リスク考慮型強化学習に向けたリターン分布推定

森村 哲郎†*, 杉山 将††, 鹿島 久嗣‡, 八谷 大岳††, 田中 利幸†‡

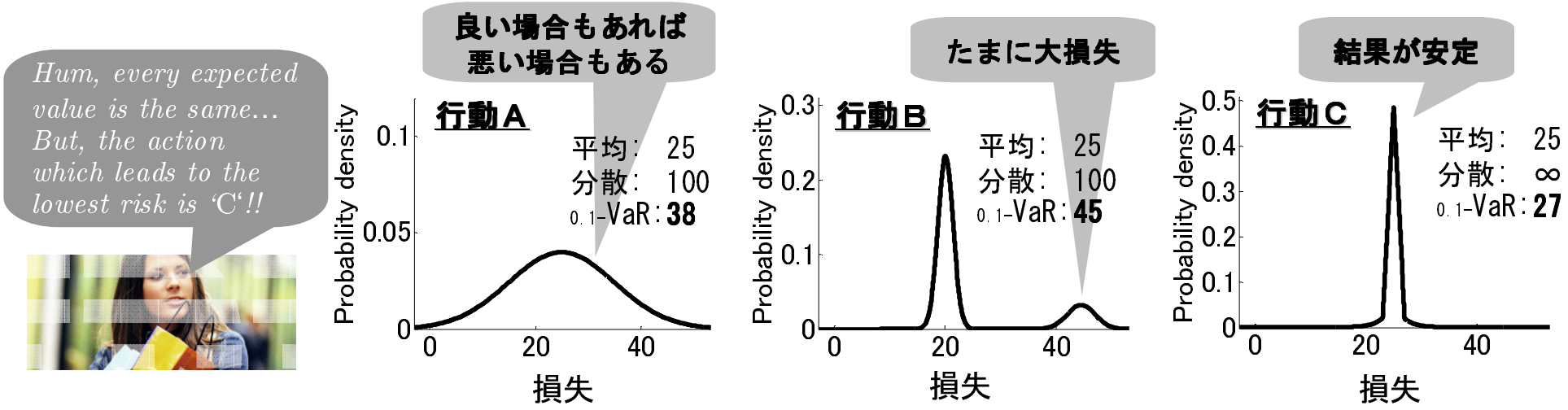
†:IBM東京基礎研究所, ††:東京工業大学, ‡:東京大学, †‡:京都大学　　*tetsuro@jp.IBM.com

➢ 背景

- 意思決定は、各選択肢のもたらす「損失」の期待値をもとに行われるのが普通だが、思いがけず起こる大損失のリスクを回避するには、損失の分布を考える必要がある
- これを逐次的な意思決定の枠組みの中で扱うには、累積損失の分布推定問題を解くことが必要。 しかし、単純な密度推定法では効率が悪く現実的ではない

➢ 成果: リスク回避型意思決定の効率的な実現手法の提案

- 強化学習におけるTD学習の様式に基づく、効率よい累積損失分布の推定法の提案



➢ アプローチ

- リターン分布 $p_\eta(\eta|s)$ におけるBellman方程式（再帰的な式）を利用
  [田中&中田, 2006]

$$p_\eta(\eta|s) = \mathrm{T}\, p_\eta(\eta|s)$$

$$\left(\begin{array}{l} \mathrm{T}: \text{分布ベルマンオペレータ} \\ \mathrm{T}\, p_\eta(\eta|s) \triangleq \frac{1}{\gamma} \sum_{s',a} p_{\mathrm{T}}(s'|s,a) p_a(a|s) \int_r p_r(r|s,a,s') p_\eta(\eta' = \frac{\eta - r}{\gamma} | s')\, dr \end{array} \right)$$

  - 近似リターン分布 $\hat{p}_\eta(\eta|s)$ を $\mathrm{T}\,\hat{p}_\eta(\eta|s)$ に近づけることで学習

- **パラメトリック推定**：パラメータ$\boldsymbol{\theta}$でリターン分布を表現 $\hat{p}_\eta(\eta|s,\boldsymbol{\theta})$
  - $\mathrm{T}\,\hat{p}_\eta(\eta|s,\boldsymbol{\theta})$ から $\hat{p}_\eta(\eta|s,\boldsymbol{\theta})$ のKullback-Leiblerダイバージェンスを小さくする。
    （自然勾配法を用いていて、パラメータ$\boldsymbol{\theta}$を最適化）

- **ノンパラメトリック推定**：パーティクルでリターン分布を表現
  - 下記の漸近性に基づき、現時刻のパーティクルを一時刻先のパーティクルで更新する。
    （パーティクル・スムージングにより、パーティクルを最適化）

$$p_\eta(\eta|s) = \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \delta(r_n + \gamma\eta'_n - \eta).$$

➢ 実験結果：6状態, 4行動 MDP

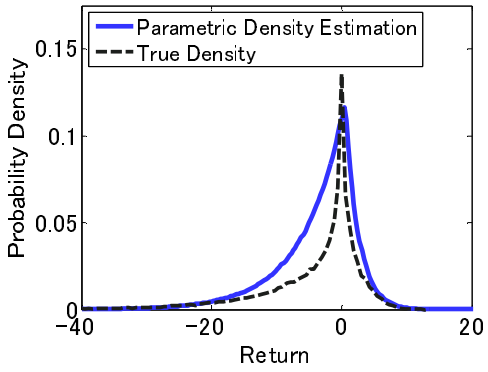
——：推定密度分布、 – – ·：真のリターン密度分布

✓ パラメトリック推定

（非対称ラプラス分布を使用）

収束に必要な試行数: **少**

モデル自由度: **低**



✓ ノンパラメトリック推定

収束に必要な試行数: **多**

モデル自由度: **高**