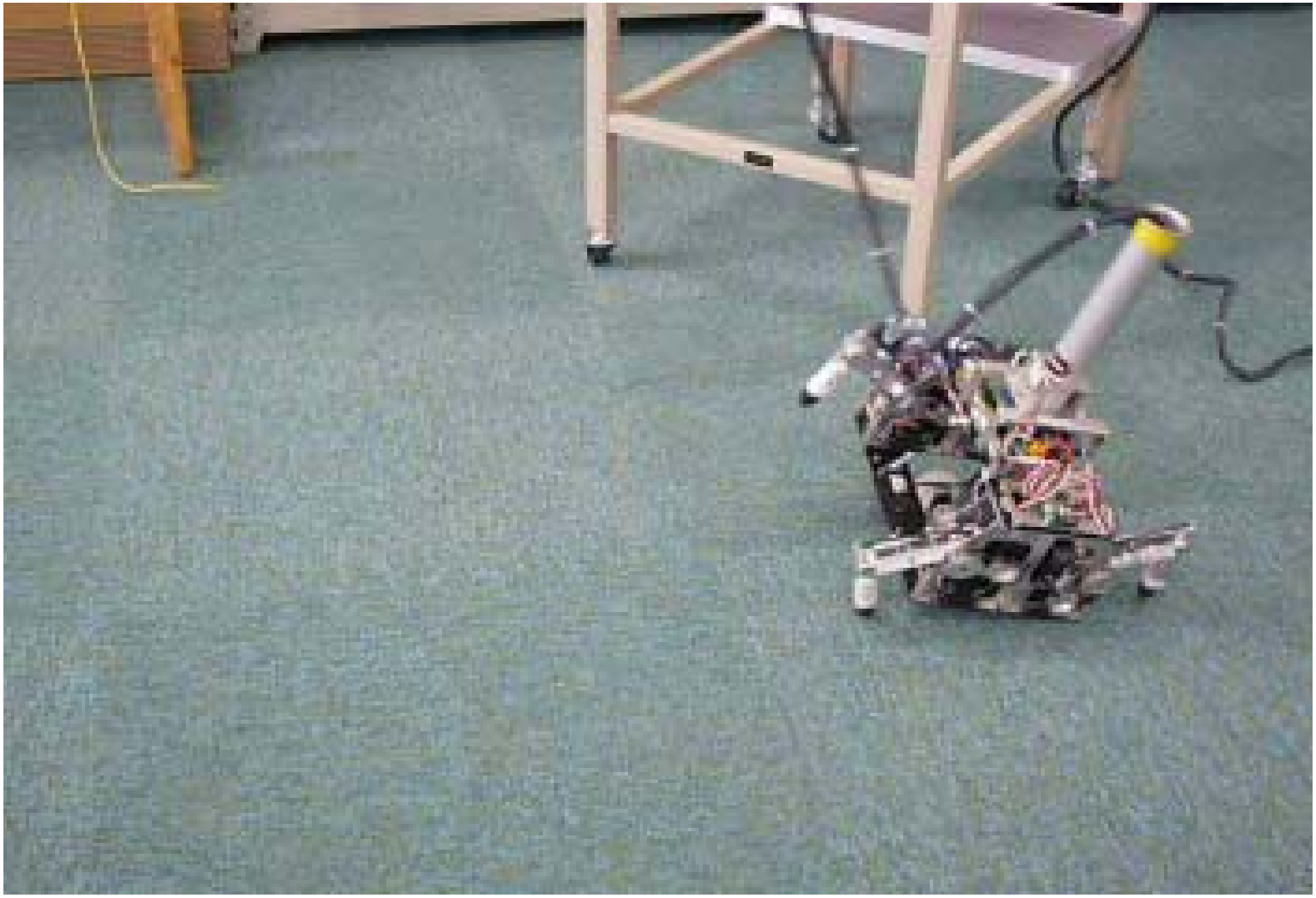


従来法（1次勾配） 15000 step 学習後

# 実機の4脚ロボットへの適用







15000 step 学習後 （約120分後）

# 確率的勾配法による接近法のまとめ

- **自然勾配Actor-Critic法に適正度の履歴を導入**した方法を提案
  SARSA（λ）に類似

- 単なる1次勾配法のActor-Critic法に比べると
  （高次元行動空間の問題において） 格段の性能向上

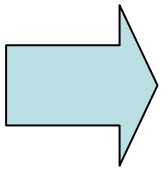
- 適正度の履歴によって**非マルコフ性**へ対処可能

- Criticが機能しなくても学習可能

- 提案手法におけるAdvantage関数の学習率 $\beta$ は、
  適正度の履歴の平均値のノルムで調節

---

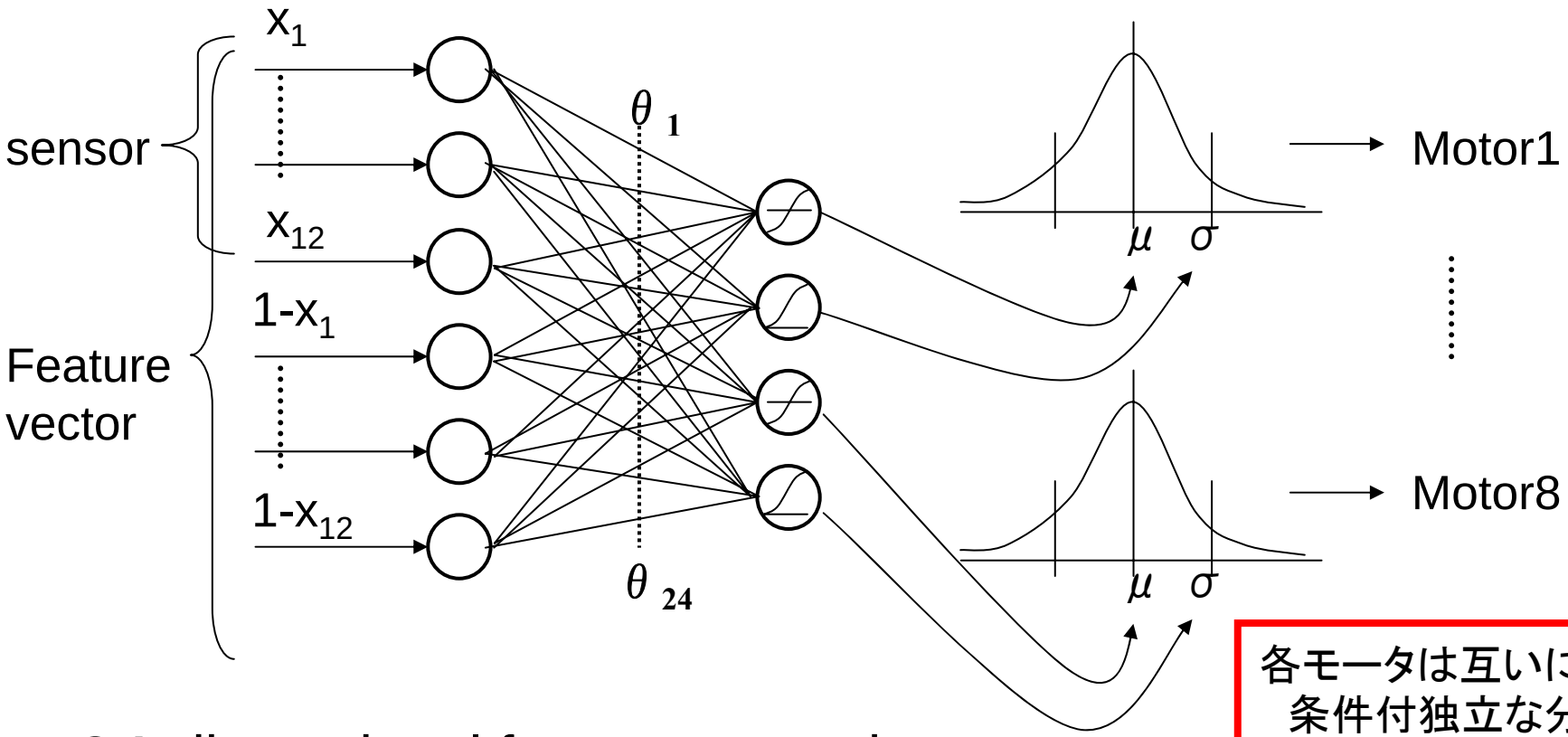
## 政策表現は予め設計者が与える



扱う問題についての事前知識が必要
制御規則の質は政策表現に大きく依存

# 【政策表現: Linear Coding】

状態価値関数の基底は
各軸3分割タイル分割




各モータは互いに
条件付独立な分布
→ これは妥当か？

- 24-dimensional feature vector is constructed from 12-dimensional sensor's information.
- Action is sampled from Cauchy distribution N($\mu$, $\sigma$).

$$\mu_i = 1 \Big/ \left(1 + \exp\left(-\sum_{k=1}^{6} \theta_{k,i} x_k\right)\right)$$

$$\sigma_i = 1 \Big/ \left(1 + \exp\left(-\theta_{7,i} x_7\right)\right) + 0.1$$

# 強化学習における多次元入出力の扱いとロボットへの適用

## 発表の流れ

（1） 強化学習とは？

（2） 確率的政策を**確率的勾配法**によって改善していく強化学習法
（自然勾配Actor-Critic法）

（3） **Q-learning アルゴリズム**を多次元状態-行動空間へ拡張
（ランダムタイリングによる多次元空間の関数近似と
Gibbsサンプリングによる行動選択を組合わせた強化学習）