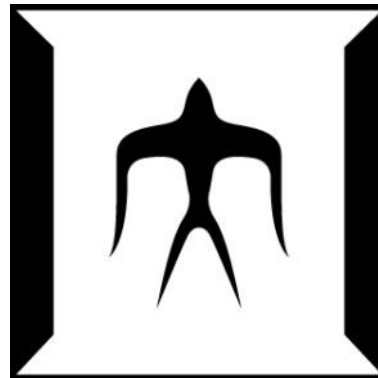


IBISML, 東京大学, 2010年6月15日

New Feature Selection Method for Reinforcement Learning: Conditional Mutual Information Reveals Implicit State-Reward Dependency

八谷大岳 杉山将

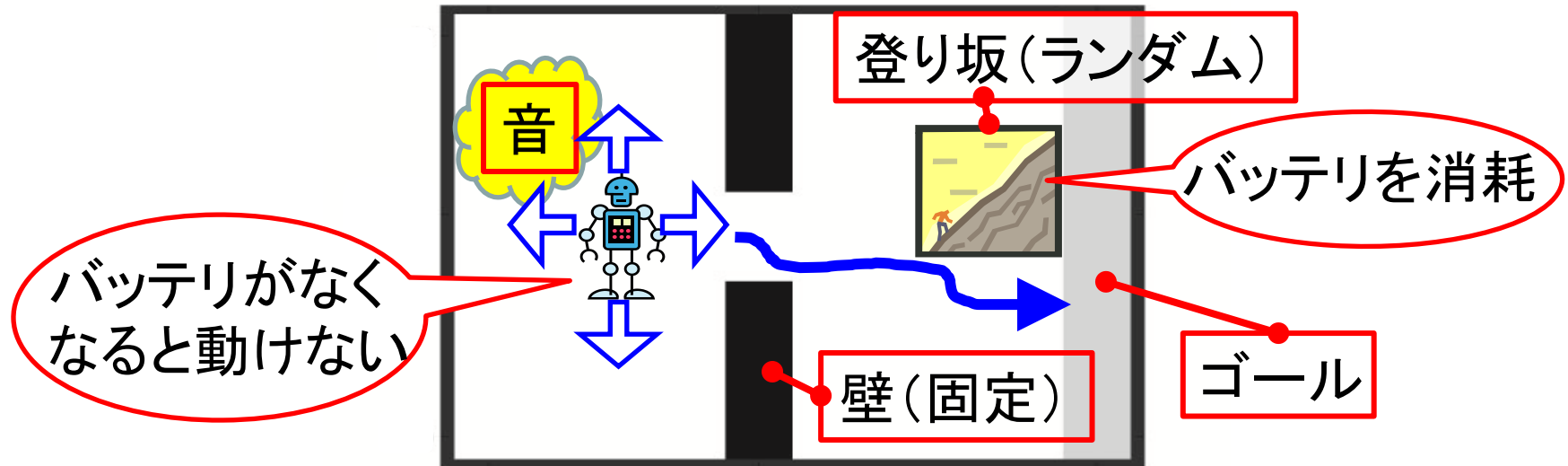
東京工業大学 計算工学専攻



強化学習とは

■ 例) ロボット誘導問題

- センサ s : ロボットの位置、バッテリー残量、傾斜、音など
- 行動 a : 上下左右に一歩移動
- 報酬 r : ゴールにいる時: 1、それ以外: 0
- 政策 $\pi(s)$: センサ s で取る行動 a を決定する関数

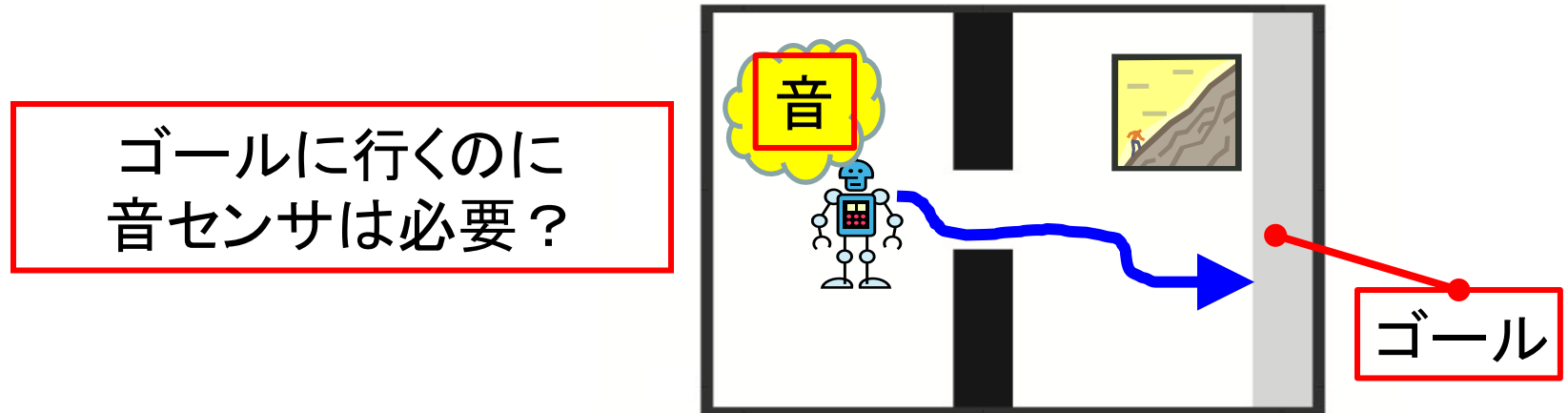


- ## ■ 目的: 報酬和 η を最大化する最適な政策を獲得する

$$\pi^* = \max_{\pi} E[\eta] \quad \eta = r_1 + r_2 + r_3 + \dots$$

本研究の目的

- **問題**: 政策 π の候補はセンサ数に対して指数関数的に増加するため、強化学習の実用化が困難
- 環境によっては、全てのセンサが必要とは限らない



- 環境は未知なので、予めセンサ選択するのは困難
- **目的**: 自動的にセンサを選択(特徴選択)する手法を提案

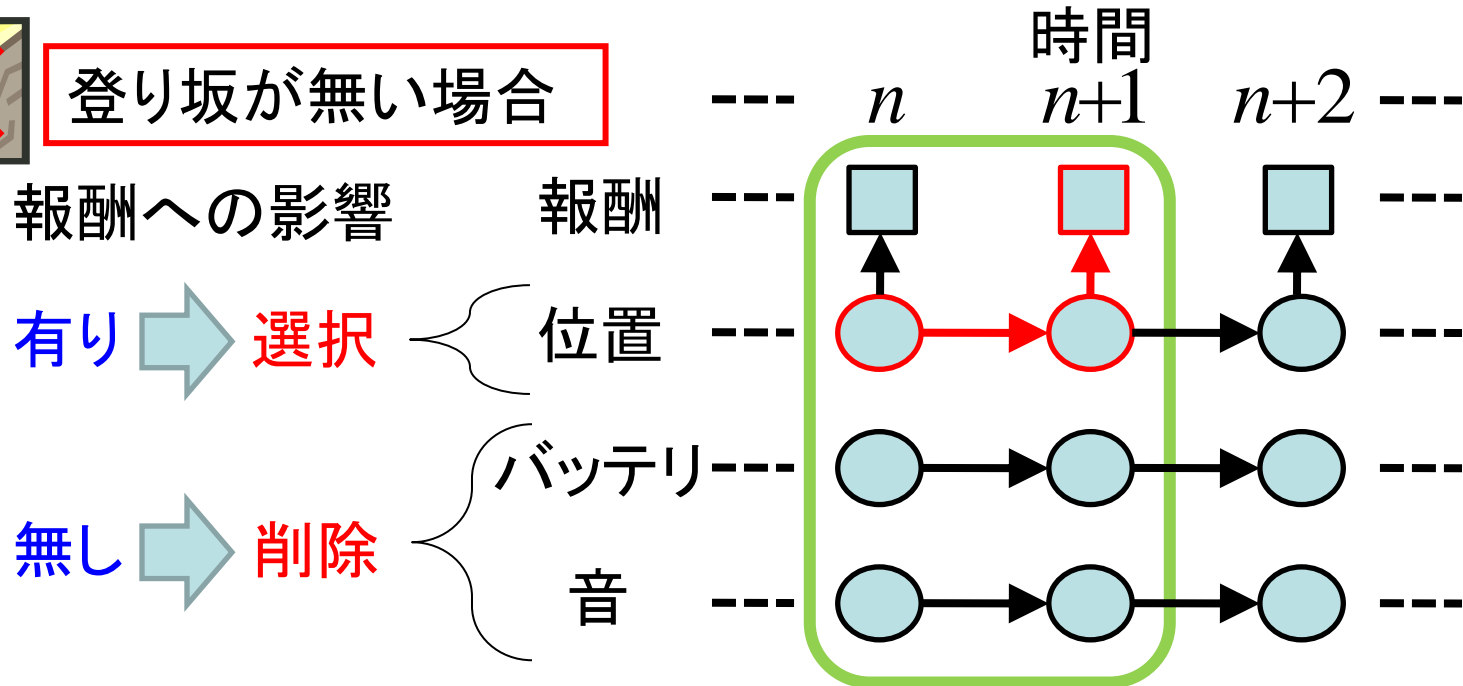
既存手法：報酬とセンサ間の独立性

(Morimoto et al. ICRA 2008)

- 1ステップ先の報酬に影響を与えるセンサを選択
- センサと報酬の独立性：条件付共分散を用いて評価



登り坂が無い場合



報酬はロボットの位置により決まる

- カーネル次元削減法により効率よく実装
(Fukumizu et al. JMLR 2004)

センサと報酬の間接的依存関係

- **問題**: 2ステップ以上の間接的な報酬への影響を考慮していない



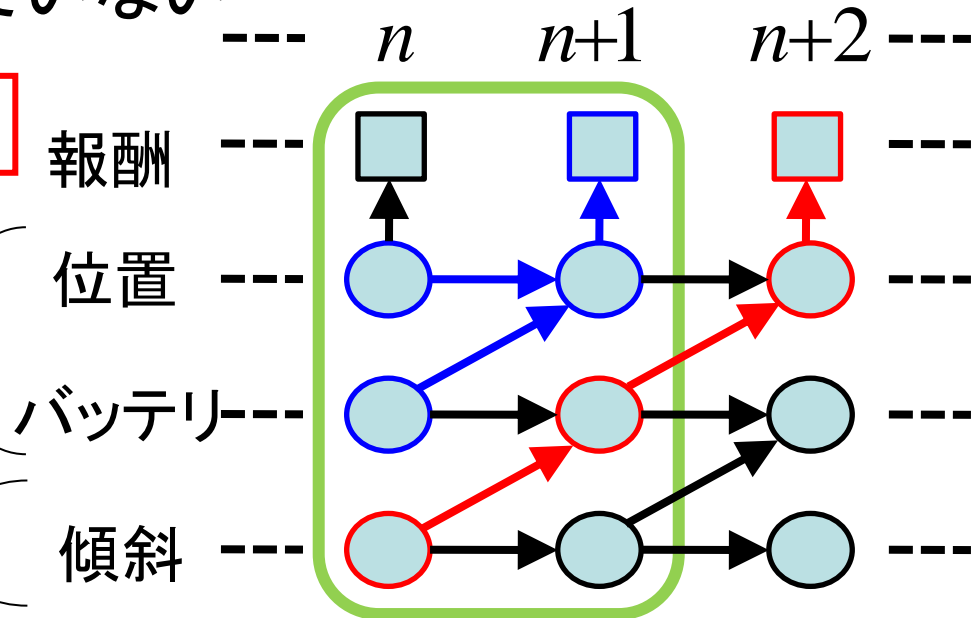
登り坂がある場合

報酬への影響

有り → 選択

無し → 削除

2ステップ後に
影響している！

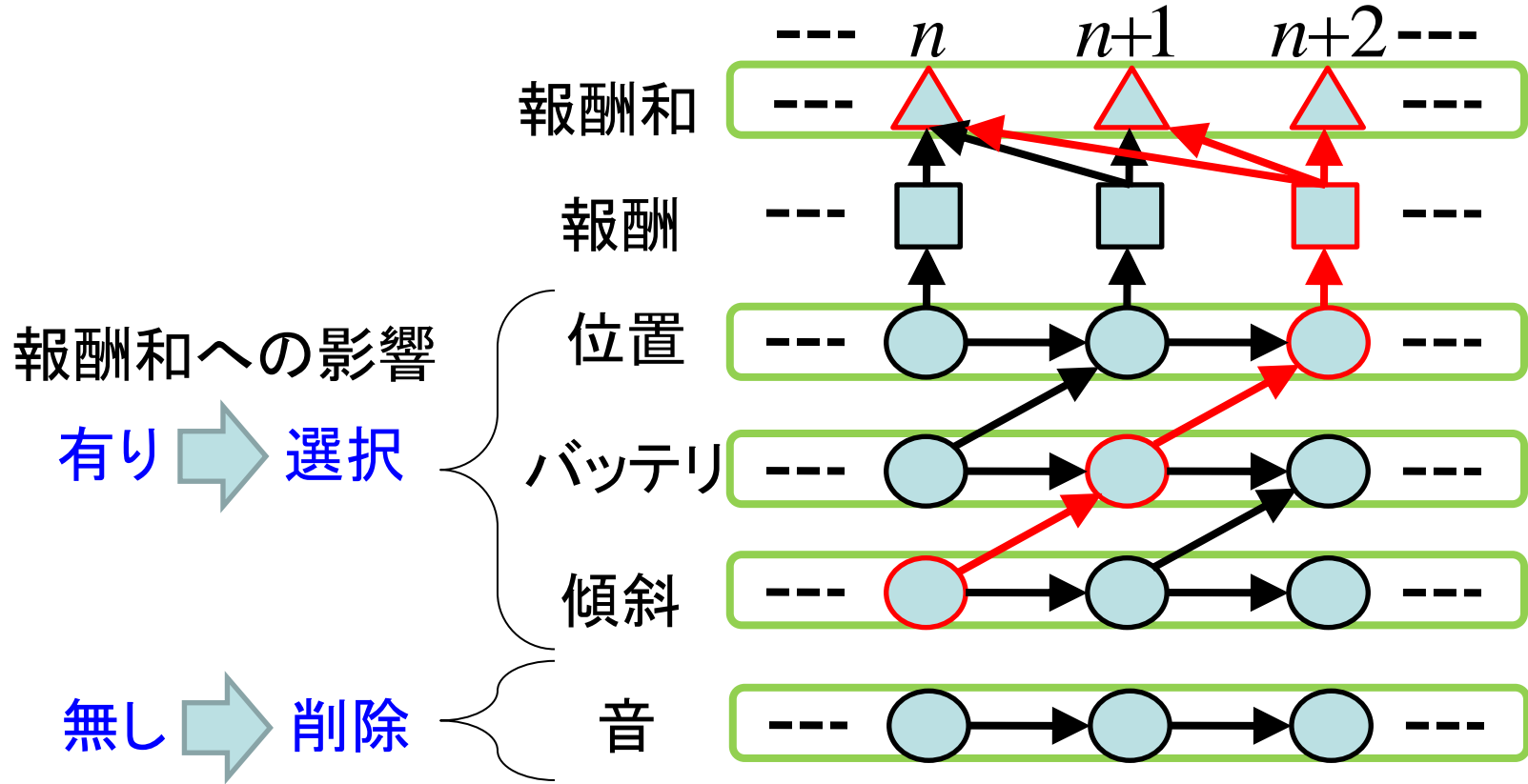


バッテリーなくなると、ロボットは動けない

- **課題**: 複数のステップを介するセンサと報酬間の依存関係を評価したい

提案手法

報酬和とセンサの系列間の独立性を考える



$$\text{報酬和: } \eta_n = r_n + r_{n+1} + r_{n+2} + \dots$$

複数のステップを介すセンサと報酬の依存関係を検出

系列間の独立性

- 条件付相互情報量を用いて評価:

$$I^\pi(\{\eta_n\}_{n=1}^N; \{s_n\}_{n=1}^N) = \frac{1}{N} \sum_{n=1}^N I^\pi(\eta_n; s_n) \quad N: \text{ステップ数}$$

$I^\pi(\eta_n; s_n)$: 政策 π に従った際の時間 n での報酬和 η とセンサ s 間の相互情報量

- 性質: 条件付相互情報量を最大化することにより、報酬和とセンサ集合の条件付独立性が得られる

$$\mathbf{z}^* = \arg \max_{\mathbf{z}} I^\pi(\{\eta_n\}_{n=1}^N; \{\mathbf{z}_n\}_{n=1}^N) \leftrightarrow \eta_n \perp \mathbf{s}_n \mid \mathbf{z}_n^*, \forall n$$

\mathbf{s} : センサの集合

\mathbf{z} : センサの部分集合

順方向特徴選択アルゴリズム

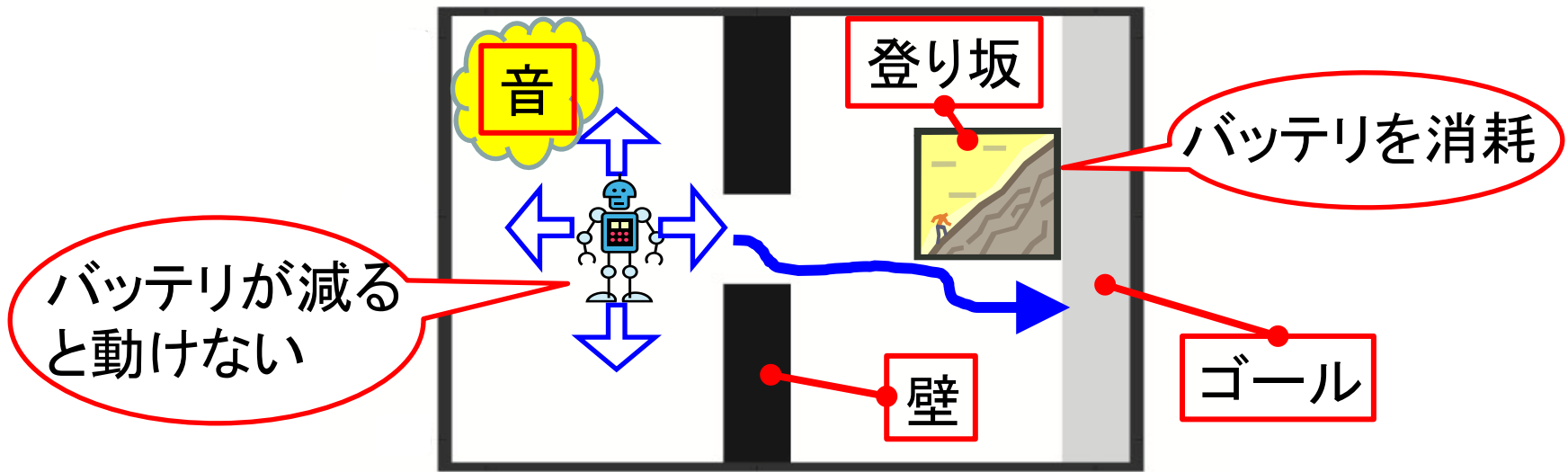
1. 初期化: $\mathbf{s} = \{s^{(1)}, s^{(2)}, \dots, s^{(V)}\}$ $\mathbf{z} = \{\}$
2. 条件付相互情報量を最大化するセンサ $s^{(*)}$ を求める

$$s^{(*)} = \arg \max_{s^{(i)}} \hat{I}^{\pi}(\{\eta\}; \{\mathbf{z} \cup s^{(i)}\})$$

3. センサ $s^{(*)}$ を \mathbf{s} から削除し、 \mathbf{z} に追加
4. 2と3を U 回繰り返す
5. $\mathbf{z} = \{z^{(1)}, z^{(2)}, \dots, z^{(U)}\}$ を出力

- 相互情報量の推定: 最小二乗相互情報量推定法を用いる (Suzuki et al. BMC Bioinformatics 2009)

実験：ロボット誘導問題

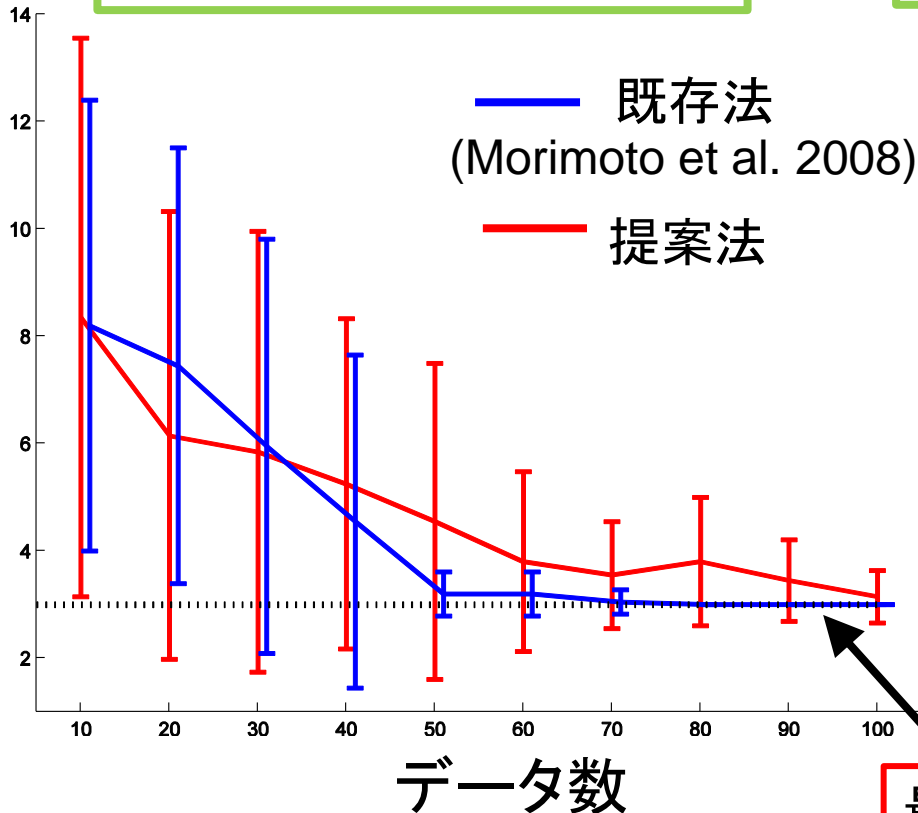


- センサ：位置 (x, y 座標)、バッテリー、傾斜に加え、10次元のガウスノイズ(音を含む)を追加
- サンプル：ランダムな政策に従い20ステップ進むことを繰り返して収集

実験結果

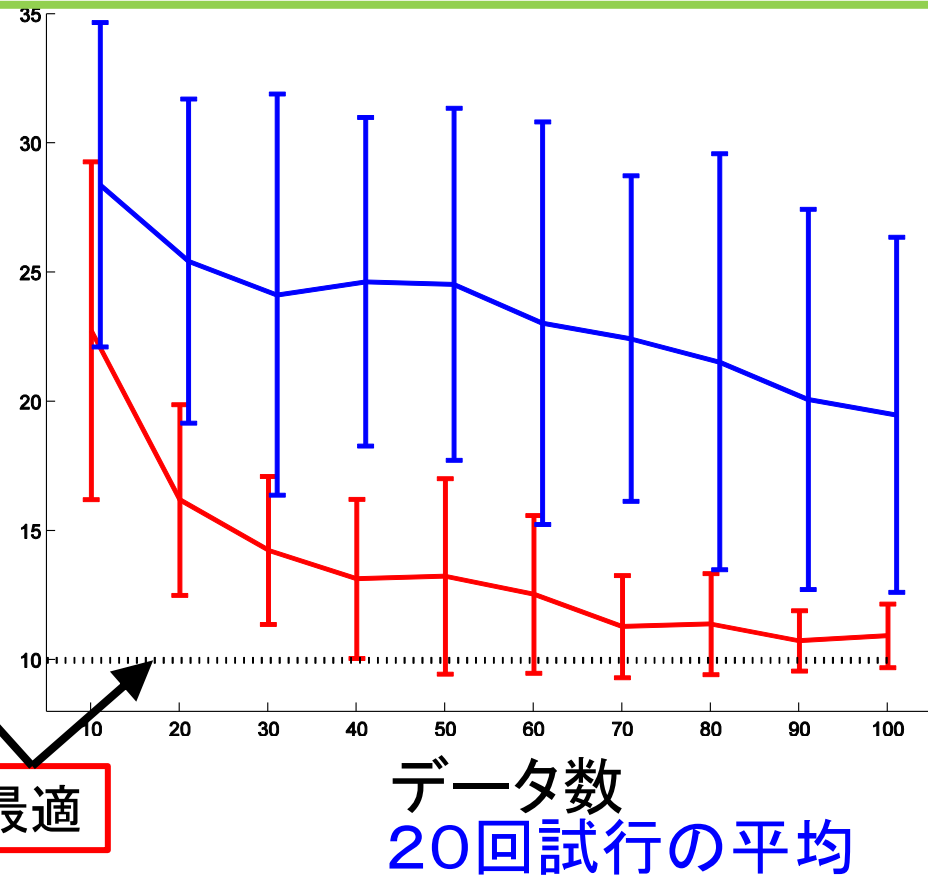
登り坂がない場合

位置(x,y座標)の順位之和



登り坂がある場合

位置(x,y座標)、バッテリー、傾斜の順位之和



- 提案法は間接的な依存関係がある場合でも性能が良いことを示した。

まとめ

- 強化学習：センサの数が多いと実用化が困難
- 既存の特徴選択法：報酬とセンサ間の2ステップ以上の間接的な依存関係を考慮していない
- 提案特徴選択法：
 - 報酬和とセンサ系列間の独立性を用いる
 - 条件付相互情報量を用いて独立性を評価
 - 順方向特徴選択アルゴリズムにより実装
- 実験を通して提案手法の有効性を確認