方策勾配法による静的局面評価関数の強化学習についての一考察

五十嵐治一，森岡祐一，山本一将

●提案手法の特徴：
①将棋の専門知識を必要しない
②強化学習として方策勾配法を採用
③「指し手評価の期待値」による行動決定
④ボルツマン分布による soft-max探索
⑤諸探索との関連が明確

将棋のルールを教え，対局させるだけでコンピュータはプロ棋士に勝てるようになるだろうか？

⇩ 強化学習

・バックギャモンでは Neural Net + TD(λ)法により実現 ['94 Tesauro]
・チェスではTDLeaf(λ)法の有効性が確認 ['98 Baxter et al.]

⇩ 方策ベースの強化学習
（マルコフ性のない報酬，方策でもOK）

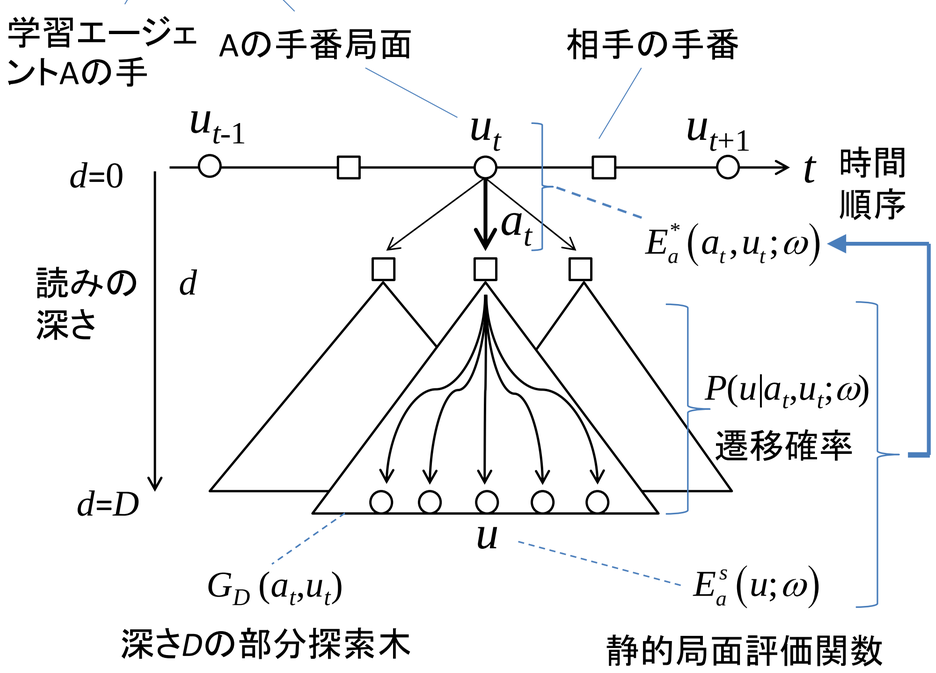
●本研究で用いた方策勾配法：[Williams92, 石原&五十嵐04]
目的関数$E(a(t);s(t),\omega)$ [離散時刻t，状態s，行動a]：方策の知識例，ルール重み　$E\left(a(t);s(t),\omega\right)=-\omega\left(s(t),a(t)\right)$ [$\omega\geqq 0$]
Boltzmann分布による確率的方策：
$$\pi\left(a(t);s(t),\omega\right)\equiv e^{+E(a(t);s(t),\omega)/T}\Big/\sum_{a\in A}e^{+E(a;s(t),\omega)/T}$$
エピソードごとの報酬期待値を極大化する学習則：
$$\Delta\omega=\varepsilon\cdot r\cdot\sum_{t=0}^{L-1}e_\omega(t)$$ [L：エピソード長，r：報酬，ε：学習係数]
特徴的適正度：
$$e_\omega(t)\equiv\partial\ln\pi/\partial\omega=+\frac{1}{T}\left[\frac{\partial E\left(a(t);s(t),\omega\right)}{\partial\omega}-\sum_{a\in A}\frac{\partial E\left(a;s(t),\omega\right)}{\partial\omega}\pi\left(a;s(t),\omega\right)\right]$$
（仮定）

⇩ 本提案

「指し手評価の期待値」を目的関数とする

$$E_a^*\left(a_t,u_t;\omega\right)\equiv\sum_{u\in U_D\left(a_t,u_t\right)}P\left(u\middle|a_t,u_t;\omega\right)E_a^s\left(u;\omega\right)$$

学習エージェントAの手　Aの手番局面　相手の手番

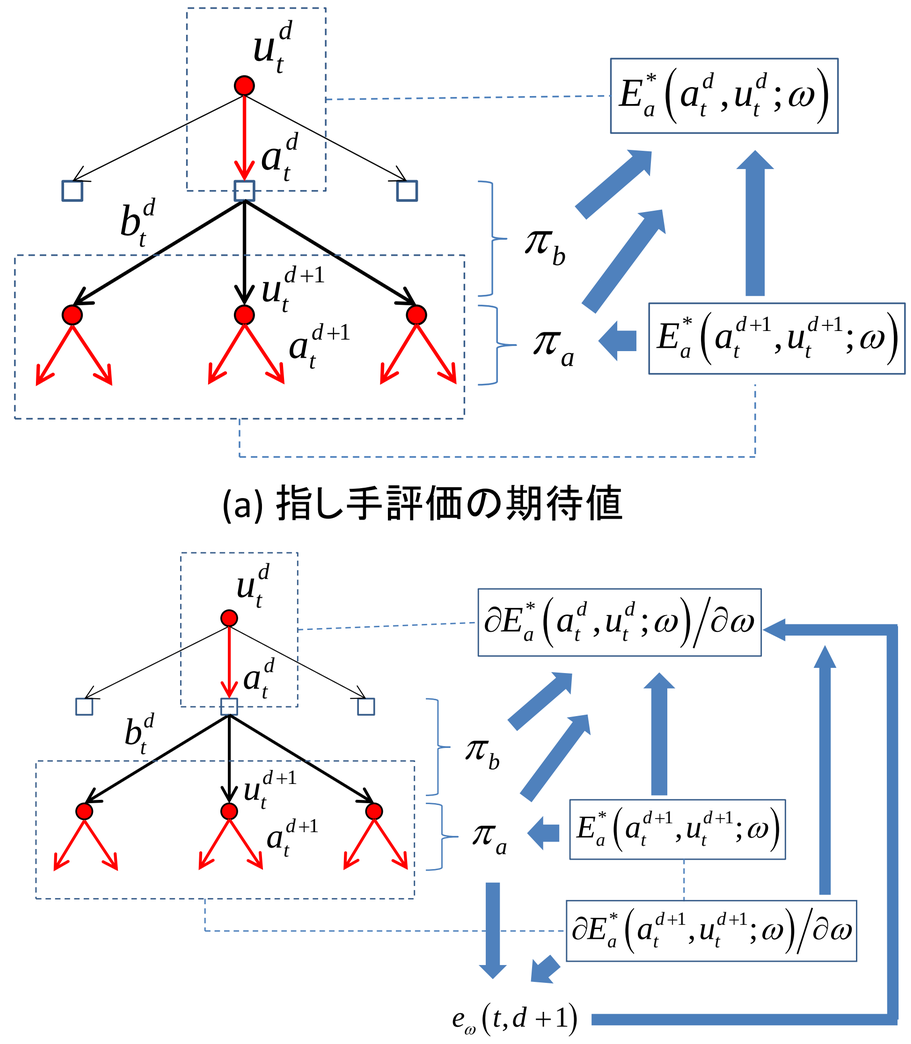


深さDの部分探索木　　静的局面評価関数

Boltzmann分布による確率的探索

従来
・min-max探索
・選択探索
・モンテカルロ探索

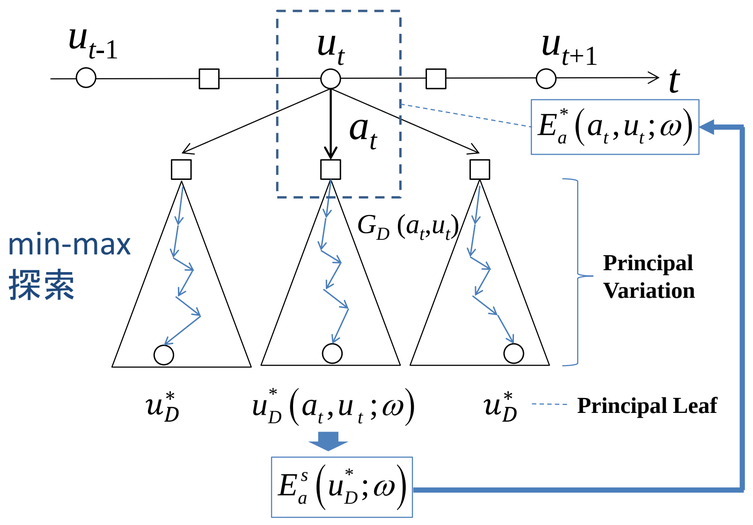
学習則に必要な再帰的計算



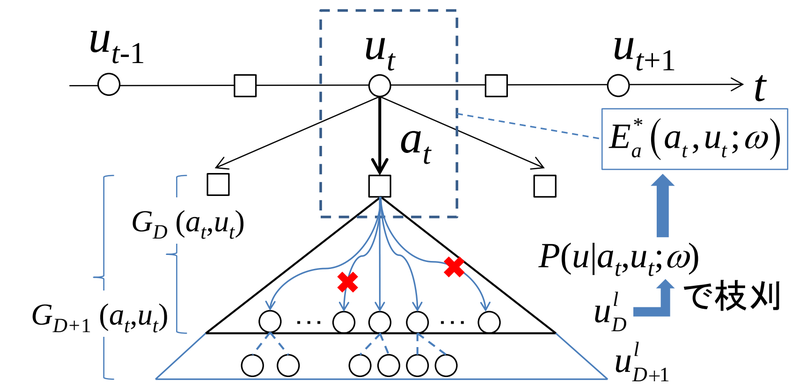
(a) 指し手評価の期待値

(b) 1階微係数

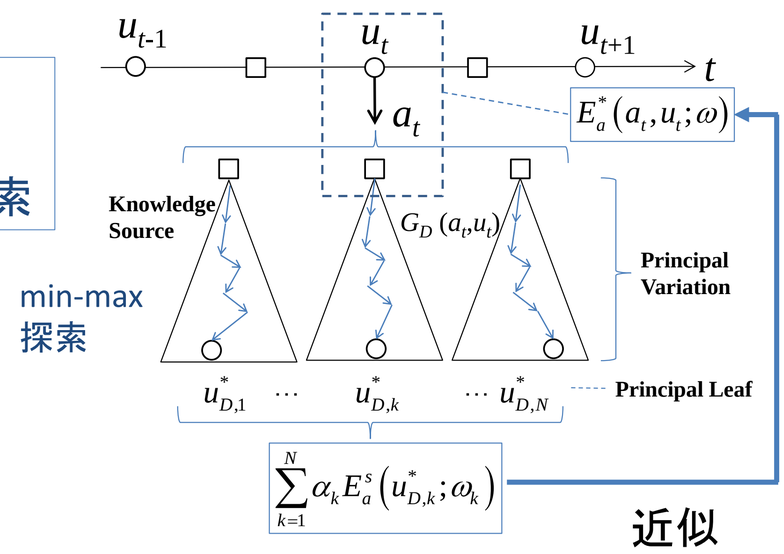
近似的計算法のアイデア



(a)PGLeaf法



(b)反復深化法の適用



(c)異なる評価関数による期待値計算の並列処理