IBM

潜在ダイナミクスにおける リスク考慮型意思決定

IBM東京基礎研究所
森村 哲郎

Joint work with
杉山 将 東京工業大学
鹿島 久嗣 東京大学
八谷 大岳 東京工業大学
田中 利幸 京都大学

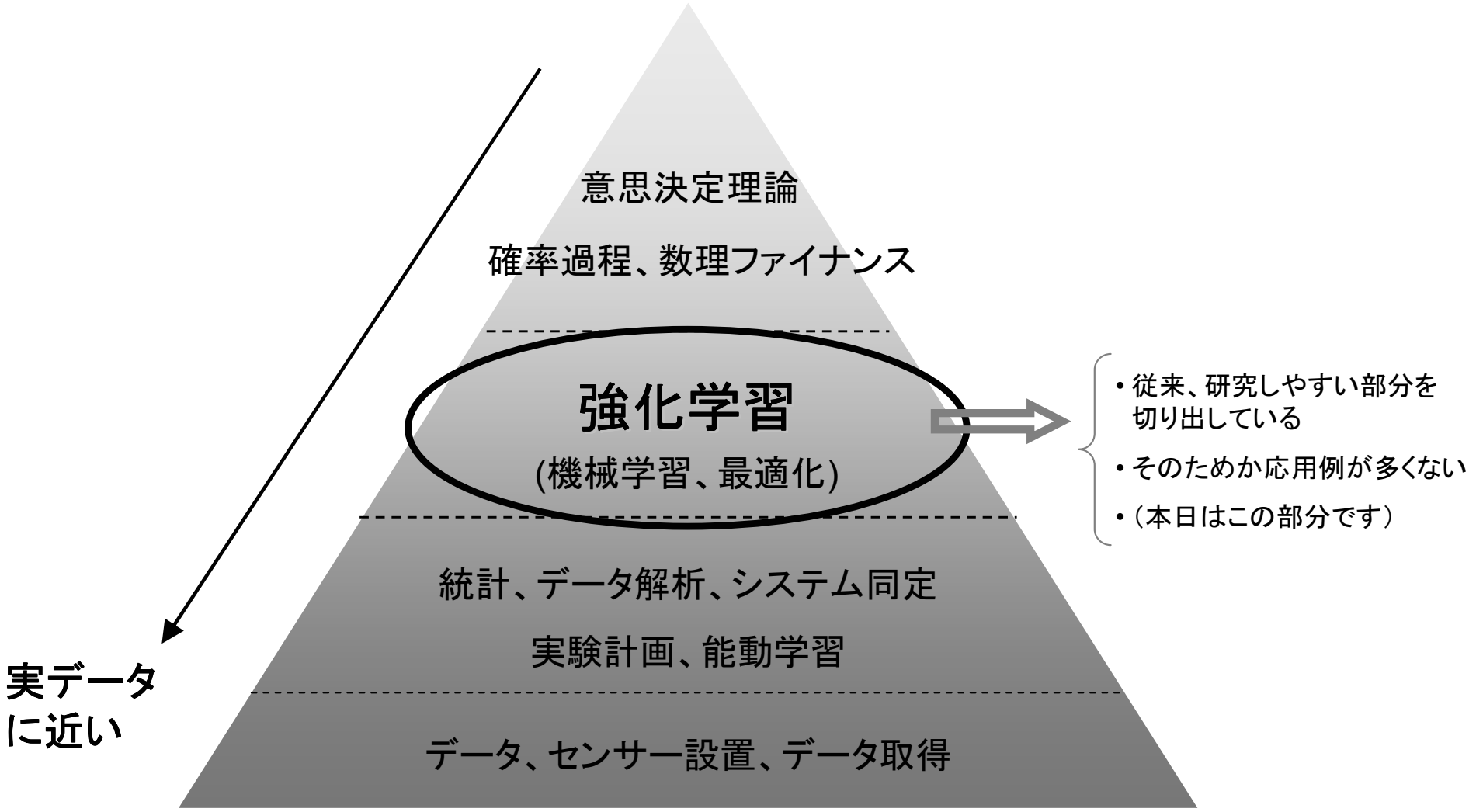
やりたいこと:データに基づく意思決定(支援)

- 未知の環境との相互作用のもたらすダイナミクスを解析し、
  意思決定を最適化する
  - 「何をすべきか(**what**)」を与えて、
    データから「どのように実現するか(**how to**)」を学習してほしい

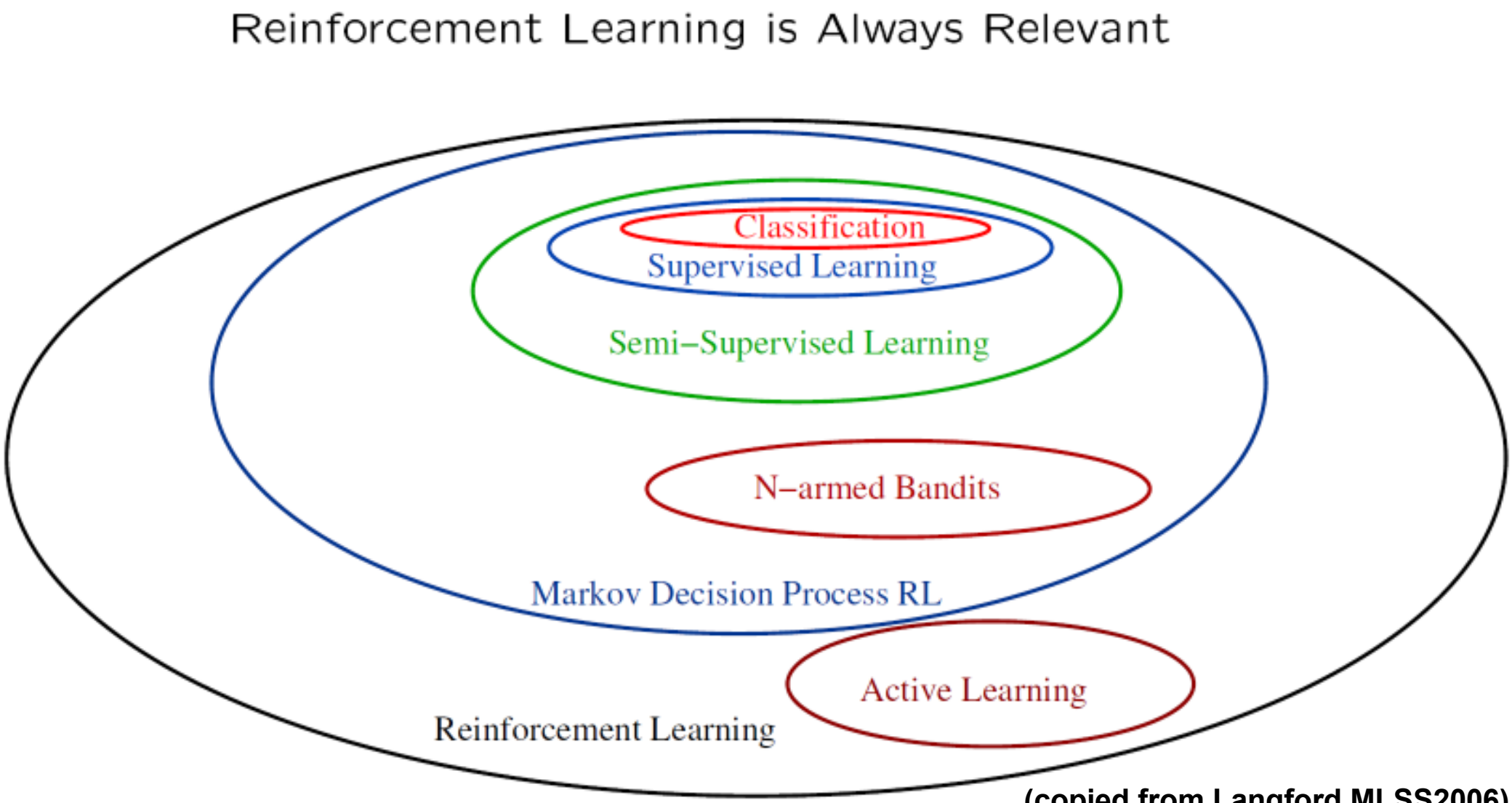
- その基盤となる理論的枠組に**強化学習**がある

といっても、「データに基づく意思決定」は多岐の研究領域にわたります

- 強化学習だけでは完結しない



[ご参考・少し異なる解釈] 強化学習は最も一般的な学習パラダイム



(copied from Langford MLSS2006)

アウトライン

- 強化学習の概要

- リスク考慮型強化学習

強化学習の位置づけ（機械学習の分類）

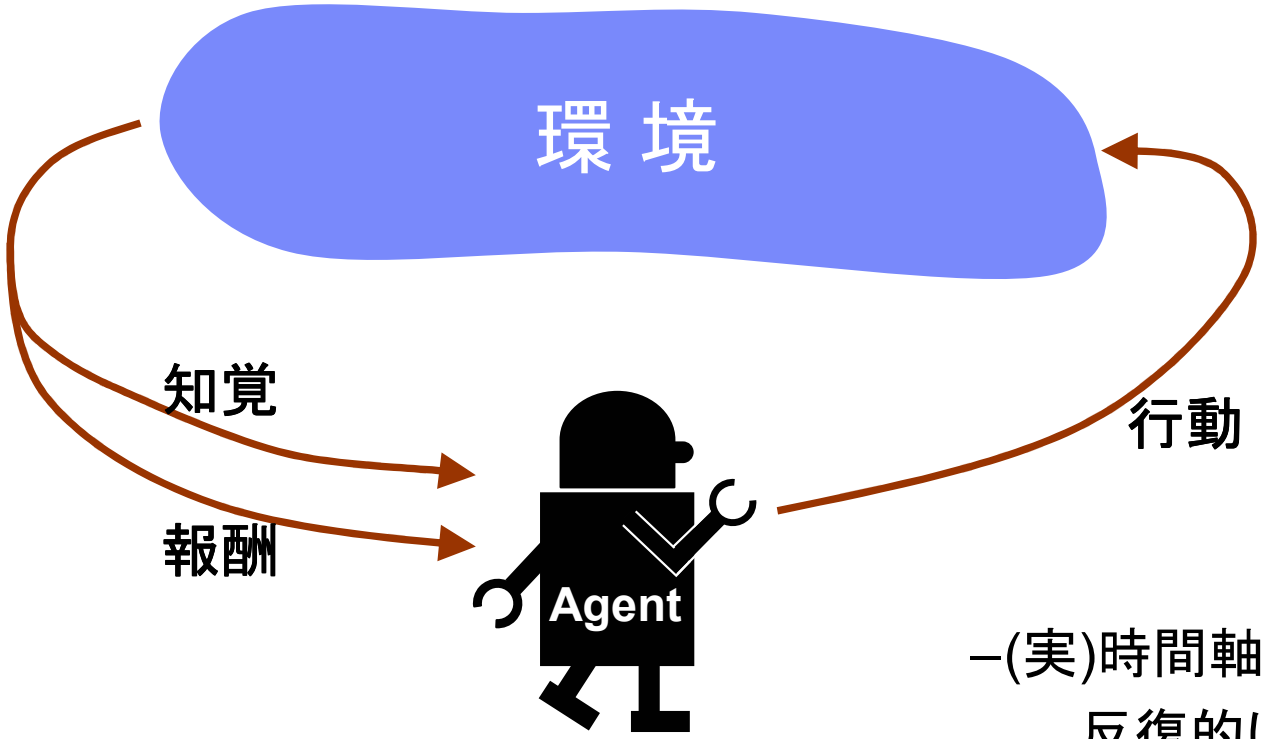
- 教師あり学習 [トレーニングデータ：特徴値、教師ラベル]
  - クラス分類
  - 回帰

- 強化学習 [トレーニングデータ：特徴値、報酬（評価値）]
  （明示的な教師信号の代わりに、報酬を利用して学習）
  - 強化学習問題: 状態遷移あり（MDPやPOMDP）
  - バンディット問題： 行動に依存した状態遷移なし

- 教師なし学習 [トレーニングデータ：特徴値のみ]
  - クラスタリング
  - 確率密度推定

強化学習は相互作用から学習する

- 強化学習 (RL) は(人工)生命のようなもの



–(実)時間軸に沿って,
　　反復的に学習や意思決定を行う
–エージェントは環境に影響を及ぼす
–環境は確率的で未知

将来の累積報酬(リターン)が最大になるように行動を選択

- エージェントの人生は経験の並び

$$o_1\ a_1\ r_1\ o_2\ a_2\ r_2\ \cdots\ o_t\ a_t\ r_t\ o_{t+1}\ \cdots$$

エージェントの一生

経験の単位

- 目的はリターン(≒累積報酬)を最大にする方策をみつけること
  - 即時報酬の最大化を目指しているわけではない

・リターンの定義:

✓非減衰の累積報酬 (もしくは平均報酬)

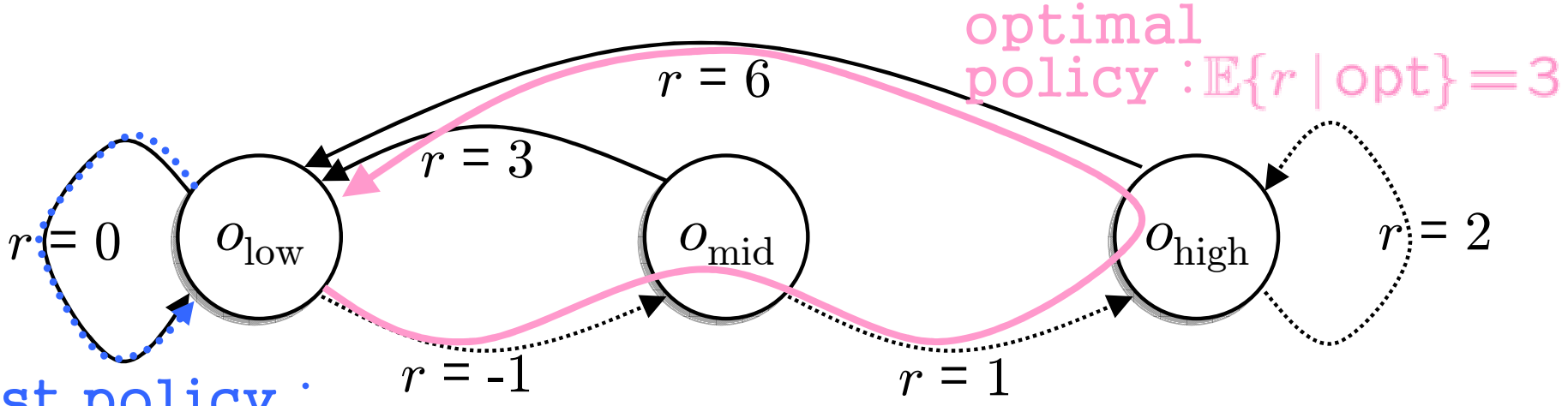
$$\mathrm{return}_t \triangleq r_t + r_{t+1} + r_{t+2} + \cdots$$

✓時間減衰率 γ の累積報酬

$$\mathrm{return}_t \triangleq r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \cdots$$

強化学習(RL)の簡単な例題:
近視眼的な方策が最悪な方策になる例

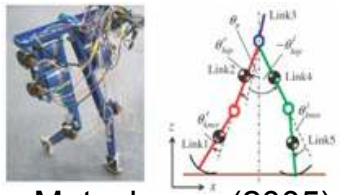
- キャンペーン・プランニング問題
  - キャンペーンを打つと、**短期売上げは上がるが**、 キャンペーン後の**カスタマーの購買意欲は下がる**
    - 観測 o ： カスタマー購買意欲 (low, mid, high)
    - 行動 ⟶ ： キャンペーンを実施
      ⋯⋯▶ ： 実施しない
    - 報酬 r : 単期の売上げ



RLは時間遅れのある大きな報酬を発見する

RLの実施例

- 従来、ロボット制御やゲーム等に使われてきた
  - –ロボティクス
    - ナビゲーション、二足歩行、ロボカップ・サッカー、ジャグリング、…
  - –制御
    - 工場プロセス制御、通信の流入制御、マルチメディアネットワークのリソース制御、ヘリコプター、エレベーター、…
  - –ゲーム
    - バックギャモン、チェス、オセロ、テトリス、…
  - –オペレーションズ・リサーチ
    - 倉庫管理、トランスポーテイション、スケジューリング、…
  - –その他
    - 対話システム、ヘルスケア、生物モデリング、…



Matsubara+ (2005)

Tesauro (1995)

Abe+ (2004)

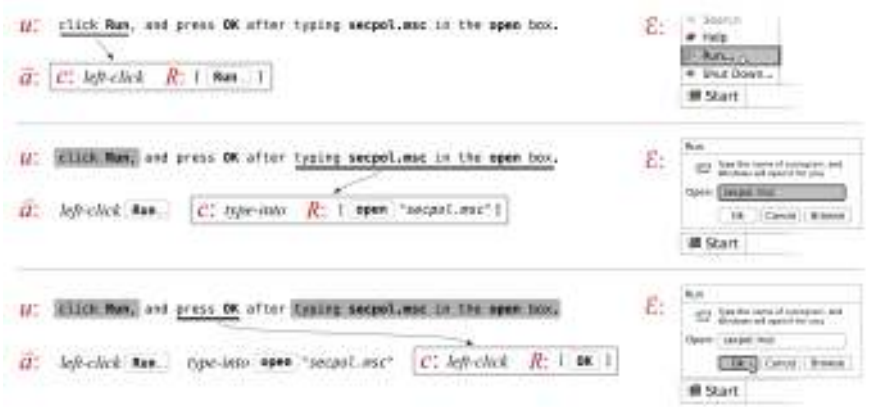
近年、現実の問題に適用され、新たな注目が集まっています

- ビジネスデータ解析や自然言語処理などの分野でRLが決定的役割を果たす実問題が次々に見出されている
  - –Abeら（KDD'10）は税金取立てに応用し、これによりNY州は3年間で100万ドルもの巨額の追加税収を得る見込み
    - 行動選択に制約のある制約付きRLを定式化
    - KDD'10 best industry/government paper

Tax Collections Optimizer

  - –Branavanら（ACL'09）はPCインストラクションの読解にRLを利用して、学習に必要な教師データ数の削減に成功
    - ACL'09 best paper

Mapping "natural language instructions"
↓
"sequences of executable actions"



強化学習法の分類

- 大きく2軸で分けられる

	価値/方策-反復型	(直接)方策-探索型
	- **価値関数**が方策を規定 - **価値関数**を学習することで、(暗に)方策が更新される	- 方策パラメータが方策を規定 - 目的関数の勾配等で、(明に)方策パラメータを更新
モデル・ベース型 環境モデルを同定し、その同定したモデルを利用して意思決定を行う	・動的計画 [Sutton & Barto '98] ・R-Max [Brafman & Tenneholz '03] ・LSTD/LSPI [Lagoudakis&Parr'03]	・線形計画 [Puterman'94, Ballo & Riano '06]
モデル・フリー型 環境の同定を経ずに、方策を学習する	・Q-learning [Watkins '89] ・Delayed Q-learning (with PAC Analysis) [Strehl '09]	・RAINFORCE [Williams '92] ・Actor-Critic [Sutton & Barto '98] - (自然)方策勾配法 [Peters+'03]

本日はここ

アウトライン

- 強化学習の背景
- リスク考慮型強化学習

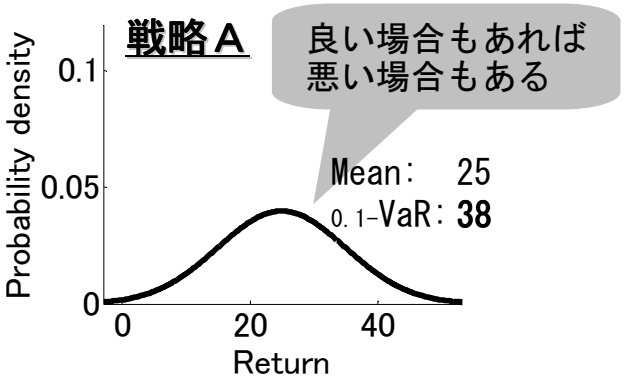
なぜリスクを考慮するのか？

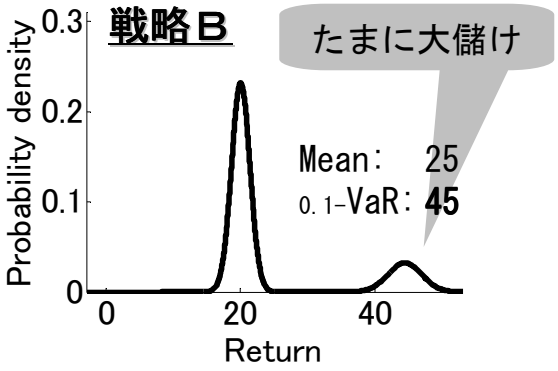
- **期待値だけでは見えない大切な情報がある**
  - 背後にあるリスクの見積もりが不可能
  - 従来の意思決定手法は、各選択肢のもたらす利得（損失）の**期待値**をもとに行われてる

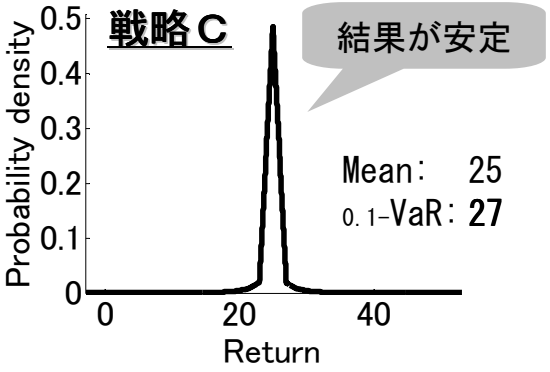
- **実問題や状況に応じて、リスク嗜好性は異なる**
  - とにかく期待リターンを最大にしたい　⇒ **risk-neutral**
  - 多少コストがかかっても、大損失することだけは避けたい　⇒ **risk-aversion**
  - 損するかもしれないが，大儲けの大チャンスに賭けたい　⇒ **risk-taking (chance-discovery)**

分布がわかれば、多種多様なリスク指標（情報）が手に入る

– **リターンの分布が求まれば**、金融工学等でよく用いられる Value-at-Risk（VaR）等、様々なリスク指標を算出でき、**リスク指標に基づいた意思決定が可能**







難点:リターンの観測まで時間遅れがあるため、その分布推定は難しい。

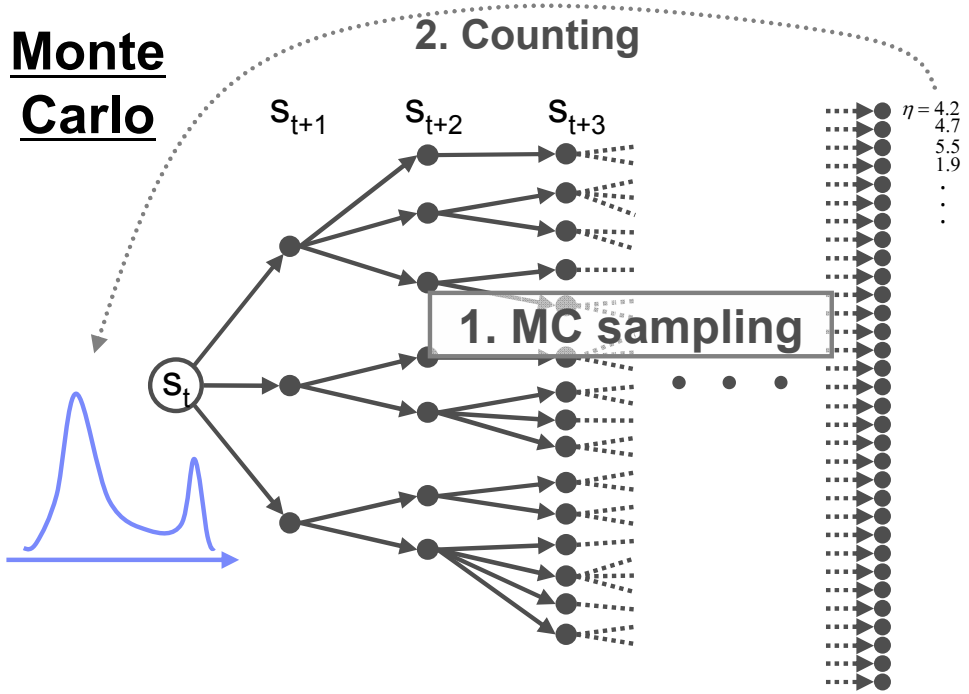
目的: 効率の良いリターン分布手法の確立

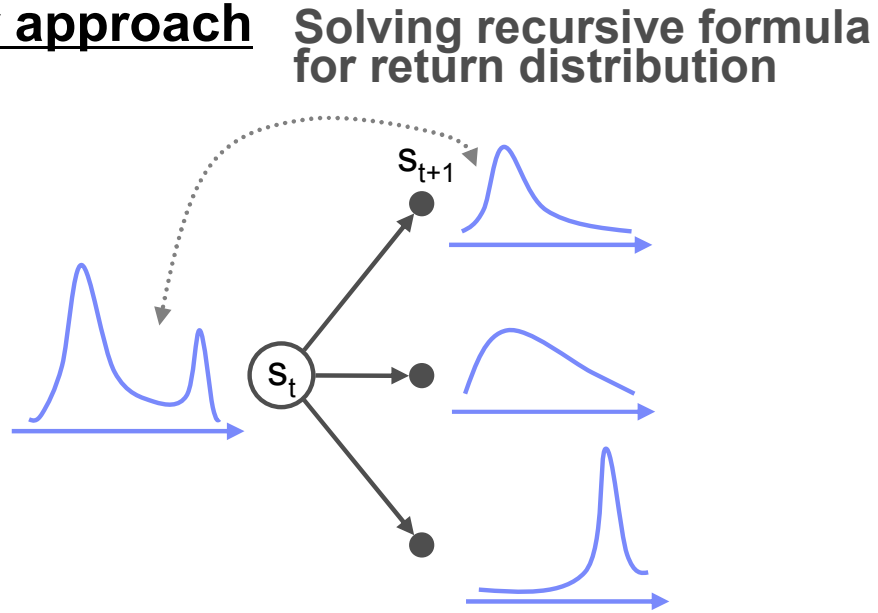
目次:

1. 二つのアプローチ
2. 分布Bellman方程式
3. 分布Bellman方程式を用いたリターン分布推定
   - パラメトリック法 [Morimura+ UAI2010]
   - ノンパラメトリック法 [Morimura+ ICML2010]
   - 実験
4. 推定リターン分布を用いたリスク考慮型意思決定
5. まとめ

リターン分布推定のためのアプローチ
価値関数(期待リターン)推定の場合同様、二通りのアプローチがあります

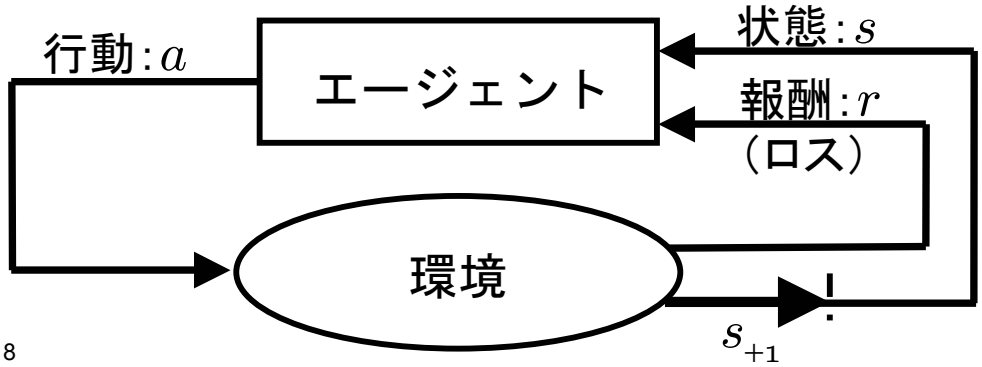
- シミュレーション・アプローチ（モンテカルロ法）
  - 直接的な方法だが、リターンまで観測に(無限の)時間遅れがあるため非効率
- 解析的アプローチ
  - リターン分布についての再帰式を導出して、その再帰式を解くことでリターン分布を推定





用語・関数の定義

- **マルコフ決定過程;** $\mathrm{MDP} \triangleq \{\mathcal{S}, \mathcal{A}, p_{\mathrm{T}}, P_{\mathrm{R}}\}$
  - 状態: $s \in \mathcal{S}$
  - 行動: $a \in \mathcal{A}$
  - 報酬: $r \in \mathbb{R}$
  - 状態遷移確率(未知):
    $p_{\mathrm{T}}(s_{+1}|s,a) \triangleq \Pr(s_{+1}|s,a)$
  - 報酬観測確率(未知):
    $P_r(r|s,a,s_{+1}) \triangleq \Pr(R \le r|s,a,s_{+1})$



- **マルコフ連鎖;** $\mathrm{M}(\pi) \triangleq \{\mathcal{S}, \mathcal{A}, p_{\mathrm{T}}, P_{\mathrm{R}}, \pi\}$
  - (確率的)方策:
    $\pi(a|s) \triangleq \Pr(a|s)$

- **リターンに関する統計量**
  - リターン (γ: 割引率)
    $\eta \triangleq \lim_{T\to\infty} \sum_{t=0}^{T} \gamma^t r_{+t}$
  - **(条件付) リターン分布関数**
    $P_\eta^\pi(\eta \,|\, s) \triangleq \Pr(E \le \eta \,|\, s, a, \mathrm{M}(\pi))$
    推定したい関数
  - 価値関数 (= 期待リターン)
    $V^\pi(s) \triangleq \mathbb{E}[\eta|s,\pi]$
    ($\mathbb{E}[x]$: x の期待値)

(リスク考慮型強化学習)

目次:

リターンに関する再帰式；ベルマン方程式

実は期待リターンに関してだけでなく、分布に関する再起式も簡単に導出できます

- **リターンの再帰式：**

$$\eta \triangleq \lim_{T\to\infty} \sum_{t=0}^{T} \gamma^t r_{+t}$$
$$= r + \gamma \eta_{+1}$$

- **期待リターンに関する再帰式（Bellman方程式）：**（$\because r \perp \eta_{+1} | s_{+1}$）

$$V(s) \triangleq \mathbb{E}[\eta | s, \pi] = \mathbb{E}[r + \eta_{+1} | s, \pi]$$
$$= \sum_{s_{+1}\in\mathcal{S}} \sum_{a\in\mathcal{A}} p_{\mathrm{T}}(s_{+1}|s,a)\pi(a|s) \left\{ \int_r r p_r(r|s,a,s_{+1}) dr + \gamma V(s_{+1}) \right\}$$

- **リターン分布に関する再帰式（分布Bellman方程式）：**（$\because r \perp \eta_{+1} | s_{+1}$）

$$P_\eta^\pi(\eta|s) = P_\eta^\pi(r + \eta_{+1}|s)$$
$$= \sum_{s_{+1}\in\mathcal{S}} \sum_{a\in\mathcal{A}} p_{\mathrm{T}}(s_{+1}|s,a)\pi(a|s) \int_{r\in\mathbb{R}} P_\eta^\pi\left(\frac{\eta - r}{\gamma} | s_{+1}\right) dP_r(r|s,a,s_{+1})$$

[中田&田中 2006]

分布Bellman方程式を用いたリターン分布推定

- **準備：分布Bellman作用素 $\mathcal{D}_\pi$ の定義**

$$\mathcal{D}_\pi[P_{\mathrm{E}}^\pi(\eta|s)] \triangleq \sum_{s_{+1}\in\mathcal{S}}\sum_{a\in\mathcal{A}} p_{\mathrm{T}}(s_{+1}|s,a)\pi(a|s)\int_r P_{\mathrm{E}}^\pi\left(\frac{\eta-r}{\gamma}|s_{+1}\right)dP_{\mathrm{R}}(r|s,a,s_{+1})$$

  – 分布Bellman方程式 ⇒ $P_{\mathrm{E}}^\pi(\eta|s) = \mathcal{D}_\pi[P_{\mathrm{E}}^\pi(\eta|s)]$

- **分布Bellman方程式を解くとは?**
  – ある累積分布関数 $F(\eta|s)$ が

$$F(\eta|s) = \mathcal{D}_\pi[F(\eta|s)],\ {}^\forall\eta\in\mathbb{R},{}^\forall s\in\mathcal{S}$$

  を満す時、$F(\eta|s)$ は分布Bellman方程式の解（=リターン分布関数 $P_{\mathrm{E}}^\pi(\eta\,|\,s)$ ）

- **リターン分布推定は、$F(\eta|s)$ と $\mathcal{D}_\pi[F(\eta|s)]$ が近くなるように $F(\eta|s)$ を学習すること**

分布Bellman方程式の解の一意性
動的計画法(DP)によるリターン分布推定は常に真の分布関数に収束する

- **DPにより、分布Bellman方程式を解く;** $dBellman\text{-}DP$
  - 各タイムステップ k で、推定リターン分布関数 $\hat{P}_k(\eta|s)$ を更新

$$\hat{P}_{k+1}(\eta|s) := \mathcal{D}_\pi[\hat{P}_k(\eta|s)],\ {}^\forall\eta \in \mathbb{R}, {}^\forall s \in \mathcal{S}$$

- **任意の初期分布から、常に真のリターン分布関数に収束**

Proposition 1 p_{T} や P_{R} は既知として，$dBllman\text{-}DP$ でリターン分布を推定した場合，解は初期近似分布 $\hat{P}_0$ に依存せず，常に真のリターン分布 P_E^π に収束する:

$$P_E^\pi(\eta|s) = \lim_{k\to\infty} \hat{P}_k(\eta|s),\ {}^\forall\eta \in \mathbb{R}, {}^\forall s \in \mathcal{S}.$$

証明: 分布Bellman方程式を特性関数化して証明される

モーメント推定量に関する収束率

準備

$$\left[\ \hat{m}_k(d\,;\,s) \triangleq \int_\eta \eta^d \hat{p}_k(\eta|s)d\eta, \quad \|m(y)\|_\infty \triangleq \max_{s\in\mathcal{S}} |m(y\,;\,s)|\ \right]$$

Proposition 2 真のリターン分布のモーメントが、全ての状態 $s \in \mathcal{S}$ で、D 次まで定義される場合、(リターン) モーメント推定誤差ベクトル

$$\boldsymbol{\Delta m}_{k,d} \triangleq \begin{bmatrix} \|\Delta m_k(1)\|_\infty \\ \vdots \\ \|\Delta m_k(d)\|_\infty \end{bmatrix} \triangleq \begin{bmatrix} \|m_{\mathrm{E}}^\pi(1) - \hat{m}_k(1)\|_\infty \\ \vdots \\ \|m_{\mathrm{E}}^\pi(d) - \hat{m}_k(d)\|_\infty \end{bmatrix}$$

に関して、以下の不等式が成立する：(ベクトルの不等号は要素ごとの不等号をを表す)

$$\boldsymbol{\Delta m}_{k+1,d} \leq \boldsymbol{A}_d \boldsymbol{\Delta m}_{k,d}, \quad 1 \leq d \leq D,\ k \geq 0.$$

ただし、

$$\boldsymbol{A}_d \triangleq \begin{bmatrix} \gamma & 0 & \cdots & 0 & 0 \\ {}_2C_1\|m_{\mathrm{R}}(1)\|_\infty\gamma & \gamma^2 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ {}_{d-2}C_1\|m_{\mathrm{R}}(d-2)\|_\infty\gamma & {}_{d-1}C_2\|m_{\mathrm{R}}(d-3)\|_\infty\gamma^2 & \cdots & \gamma^{d-1} & 0 \\ {}_dC_1\|m_{\mathrm{R}}(d-1)\|_\infty\gamma & {}_dC_2\|m_{\mathrm{R}}(d-2)\|_\infty\gamma^2 & \cdots & {}_dC_{d-1}\|m_{\mathrm{R}}(1)\|_\infty\gamma^{d-1} & \gamma^d \end{bmatrix}.$$

⇒ 低次のモーメント推定誤差が大きいほど、高次のモーメント推定は非効率に

モーメント推定量に関する収束率は$O(\gamma^k)$

- 線形変換されたモーメント推定誤差ベクトル $U_d^{-1}\Delta m_{k,d}$ の各要素は
  1DPステップで少なくても γ (<1) 減衰

Proposition 真のリターン分布のモーメントが、全ての状態 $s \in \mathcal{S}$ で、D 次まで定義される場合、リターン・モーメント推定誤差ベクトル $\Delta m_{k,d}$ に関して、以下の不等式が成立する：(ベクトルの不等号は要素ごとの不等号をを表す)

$$U_d^{-1}\Delta m_{k+1,d} \leq \gamma U_d^{-1}\Delta m_{k,d}, \quad 1 \leq d \leq D,\ k \geq 0.$$

減衰率

ここで、$U_d \in \mathbb{R}^{d\times d}$ は行列 A_d (式(16)) の固有ベクトル $v_{d,1}, v_{d,2}, \ldots, v_{d,d}$ からなる行列

$$U_d \triangleq [v_{d,1}, \ldots, v_{d,d}]$$

である.

モーメント推定量に関する収束率は$O(\gamma^k)$
Special case: 一次モーメント

- **分布Bellman方程式におけるDPでの1次のモーメントの収束率は、従来のBellman方程式におけるDPの収束率と同じ**

Corollary 全ての状態 $s \in \mathcal{S}$ で、リターン分布の1次のモーメントが定義できる場合、一次のモーメント推定誤差の減衰速度は

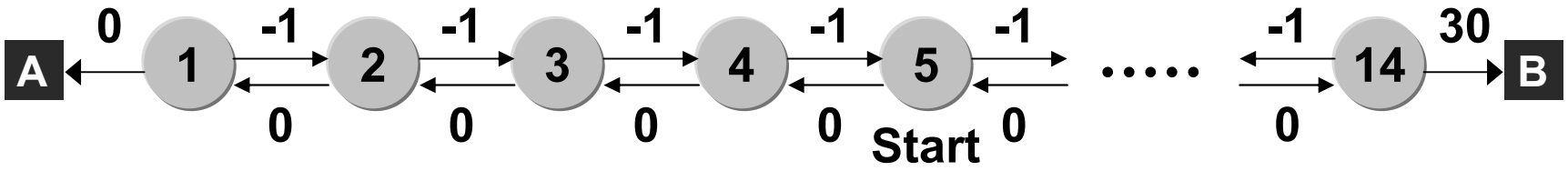
$$\|m_{\mathrm{E}}^{\pi}(1) - \hat{m}_{k+1}(1)\|_{\infty} \leq \gamma \|m_{\mathrm{E}}^{\pi}(1) - \hat{m}_{k}(1)\|_{\infty}$$

となり、リターンの期待値に関する動的計画法 *[1]* と同じである

⇒ *dBellman-DP*は、*Bellman-DP*の期待値推定から分布への自然な拡張

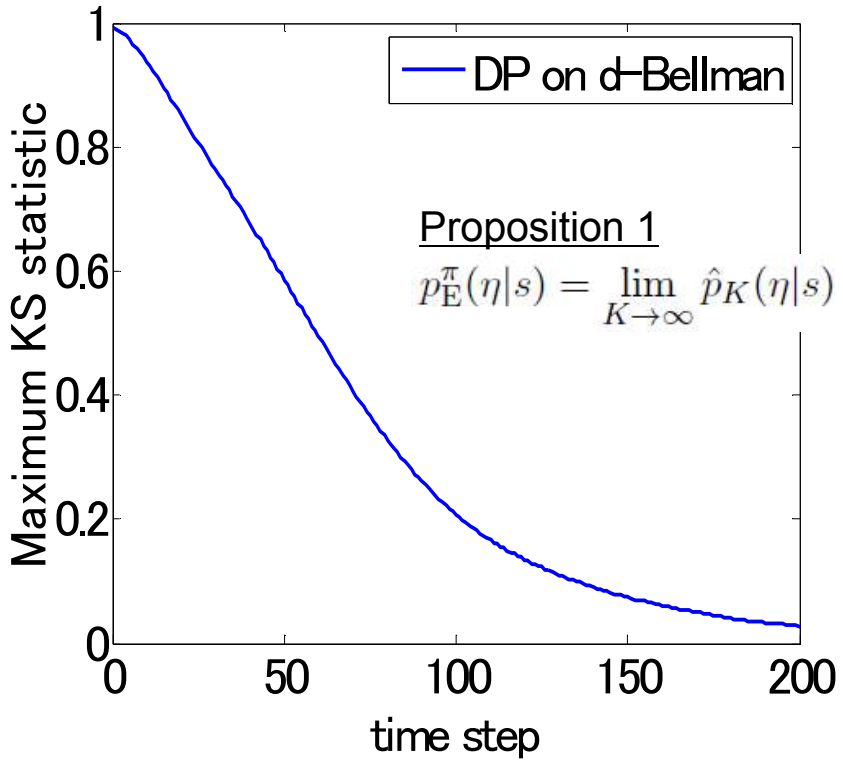
[ご参考] 数値実験でDPの収束性を検証

- **14状態のランダムウォーク**



- **KS統計量で分布Bellman方程式に基づくDPの収束性を評価**
  - –KS (Kolmogorov-Smirnov) 統計量：
    - 2つの分布 p と q を差異を計る最も有効かつ一般的な統計量の一つ

$$D_{\mathrm{KS}}(p(x), q(x)) \triangleq \sup_x \left| \int_{-\infty}^{x} p(x)dx - \int_{-\infty}^{x} q(x)dx \right|$$



(リスク考慮型強化学習)

目次:

分布モデルを仮定して分布Bellman方程式を解く

- **分布Bellman方程式は汎関数の自由度持つため、そのままでは解きにくい**
  - –リターン分布のモデルを仮定する

- **方針**
  - –以下の繰り返して、少しずつ分布Bellman方程式の再帰（右辺と左辺の）関係を満たすようにする

$$\hat{P}_\eta(\eta|s) \overset{\text{近似}}{\Longrightarrow} \mathcal{D}_\pi[\hat{P}_\eta(\eta|s)]$$

推定リターン分布
（分布Bellman方程式の**左辺**に対応）

ターゲット分布
（分布Bellman方程式の**右辺**に対応）

(リスク考慮型強化学習)

目次:

パラメトリック・リターン分布推定:
KLダイバージェンスを(確率的)自然勾配法により最小化

- リターン分布をパラメータθをもつパラメトリック分布 $\hat{p}_\eta(\eta|s,\boldsymbol{\theta})$ で表現
- ターゲット分布 $\mathcal{D}^\pi[\hat{p}_\eta(\eta|s,\boldsymbol{\theta})]$ から $\hat{p}_\eta(\eta|s,\boldsymbol{\theta})$ の擬距離にKLダイバージェンスを使用

$$\mathrm{D}_{\mathrm{KL}}\left\{\mathcal{D}^\pi[\hat{p}_\eta(\eta|s,\boldsymbol{\theta})],\hat{p}_\eta(\eta|s,\boldsymbol{\theta})\right\} \triangleq \int_\eta \mathcal{D}^\pi[\hat{p}_\eta(\eta|s,\boldsymbol{\theta})]\log\frac{\mathcal{D}^\pi[\hat{p}_\eta(\eta|s,\boldsymbol{\theta})]}{\hat{p}_\eta(\eta|s,\boldsymbol{\theta})}d\eta$$

- θを調整して D_{KL}を(局所)最小化することで、リターン分布を推定
  - D_{KL}の勾配: $\nabla_{\theta_i}\mathrm{D}_{\mathrm{KL}}[\mathcal{D}\hat{p}_\eta,\hat{p}_\eta] \triangleq \lim_{\varepsilon\to 0}\frac{\mathrm{D}_{\mathrm{KL}}[\mathcal{D}\hat{p}_\eta(\eta|s,\boldsymbol{\theta}),\hat{p}_\eta(\eta|s,\boldsymbol{\theta}+\varepsilon\boldsymbol{e}_i)]-\mathrm{D}_{\mathrm{KL}}[\mathcal{D}\hat{p}_\eta(\eta|s,\boldsymbol{\theta}),\hat{p}_\eta(\eta|s,\boldsymbol{\theta})]}{\varepsilon}$

$$= -\int_\eta \mathcal{D}\hat{p}_\eta(\eta|s,\boldsymbol{\theta})\frac{\partial}{\partial\theta_i}\log\hat{p}_\eta(\eta|s,\boldsymbol{\theta})d\eta$$

$$= \frac{1}{\gamma}\mathbb{E}\left[-\int_\eta \hat{p}_\eta\left(\frac{\eta-r}{\gamma}|s,\boldsymbol{\theta}\right)\frac{\partial}{\partial\theta_i}\log\hat{p}_\eta(\eta|s,\boldsymbol{\theta})d\eta \mid s,\pi\right]$$

  - (確率的)自然勾配法により最小化: ←指数分布族を使えばモーメントが一致

$$\boldsymbol{\theta} := \boldsymbol{\theta} + \alpha\boldsymbol{F}_{\hat{p}_\eta}(s,\boldsymbol{\theta})^{-1}\int_\eta \hat{p}_\eta\left(\frac{\eta-r}{\gamma}|s,\boldsymbol{\theta}\right)\frac{\partial}{\partial\theta_i}\log\hat{p}_\eta(\eta|s,\boldsymbol{\theta})d\eta$$

学習率

$\hat{p}_\eta(\eta|s,\boldsymbol{\theta})$ のフィッシャー情報行列

使用するパラメトリック分布

- 解析的に自然勾配とVaRを計算できる分布を利用
  - ガウス分布： $\boldsymbol{\theta}^{\mathrm{g}} \triangleq [\mu, \sigma]^\top$ ← 指数分布族のため、モーメント一致性が保障される

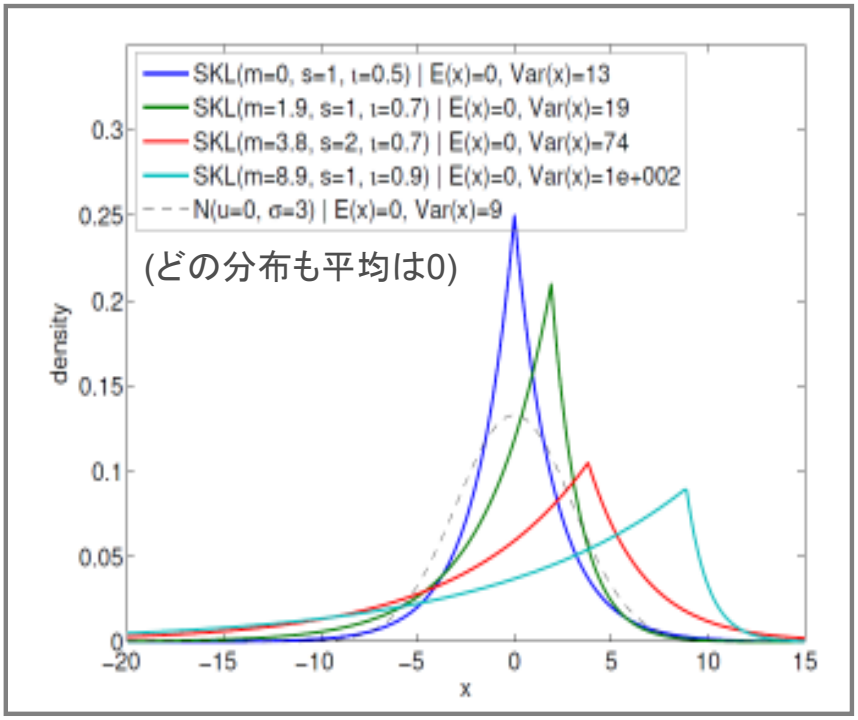
$$p^{\mathrm{g}}(x|\mu,\sigma) \triangleq \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

  - ラプラス分布： $\boldsymbol{\theta}^{\mathrm{l}} \triangleq [m, b]^\top$ ← 裾野の重たい対称分布

$$p^{\mathrm{l}}(x|m,b) \triangleq \frac{1}{2b}\exp\left(-\frac{1}{b}|x-m|\right)$$

  - 歪ラプラス分布：$\boldsymbol{\theta}^{\mathrm{skl}} \triangleq [m, b, c]^\top$ ← 裾野の重たい非対称分布

$$p^{\mathrm{skl}}(x|m,b,c) \triangleq \frac{c(1-c)}{b}\begin{cases}\exp\left(\frac{1-c}{b}(x-m)\right) & \text{if } x < m\\ \exp\left(-\frac{c}{b}(x-m)\right) & \text{otherwise}\end{cases}$$



各パラメトリック分布の更新式

- **ガウスモデル**: $[\ \mu \triangleq \mu(s;\boldsymbol{\theta}),\ \sigma \triangleq \sigma(s;\boldsymbol{\theta}),\ \mu' \triangleq \mu(s_{+1};\boldsymbol{\theta}),\ \sigma' \triangleq \sigma(s_{+1};\boldsymbol{\theta}),\ \delta \triangleq r + \gamma\mu' - \mu\ ]$

TD誤差

$$\mu := \mu + \alpha\delta \quad \longrightarrow \text{従来のTD学習と同じ更新式}$$

$$\sigma := \sigma + \alpha\{\delta^2 + \gamma^2\sigma'^2 - \sigma^2\}/2\sigma \quad \longrightarrow \text{[Dearden 1998, Sato \& Kobayashi 2001] の分散の更新式と同様}$$

- **ラプラスモデル**: $[\ m \triangleq m(s;\boldsymbol{\theta}),\ b \triangleq b(s;\boldsymbol{\theta}),\ m' \triangleq m(s_{+1};\boldsymbol{\theta}),\ b' \triangleq b(s_{+1};\boldsymbol{\theta}),\ \delta \triangleq r + \gamma m' - m:\ ]$

$$m := \begin{cases} m + \alpha\{-1 + \exp(\frac{1}{\gamma b'}\delta)\}b & \text{for } \delta \le 0, \\ m + \alpha\{1 - \exp(-\frac{1}{\gamma b'}\delta)\}b & \text{for } \delta > 0, \end{cases}$$

$$b := b + \alpha\left\{-b + |\delta| + \gamma b' \exp(-\frac{1}{\gamma b'}|\delta|)\right\} b/2.$$

ガウスモデルと異なり、更新量が bound される

↓

ロバストRLの更新式と類似

[Mihatsch & Neuneier 2002, Sugiyama+ 2010]

- **歪ラプラスモデル**: $[\ m \triangleq m(s;\boldsymbol{\theta}),\ b \triangleq b(s;\boldsymbol{\theta}),\ c \triangleq c(s;\boldsymbol{\theta}),\ m' \triangleq m(s_{+1};\boldsymbol{\theta}),\ b' \triangleq b(s_{+1};\boldsymbol{\theta}),$
$c' \triangleq c(s_{+1};\boldsymbol{\theta}),\ \delta \triangleq r + \gamma m' - m\ ]$

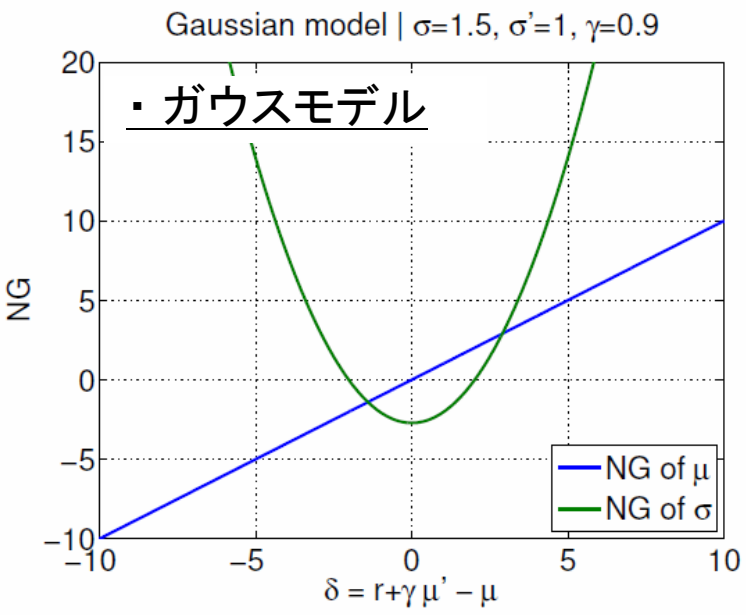
・$\delta \leqq 0$

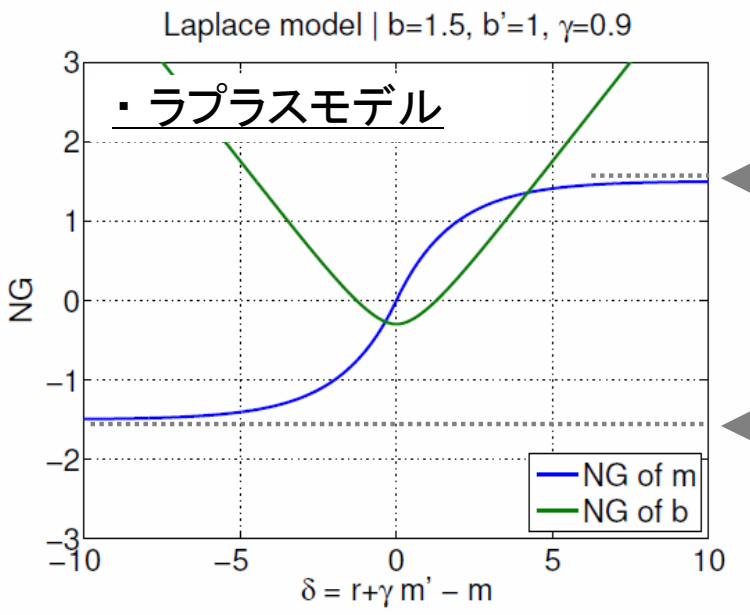
$$\begin{cases} m := m + \frac{\alpha}{2c}\left[-2b - \frac{\gamma b'(1-c)(1-2c')}{c'(1-c')} - (1-c)\delta + \frac{1-c'}{1-c}\left\{2b + \frac{\gamma b'(1-2c)}{c'}\right\}\exp(\frac{c'}{\gamma b'}\delta)\right] \\ b := b + \frac{\alpha}{2c}\left[-b(1-c) - \frac{\gamma b'(1-c)^2(1-2c')}{c'(1-c')} - (1-c)^2\delta + \frac{1-c'}{1-c}\left\{b(1-2c) + \frac{\gamma b'(1-3c+3c^2)}{c'}\right\}\exp(\frac{c'}{\gamma b'}\delta)\right] \\ c := c + \frac{\alpha}{2b}\left[-b(1-c) - \frac{\gamma b'(1-c)^2(1-2c')}{c'(1-c')} - (1-c)^2\delta + (1-c')\left\{b + \frac{\gamma b'(1-2c)}{c'}\right\}\exp(\frac{c'}{\gamma b'}\delta)\right] \end{cases}$$

・$\delta > 0$

$$\begin{cases} m := m + \frac{\alpha}{2(1-c)}\left[2b - \frac{\gamma b'c(1-2c')}{c'(1-c')} - c\delta - \frac{c'}{c}\left\{2b - \frac{\gamma b'(1-2c)}{1-c'}\right\}\exp(-\frac{1-c'}{\gamma b'}\delta)\right] \\ b := b + \frac{\alpha}{2(1-c)}\left[-bc + \frac{\gamma b'c^2(1-2c')}{c'(1-c')} + c^2\delta + \frac{c'}{c}\left\{-b(1-2c) + \frac{\gamma b'(1-3c+3c^2)}{1-c'}\right\}\exp(-\frac{1-c'}{\gamma b'}\delta)\right] \\ c := c + \frac{\alpha}{2b}\left[-bc - \frac{\gamma b'c^2(1-2c')}{c'(1-c')} - c^2\delta + c'\left\{b + \frac{\gamma b'(1-2c)}{1-c'}\right\}\exp(-\frac{1-c'}{\gamma b'}\delta)\right] \end{cases}$$

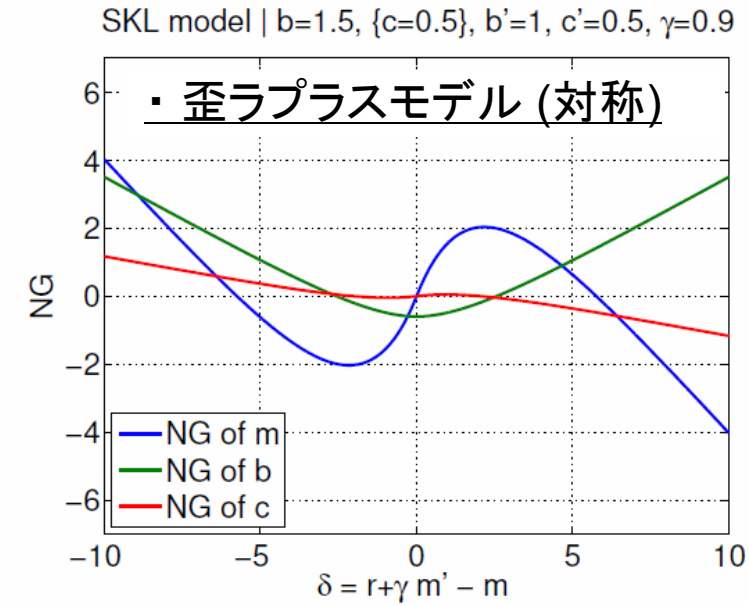
TD誤差（δ）対する、更新値

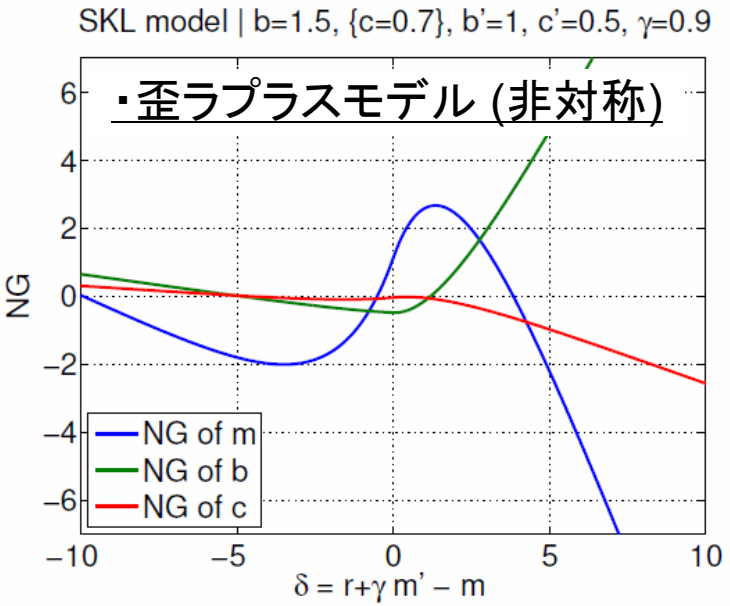




bound されてる

bound されてる





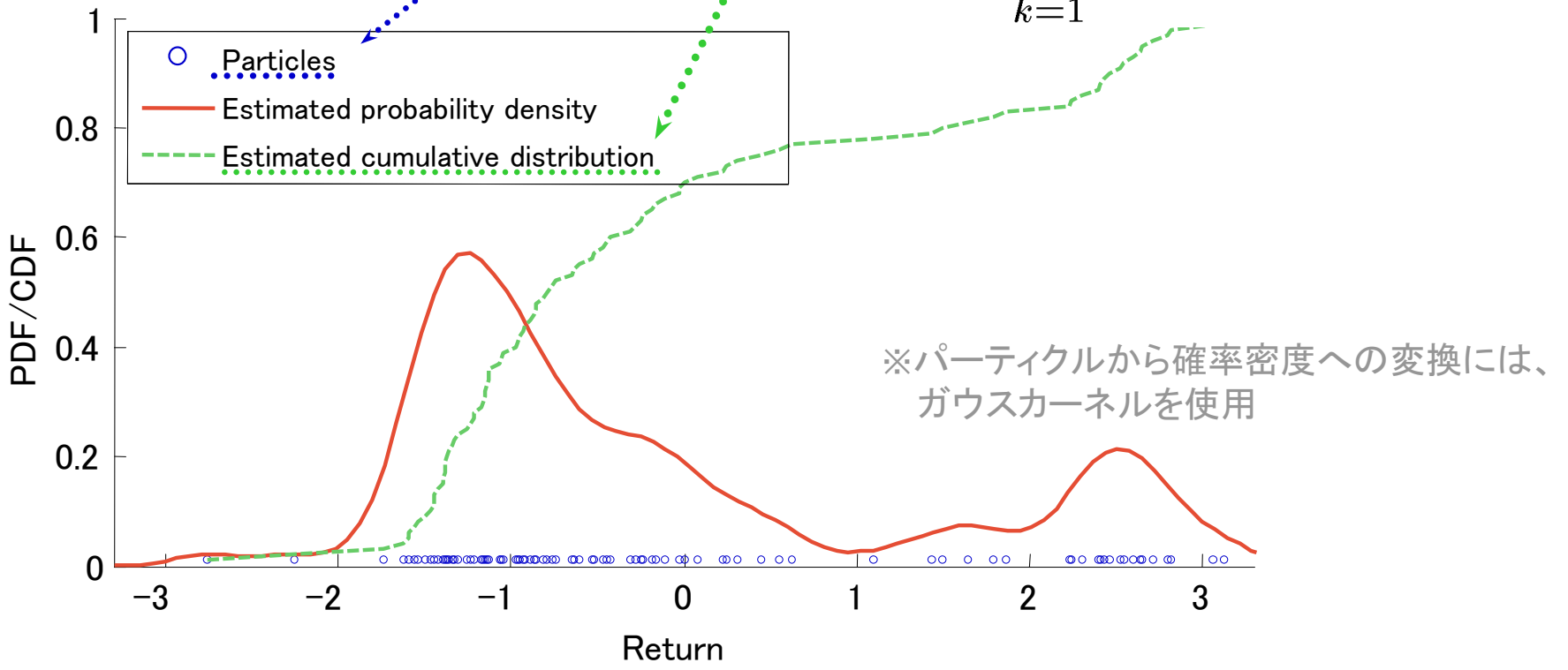
(リスク考慮型強化学習)

目次:

ノンパラメトリック・リターン分布推定：
パーティクルでリターン分布を近似

- N個のパーティクル $v_s = \{v_s^{(1)}, \ldots, v_s^{(N)}\}$ でリターン分布 $P_\eta^\pi(\eta|s)$ を近似
  - ランダムにパーティクル η_i を選ぶことは、近似分布 $\hat{P}_\eta(\eta|s, \boldsymbol{V})$ から標本を1つ生成することと同義：

$$\eta_s^{(i)} \sim \hat{P}_\eta(\eta|s;\boldsymbol{V}) \triangleq \frac{1}{K}\sum_{k=1}^{K} I(v_{s,k} \le \eta)$$



※パーティクルから確率密度への変換には、ガウスカーネルを使用

パーティクル・スムージングによる分布推定:
1時刻先の状態のパーティクルを利用して、現状態のパーティクルを更新

- **リターン分布の再帰式**

$$p_\eta^\pi(\eta|s) \propto \sum_{s_{+1}\in\mathcal{S}} \sum_{a\in\mathcal{A}} p_{\mathrm{T}}(s_{+1}|s,a)\pi(a|s) \int_r p_r(r|s,a,s_{+1}) p_\eta^\pi\left(\frac{\eta - r}{\gamma}|s_{+1}\right) dr$$

より、互いに独立な標本 $(r^{(1)}, v_{+1}^{(1)}), \ldots, (r^{(N)}, v_{+1}^{(N)})$ **を用いて:**

(I は指示関数)

$$P_\eta^\pi(\eta|s) = \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} I(r^{(n)} - \gamma v_{+1}^{(n)} \le \eta)$$

$$\Leftrightarrow \quad (r + \gamma\eta_{+1}) \sim P_\eta^\pi(\eta|s) \quad \text{<分布の平衡式>}$$

- **パーティクル・スムージング (Particle Smoothing; PS)**
  - $\hat{P}_\eta(\eta|s)$**を分布の平衡式**に従わせるには、乱択の**状態**s **のパーティクル** $v_s^{(i)}$ **を、一時刻先の状態**s_{+1} **の乱択パーティクル** $v_{s_{+1}}^{(j)}$ **を用いて更新すればいい:**

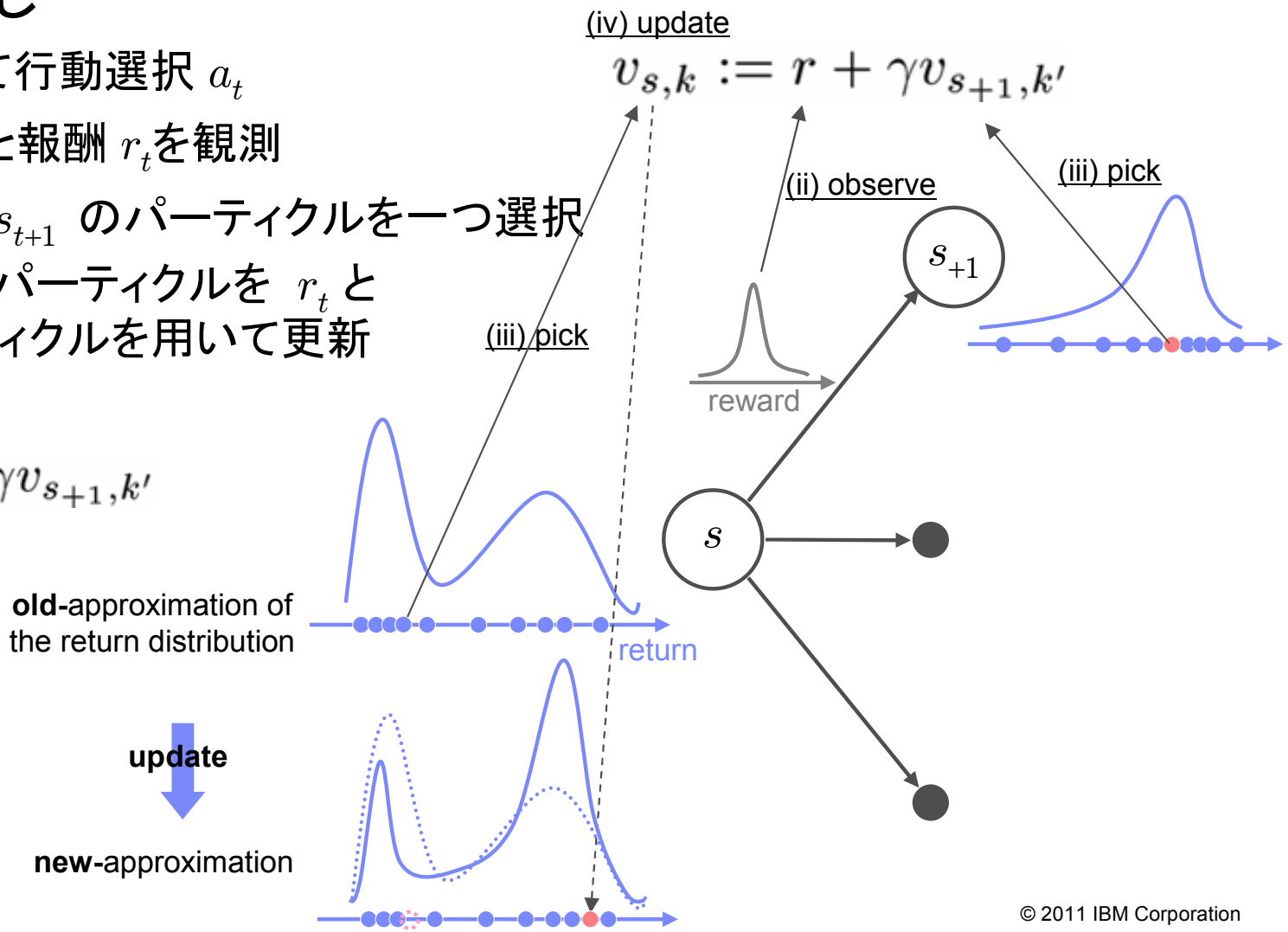
$$v_s^{(i)} := r + \gamma v_{s_{+1}}^{(j)}$$

Particle Smoothing Return Distribution approximation (RDPS) アルゴリズム

- ダイナミクス(環境)に関する知識は必要としない
  - 以下の繰り返し
    - (i) 方策に従って行動選択 a_t
    - (ii) 次状態 s_{t+1} と報酬 r_tを観測
    - (iii) 乱択で s_t と s_{t+1} のパーティクルを一つ選択
    - (iv) 選んだ s_t のパーティクルを r_t と s_{t+1}のパーティクルを用いて更新

$$v_{s,k} := r + \gamma v_{s_{+1},k'}$$

Only using observations



RDPSアルゴリズムの精度保証

- Kolmogorov-Smirnov (KS) 統計量で $\hat{P}_\eta(\eta|s)$ と $\mathcal{D}^\pi[\hat{P}_\eta(\eta|s)]$ の相違度を測る

$$\mathrm{D_{KS}}\{\hat{\mathrm{P}}_\eta(\eta|\mathrm{s}), \mathcal{D}^\pi[\hat{\mathrm{P}}_\eta(\eta|\mathrm{s})]\} \triangleq \max_\eta \left|\hat{\mathrm{P}}_\eta(\eta|\mathrm{s}) - \mathcal{D}^\pi[\hat{\mathrm{P}}_\eta(\eta|\mathrm{s})]\right|$$

- **Proposition** (概要): ***extending Kolmogoroff (1941) result***

  PDPSのパーティクル更新を十分に繰り返せば、以下が成り立つ

$$\mathbb{E}\left[\lim_{K\to\infty} \sqrt{K} D^*_{\mathrm{KS}}\{\hat{P}_{\mathrm{E}}(\eta|s;\boldsymbol{V}_K), \Pi\hat{P}_{\mathrm{E}}(\eta|s;\boldsymbol{V}_K)\}\right] \leq \sqrt{\frac{\pi}{2}}\ln(2)$$

$$\mathbb{V}\left[\lim_{K\to\infty} \sqrt{K} D^*_{\mathrm{KS}}\{\hat{P}_{\mathrm{E}}(\eta|s;\boldsymbol{V}_K), \Pi\hat{P}_{\mathrm{E}}(\eta|s;\boldsymbol{V}_K)\}\right] \leq \frac{1}{12}\pi(\pi - 6\ln^2(2))$$

分散

⇒ パーティクル数を増やすほど、分布Bellman方程式の残差を減らせる

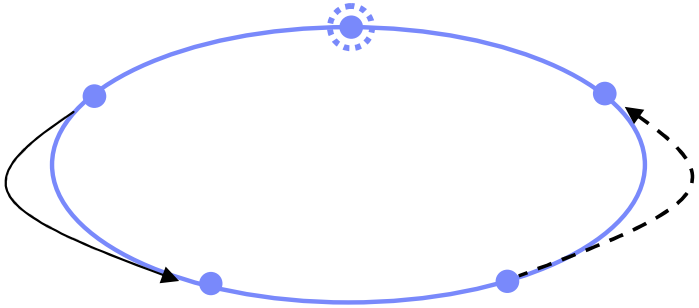
(リスク考慮型強化学習)

目次:

数値実験:
リターン分布推定能を評価 ※エージェントはランダムウォーク

- **無限期間、5状態2行動MDP**

● :状態
— :リンク



報酬の設定

——▶ : $r \sim \mathrm{N}(\mu{=}20,\ \sigma^2{=}2)$
（N:ガウス分布）

- - - -▶ : $r \sim \mathrm{G}(k{=}2,\ \theta{=}5){+}30$
（G:ガンマ分布）

その他の状態遷移 : $r = 0$

- **有限期間、5状態2行動MDP**

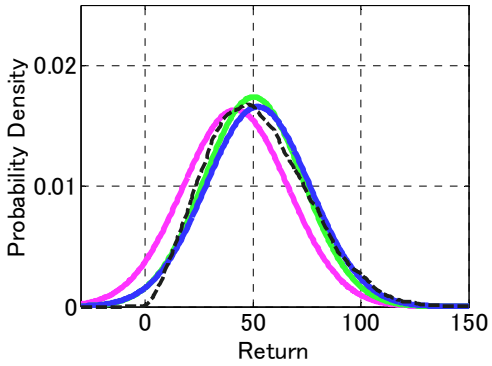
○ :終端状態

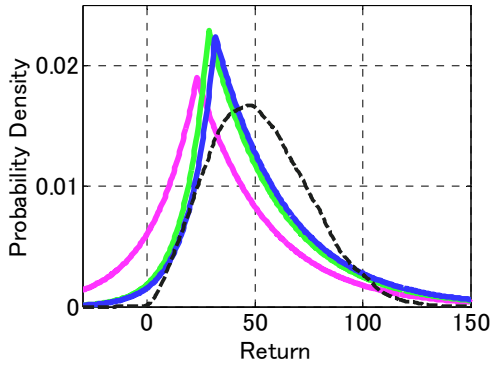
- **無限期間、30状態2行動MDP**
  - –状態遷移確率と報酬は乱数で決定 [Morimura+ '09]
    - 状態遷移はDirichlet分布で初期化
    - 報酬はガウス分布で初期化

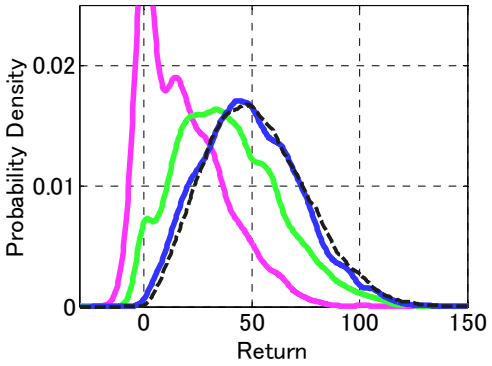
■ 状態 ◉ のリターン分布推定結果　[- · - · - ：真のリターン分布(モンテカルロにより推定)]

□ 無限期間、5状態MDP：——: $6*10^3$ステップ時，——: $15*10^3$ステップ時，——: $30*10^3$ステップ時

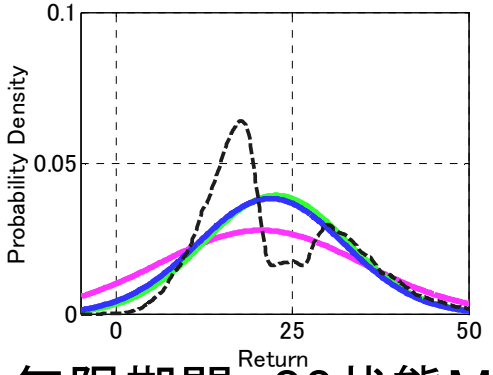
- **ガウスモデル**(パラメトリックモデル)
- **歪ラプラスモデル**(パラメトリックモデル)
- **パーティクルモデル**(ノンパラメトリックモデル)

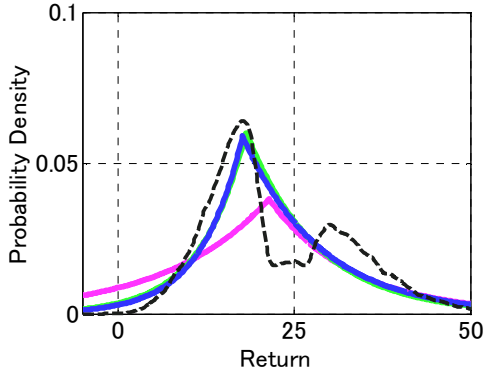


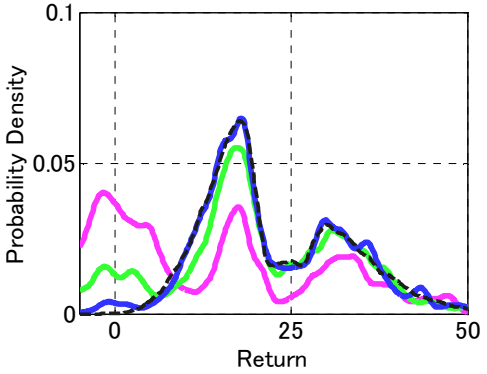




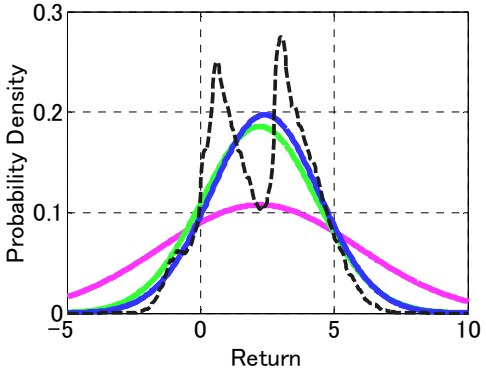
□ 有限期間、5状態MDP:——: $2*10^3$ステップ時，——: $5*10^3$ステップ時，——: $10*10^3$ステップ時

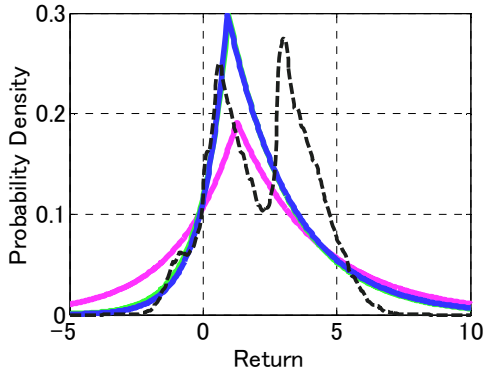


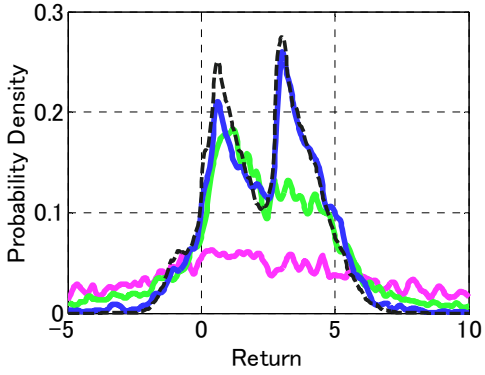




□ 無限期間、30状態MDP——: $6*10^4$ステップ時，——: $15*10^4$ステップ時，——: $30*10^4$ステップ時







(リスク考慮型強化学習)

目次:

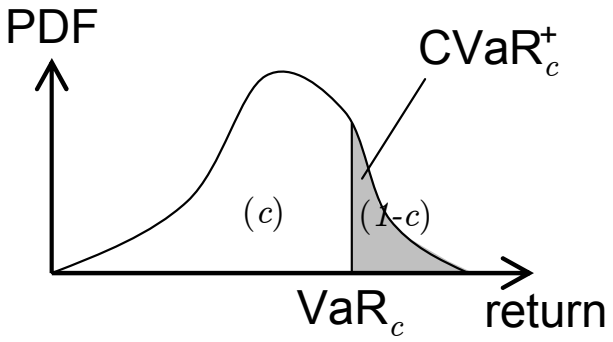
1. 二つのアプローチ
2. 分布Bellman方程式
3. 分布Bellman方程式を用いたリターン分布推定
   - パラメトリック法 [Morimura+ UAI2010]
   - ノンパラメトリック法 [Morimura+ ICML2010]
   - 実験
4. **推定リターン分布を用いたリスク考慮型意思決定**
5. まとめ

推定リターン分布を用いたリスク考慮型意思決定の例

- **RDPSによりリターン分布を求められるので、分布から規定される任意のリスク指標 $\mathcal{F}_0[\eta|\pi], \mathcal{F}_1[\eta|\pi], \ldots, \mathcal{F}_k[\eta|\pi]$ を用いた最適化問題を扱える:**

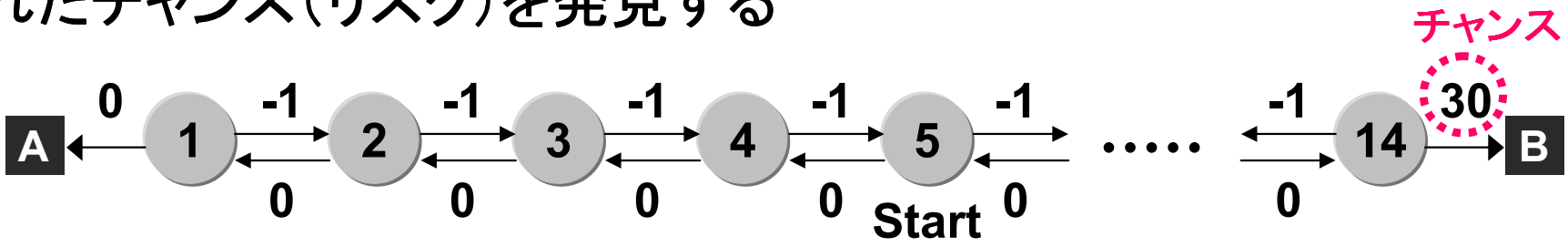
$$\begin{aligned} \max_{\pi \in \Pi} \quad & \mathcal{F}_0[\eta|\pi], \\ \text{s.t.} \quad & \mathcal{F}_1[\eta|\pi] \geq \varepsilon_1, \ldots, \mathcal{F}_k[\eta|\pi] \geq \varepsilon_k, \end{aligned}$$

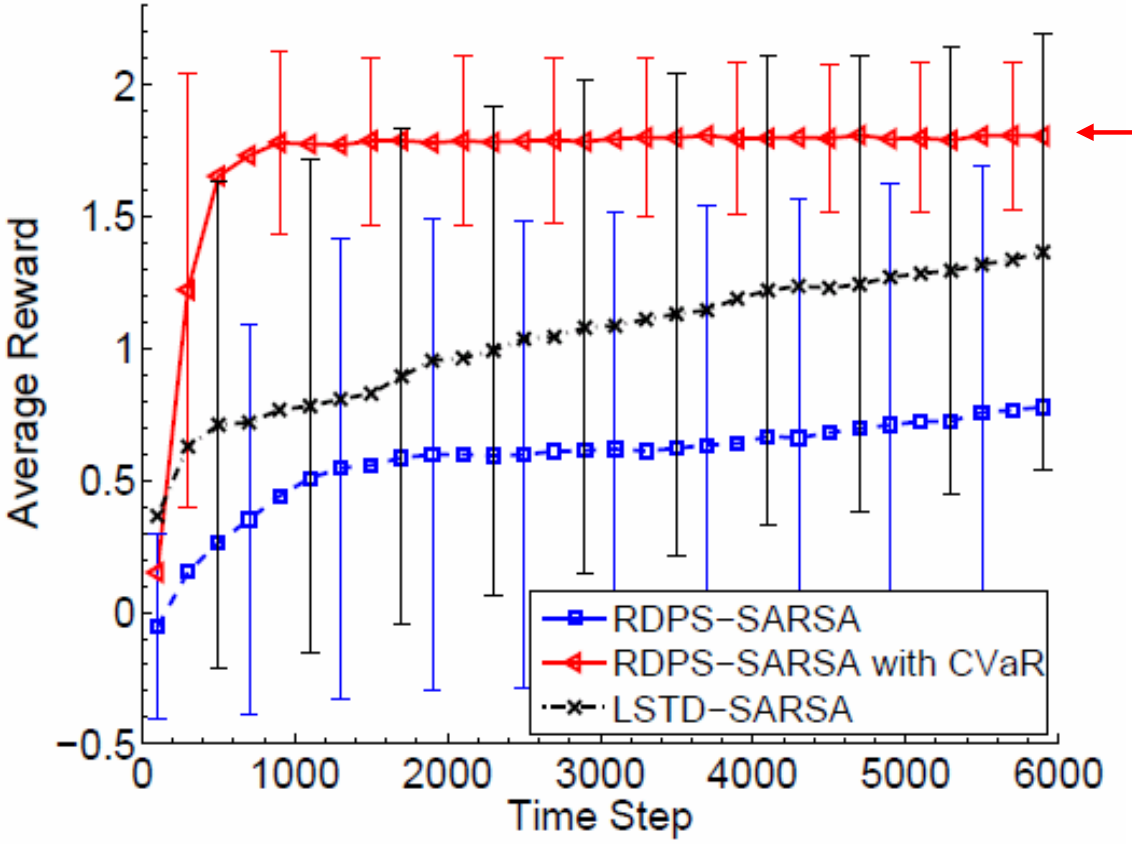
- リスク嗜好性は探索・搾取のトレードオフをバランスする **[Bagnell 2004]**
  - Risk-aversion → 搾取 (robust RL)
  - Risk-taking → 探索
- **今回は、CVaRを用いて探索の効率化を目指す**
  - 実方策(実際の行動選択に用いる探索用方策)と目的方策の二種類の方策を用いる
    - 実方策の目的関数: $\mathrm{CVaR}^{+}\mathbb{E}^{\pi}\left[\eta \mid P_{\mathrm{E}}^{\pi}(\eta|s,a) \geq 1-c\right]$
    - 目的方策の目的関数: 期待リターン
    - (両方策のバランスに重点サンプリングを利用)



リスクを活用すれば，学習の効率化が実現できます

- 隠れたチャンス(リスク)を発見する





Proposed algorithm (chance-discovery RL)

(リスク考慮型強化学習)

目次:

まとめ

- リターン分布の再帰式である分布Bellman方程式をみた
  - –その性質、解の一意性などを明らかにした

- 分布Bellman方程式を用いた二通りのリターン分布法を提案した
  - –パラメトリック法：自然勾配によりKLダイバージェンスを(局所)最小化
  - –ノンパラメトリック法：particle smoothingによりKS統計量を小さくする

	収束までに要する試行数 （学習の効率）	モデルの自由度 （≒VaR等の推定精度）
パラメトリック・アプローチ	小ない	低い
ノンパラ・アプローチ	多い	高い

- リスク考慮によって、効率の良い探索を達成できることを示した

参考文献

- N. Abe, N. K. Verma, C. Apte, and R. Schroko. Cross channel optimized marketing by reinforcement learning. In *International Conference on Knowledge Discovery and Data Mining*, pages 767-772, 2004.
- D Bello and G Riano. Linear programming solvers for markov decision processes. In *IEEE Systems and Information Engineering Design Symposium*, pages 90-95, 2006.
- R. I. Brafman and M. Tennenholtz. R-max { a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3:213-231, 2003.
- A. Kolmogoroff. Condence limits for an unknown distribution function. *The Annals of Mathematical Statistics*, 12(4):461-463, 1941.
- J. Langford. Reinforcement Learning Theory. Machine Learning Summer School, 2006.
- M. G. Lagoudakis and R. Parr. Least-squares policy iteration. *Journal of Machine Learning Research*, 4:1107-1149, 2003.
- T. Morimura, M. Sugiyama, H. Kashima, H. Hachiya, and T. Tanaka. Nonparametric return distribution approximation for reinforcement learning. In *International Conference on Machine Learning*,2010.
- T. Morimura, M. Sugiyama, H. Kashima, H. Hachiya, and T. Tanaka. Parametric return density estimation for reinforcement learning. In *Conference on Uncertainty in Articial Intelligence*, 2010.
- J. Peters, S. Vijayakumar, and S. Schaal. Reinforcement learning for humanoid robotics. In *IEEE-RAS International Conference on Humanoid Robots*, 2003.
- M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley and Sons, 1994.
- A. L. Strehl, L. Li, and M. L. Littman. Reinforcement learning in nite mdps: Pac analysis. *Journal of Machine Learning Research*, 10:2413-24443, 2009.
- R. S. Sutton and A. G. Barto. *Reinforcement Learning*. MIT Press, 1998.
- G. Tesauro. Temporal difference learning and td-gammon. *Communications of the ACM*, 38(5): 58-68, 1995.
- C. J. C. H. Watkins and P. Dayan. Technical note: Q-learning. *Machine Learning*, 8:279{292, 1992.
- R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229-256, 1992.
- 中田浩之and 田中利幸. マルコフ決定過程における収益分布の評価. In *情報論的学習理論ワークショップ(IBIS),* 2006.
- 森村哲郎, 杉山将, 八谷大岳, 鹿島久嗣, and 田中利幸. 動的計画法によるリターン分布推定. In *報論的学習理論ワークショップ(IBIS),* 2010.