# 方策勾配法とα β探索を組み合わせた強化学習アルゴリズムの提案

森岡 祐一, 五十嵐 治一

【従来手法】
コンピュータ将棋の評価関数パラメータの調整には教師有り学習の一種であるBonanza Methodが主に用いられてきた。
↓
一方、コンピュータチェスではTDLeaf(λ)で学習に成功したとの報告がある。しかし、コンピュータ将棋でのTDLeaf(λ)の成功例はまだ公表されていない。

**【提案手法】**
**TDLeaf(λ)とは異なる強化学習アルゴリズムである方策勾配法とα β探索を組み合わせる事により、報酬設定の自由度が高い学習アルゴリズムが作れるのではないか。**
**↓**
**方策勾配法の一アルゴリズムであるREINFORCEとα β探索を組み合わせるPGLeafを提案する。**

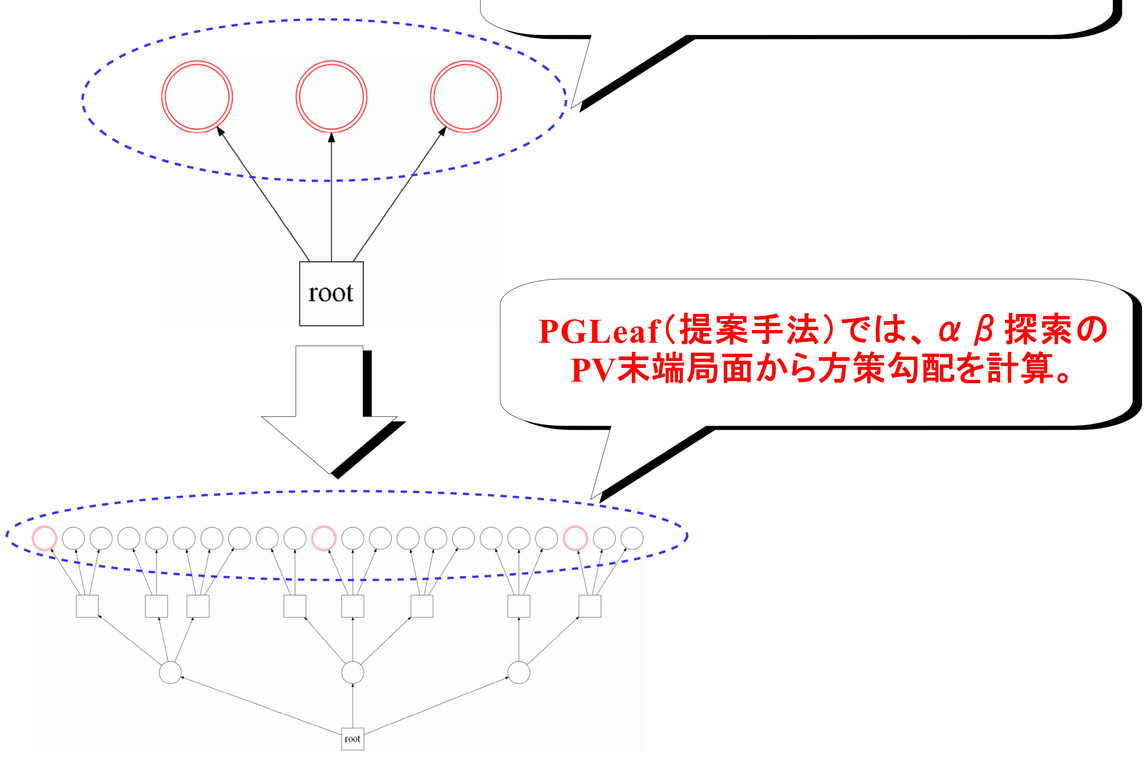
## PGLeafの学習の流れ

i) 評価関数のパラメータをごく小さな乱数で初期化。
ii) 以下(a)～(b)を繰り返す。
(a) 自己対戦で一定回数対局。
(b) 方策勾配を元にパラメータ修正。

## 方策勾配の計算方法

REINFORCEでは現局面で手を指した直後の局面から方策勾配を計算。

**PGLeaf(提案手法)では、α β探索のPV末端局面から方策勾配を計算。**




**α β探索と相性の良い評価関数パラメータの学習を目的とする。**

### パラメータθの更新則

$$\theta = \theta + \alpha \widehat{\nabla_\theta J(\theta)} \quad (1)$$

$$\widehat{\nabla_\theta J(\theta)} = \frac{1}{M}\sum_{m=1}^{M}\left(\gamma^{T_m-1} r_m - b*\right) g(m) \quad (2)$$

$$b* = \frac{\sum_{m=1}^{M}\gamma^{T_m-1} r_m g(m)^2}{\sum_{m=1}^{M} g(m)^2} \quad (3)$$

$$g(m) = \sum_{t=1}^{T_m} \nabla_\theta log\pi\left(a_{m,t}|s_{m,t};\theta\right) \quad (4)$$

$$\nabla_\theta log\left(\pi\left(a|s;\theta\right)\right) = \frac{1}{T}\left(\phi(s,a) - \sum_{a'\in A}\left(\pi\left(a'|s;\theta\right)\phi\left(s,a'\right)\right)\right) \quad (5)$$

【対局実験】

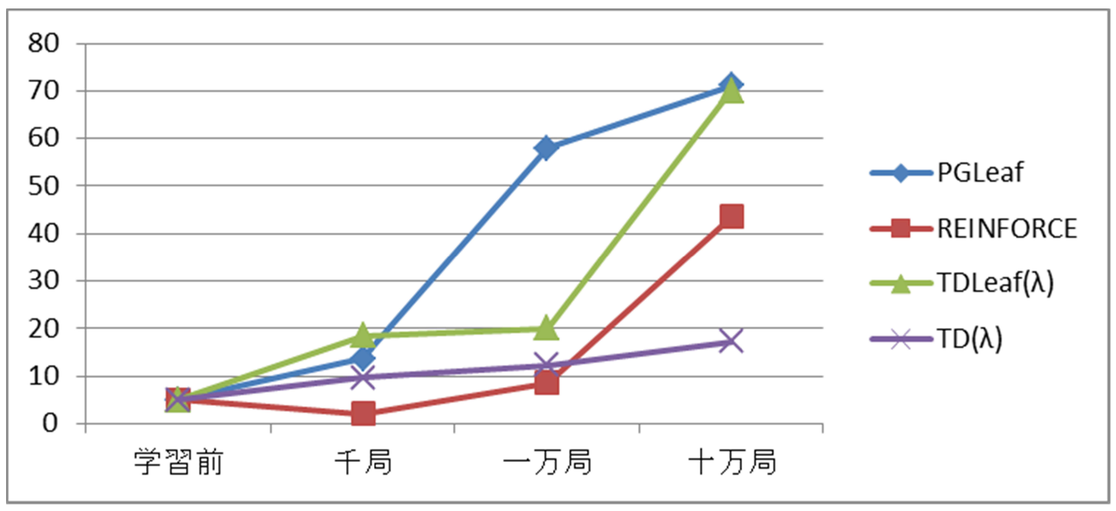
5五将棋を対象として以下の条件で対局実験を行い、提案手法の有効性を検証した。

1. PGLeaf、REINFORCE、TDLeaf(λ)、TD(λ)の4アルゴリズムでの比較を行った。
2. 学習は自己対戦を用いて行い、各アルゴリズムで10万局の対局・学習を行った。
3. 学習・対局時共にα β探索と線形の評価関数を組み合わせて先読みを行い、指し手を決定した。
4. 報酬設定はMDP的な環境(勝敗に応じて報酬を与える)と、非MDP的な環境(終局時の王将周辺の駒の有無に応じて報酬を増減する)の2通りの設定で比較を行う。
5. 対局実験時は思考時間を1手1秒とし、探索は反復深化を行った。
6. 対局実験は4アルゴリズムのリーグ戦と、ssp(フリーで公開されている5五将棋対応の思考ルーチン)との対局を行った。

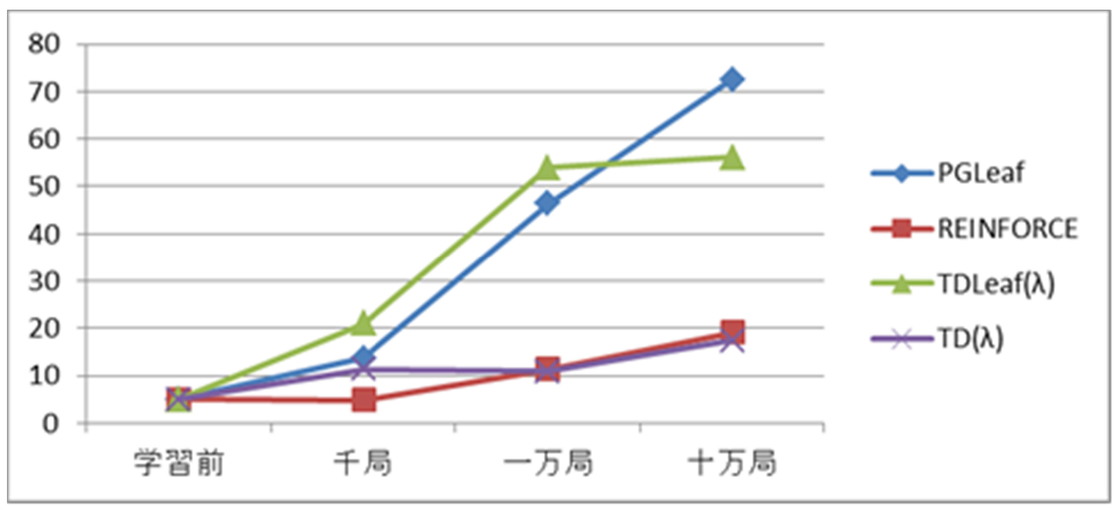
## 強化学習の4アルゴリズムでのリーグ戦(MDP環境、数値は勝率(%))

| | PGLeaf | REINFORCE | TDLeaf(λ) | TD(λ) |
|---|---|---|---|---|
| PGLeaf | - | 76.2 | 56.2 | 87.9 |
| REINFORCE | 23.8 | - | 16.0 | 31.3 |
| TDLeaf(λ) | 43.8 | 84.0 | - | 87.0 |
| TD(λ) | 12.1 | 68.7 | 13.0 | - |

## sspとの対局実験(MDP環境)




## sspとの対局実験(非MDP環境)




**【まとめ】**
**・提案手法は、α β探索を用いないREINFORCE・TD(λ)に対して、優位に勝ち越した。**
**・TDLeaf(λ)との比較では、僅かながらも高い勝率を示した。また、非MDP環境において性能劣化が無い事を確認した。**
**・今後は、本将棋に提案手法を適応していく。その際、報酬設定が重要となるので、有効な設定を検討していく。**