限られたデータを用いた強化学習の成功条件について

オフライン強化学習に関する最近の理論的進展

LINEヤフー研究所

宮口航平

2024年度CIGS経済・社会の分野横断的研究会（2024/12/26）

自己紹介

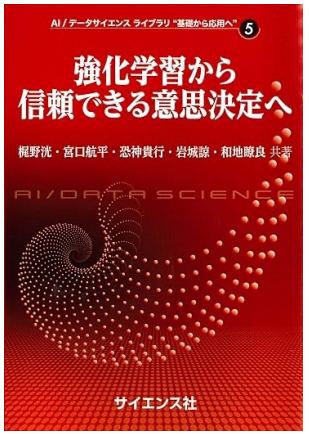
宮口航平

~ 2019年　　東京大学情報理工学系研究科（博士）
- 数理情報学専攻 第６研究室

~ 2024年 12月　IBM東京基礎研究所

~ 現在　　LINEヤフー研究所

研究分野
- 学習理論（に基づくエキゾチックなデータ分析）
- 社会で役に立つ強化学習
  - 理論と応用のギャップをいかに埋めるか



Agenda

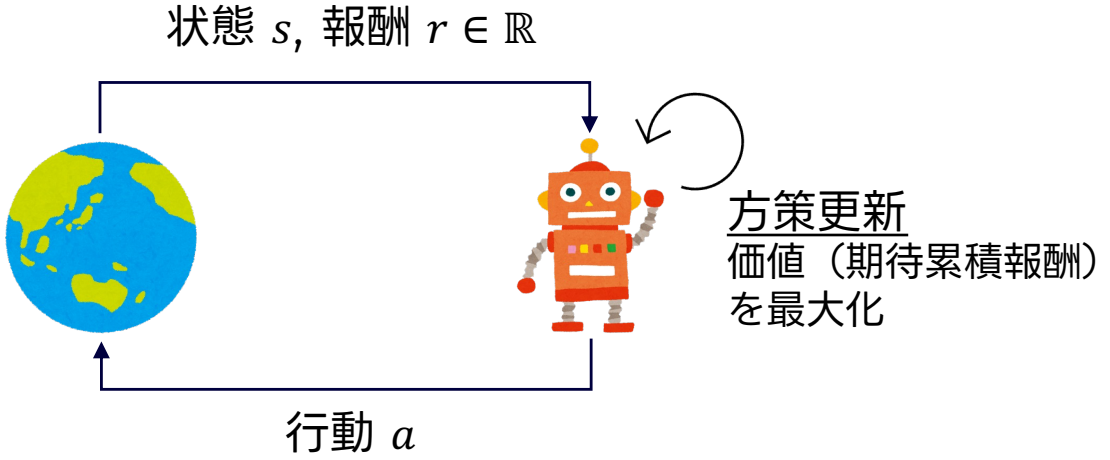
1. 強化学習：意思決定最適化のための機械学習

2. オフライン強化学習：限られたデータを用いた強化学習

3. オフライン強化学習の成功条件に関する理論的進展
   - M. (2024). Worst-Case Offline Reinforcement Learning with Arbitrary Data Support. *NeurIPS'24*.

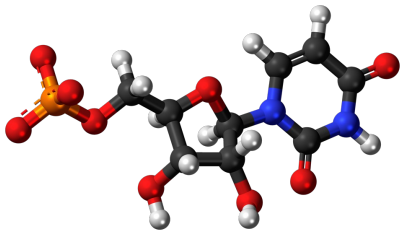
強化学習

意思決定最適化のための機械学習

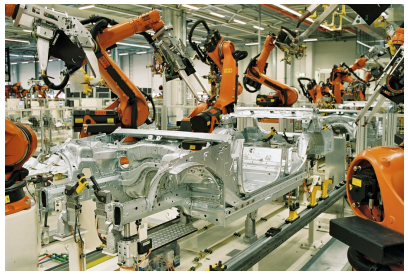


- ✓ 環境とそれに適応するエージェントのモデル
- ✓ 推定/計画/実行といった複数能力の統合が要

強化学習の応用先



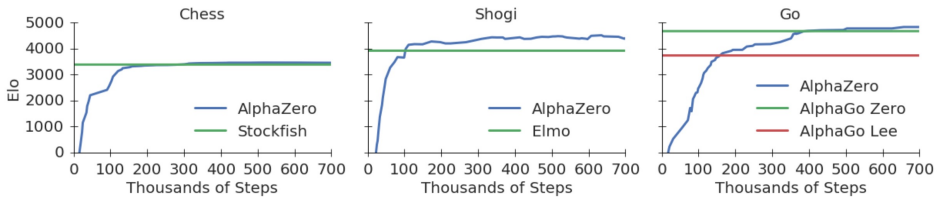



✓ 特に長期的な目標の達成が絡む分野で役立つ

（ゲーム、組み立て、制御、ユーザー体験、運用実績、治療効果、アラインメント、論理的思考…）

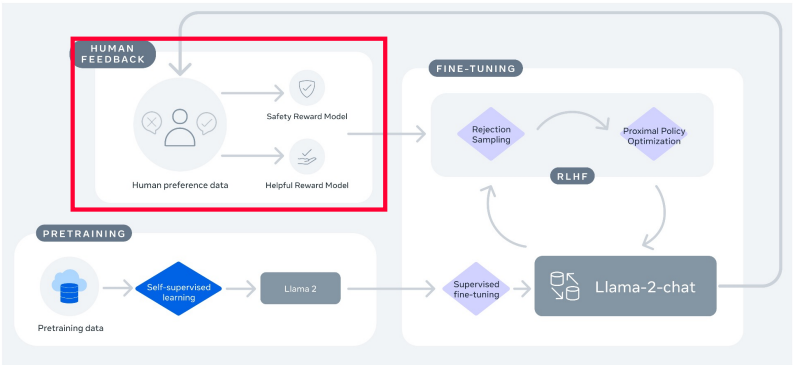
これまでの強化学習の成功条件：大量の試行錯誤

AlphaZeroの訓練 [S+'17]



	Chess	Shogi	Go
Mini-batches	700k	700k	700k
Training Time	9h	12h	34h
Training Games	44 million	24 million	21 million
Thinking Time	800 sims 40 ms	800 sims 80 ms	800 sims 200 ms

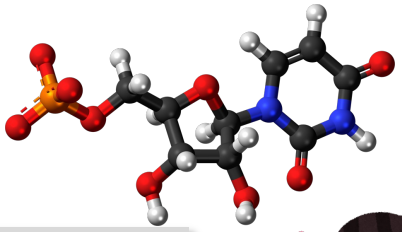
Llama2のRLHFデータ [T+'23]



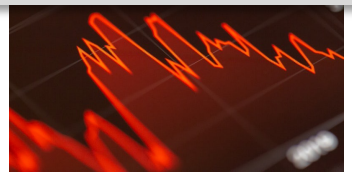
Dataset	Num. of Comparisons	Avg. # Turns per Dialogue	Avg. # Tokens per Example	Avg. # Tokens in Prompt	Avg. # Tokens in Response
Meta (Safety & Helpfulness)	1,418,091	3.9	798.5	31.4	234.1
Total	2,919,326	1.6	595.7	108.2	216.9

- ✓ 今日の強化学習の成功はオンライン（本番環境）での多くの試行錯誤が前提
- ✓ 理論的にも強化学習を解くためのカギは試行錯誤の幅と数 [W'89,WD'92]

実際には「大量の試行錯誤」は困難



実社会に強化学習を役立てる上で大きな障壁の一つ



実行可能性の問題

安全性・倫理・規制の観点

コストの問題

お金・時間・品質の観点

Agenda

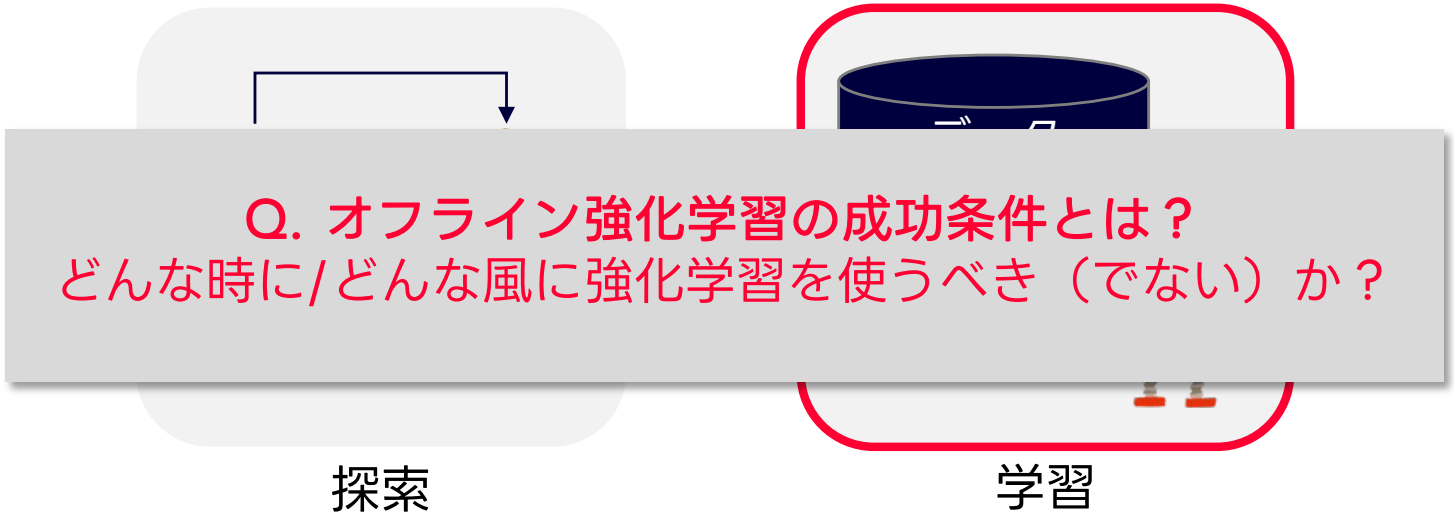
1. 強化学習：意思決定最適化のための機械学習

2. オフライン強化学習：限られたデータを用いた強化学習

3. オフライン強化学習の成功条件に関する理論的進展
   - M. (2024). Worst-Case Offline Reinforcement Learning with Arbitrary Data Support. *NeurIPS'24*.

オフライン強化学習

限られたデータを用いた強化学習（2020年~）

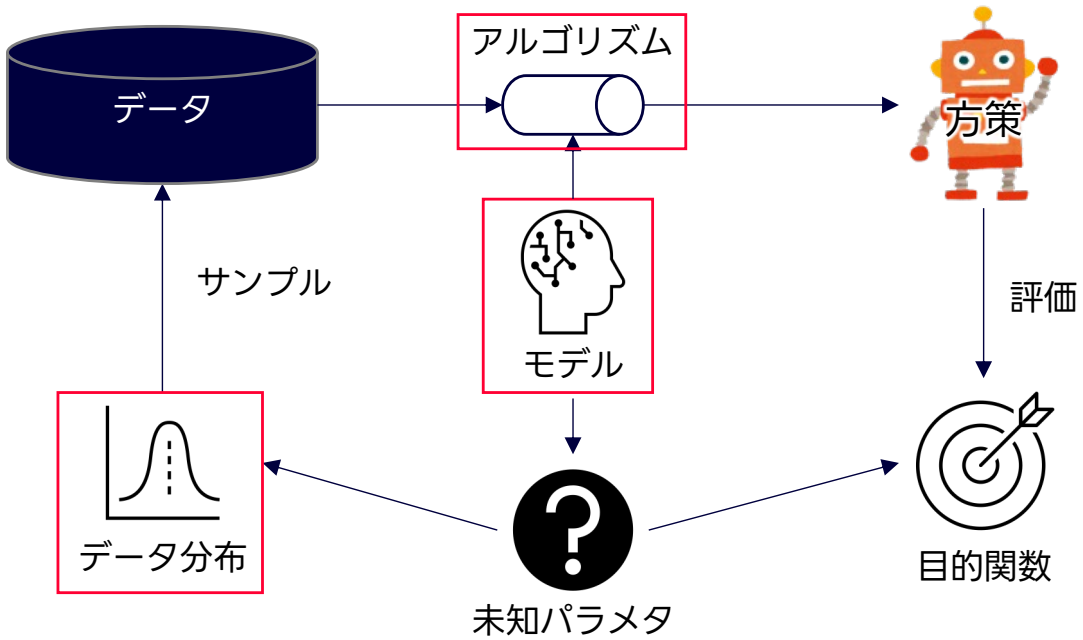


Q. オフライン強化学習の成功条件とは？
どんな時に/どんな風に強化学習を使うべき（でない）か？

探索　　学習

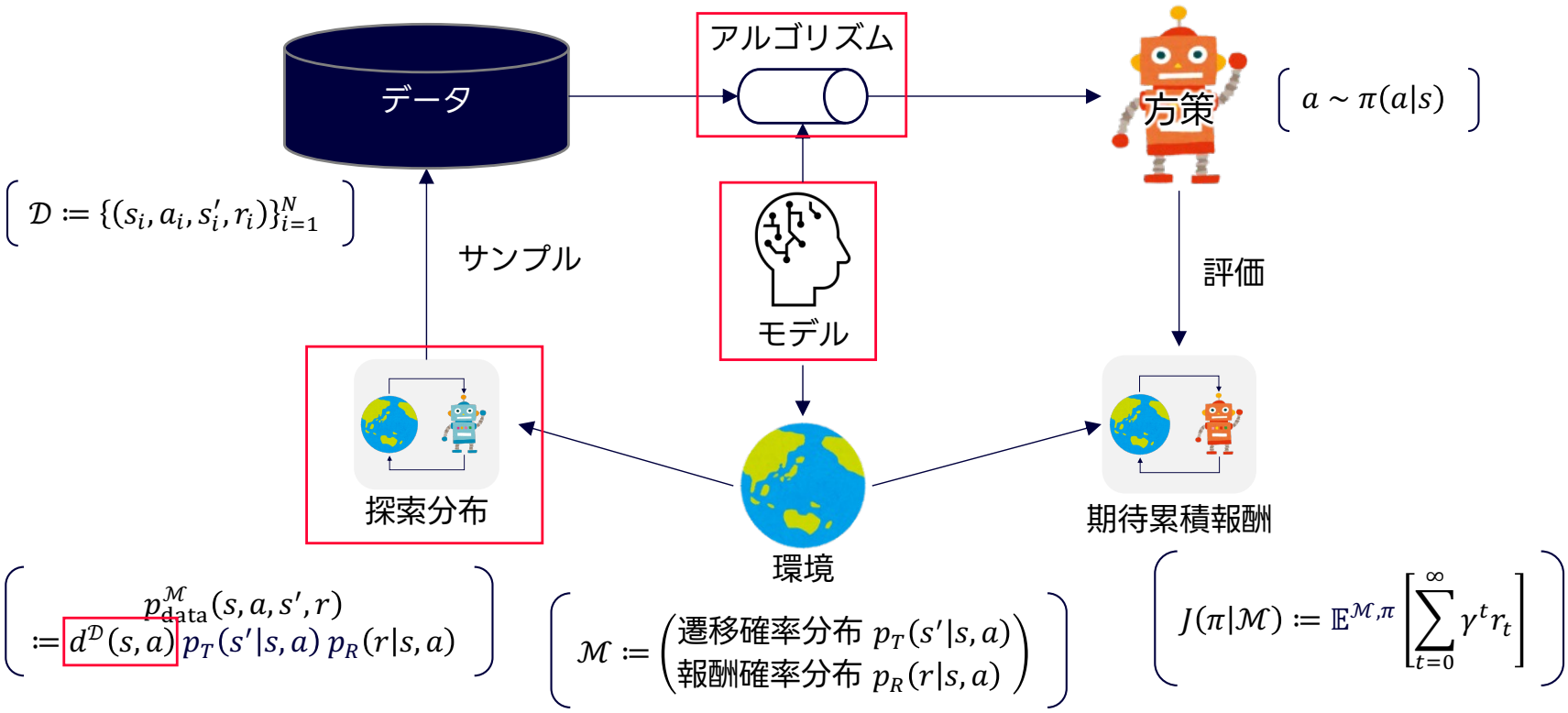
- ✓ 強化学習を探索フェーズと学習フェーズに分解
- ✓ 探索固定で勝負する強化学習

統計的学習理論による定式化



✓ 学習が成功するかどうかはモデル・データ分布・アルゴリズム次第

統計的学習理論による定式化



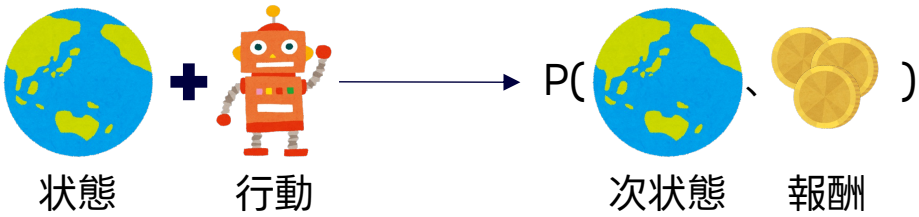
オフライン強化学習の（代表的な）成功条件

	モデル	探索分布	アルゴリズム	誤差保証
従来結果 1 [US'21]	環境	網羅的	適度に保守的	$O\left(\frac{1}{\sqrt{N}}\right)$
従来結果 2 [Z+'22]	価値	網羅的	適度に保守的	$O\left(\frac{1}{\sqrt[6]{N}}\right)$

✓ オフライン強化学習の盛り上がりと共にここ数年で急速に理論的理解が進んだ

モデルの条件：表現可能性

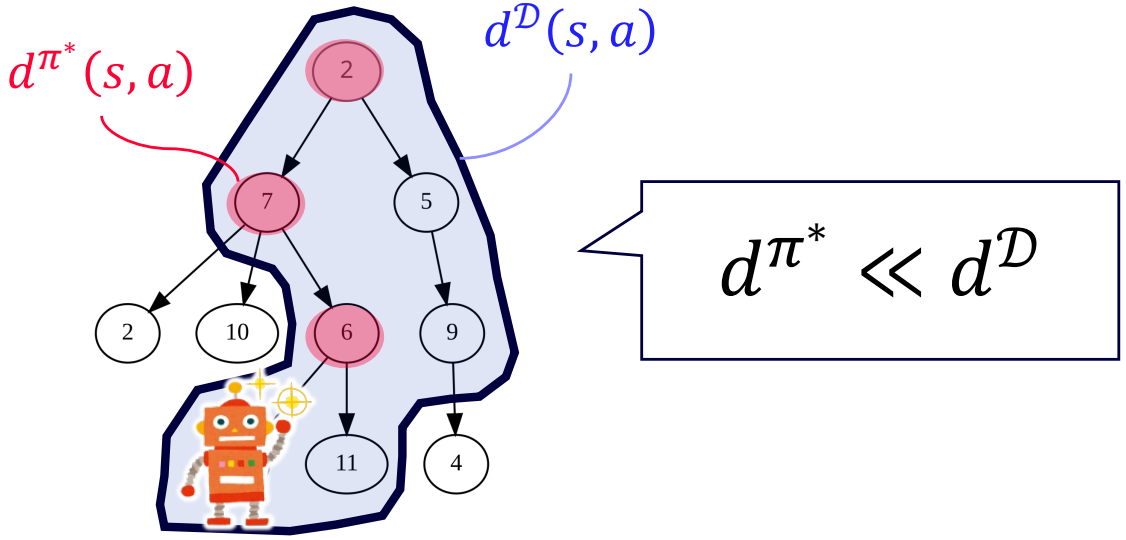
1. 環境モデル
（モデルベースRL）



2. 価値モデル
（モデルフリーRL）

✓ 真の環境 or 真の価値関数を近似的に表現可能であること
（難易度：環境モデルの表現 > 価値モデルの表現）

探索分布の条件：網羅性



✓ <u>最適方策の訪問先</u>を網羅していること

（全状態を網羅する必要はない）

アルゴリズムの条件：保守的正則化

1. 方策正則化

$$\tilde{J}(🤖) = J(🤖) - \alpha D(🤖, 🤖)$$

目的関数

探索方策からのズレ

2. 価値正則化

$$\tilde{Q}(🌍, 🤖) = Q(🌍, 🤖) - \frac{\alpha}{\sqrt{N(🌍, 🤖)}}$$

行動価値

訪問回数

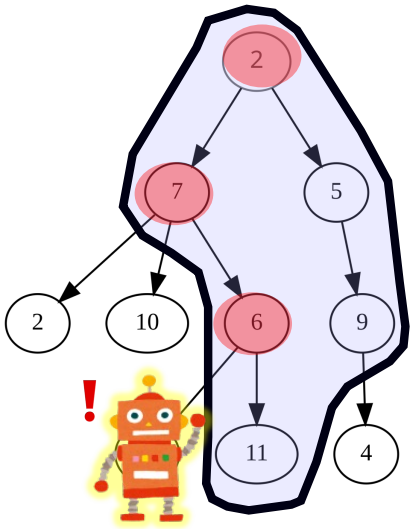
✓ 適度に保守的な正則化によって、不確実な情報に基づく過度な最適化を回避すること

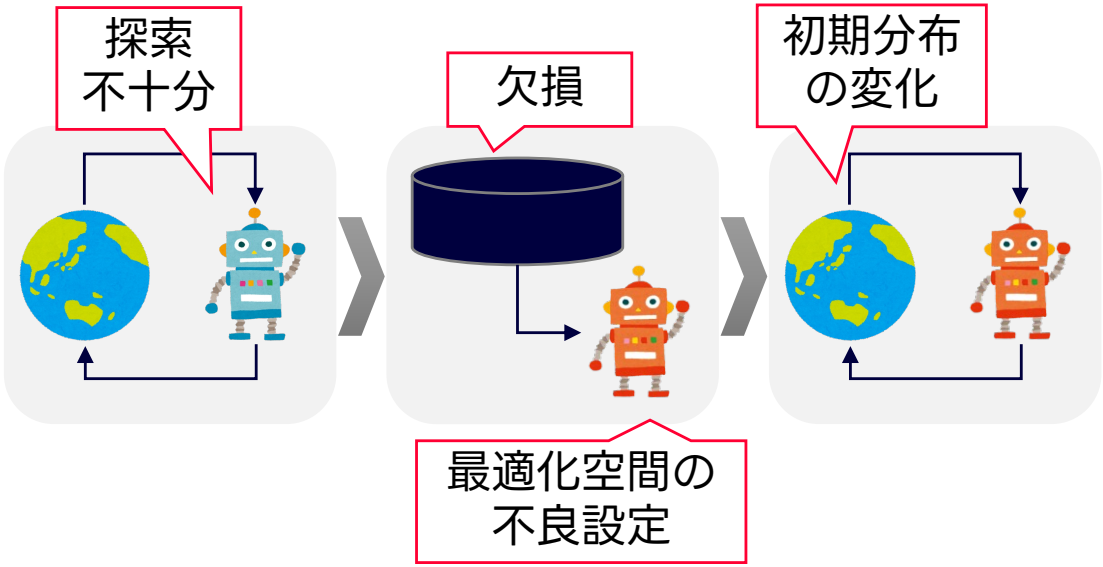
従来条件の課題

	モデル	探索分布	アルゴリズム	誤差保証
従来結果 1 [US'21]	環境	網羅的	適度に保守的	$O\left(\frac{1}{\sqrt{N}}\right)$
従来結果 2 [Z+'22]	価値	網羅的	適度に保守的	$O\left(\frac{1}{\sqrt[6]{N}}\right)$

- ✓ 探索分布の網羅性がないと何も言えない
- ✓ 保守的正則化の重み α はオンラインで検証しながら調整する必要あり
- ✓ 価値関数をモデル化する場合に誤差保証が悪化

探索分布の網羅性が破れる要因は多い





- ✓ 追加の探索なしに網羅性を回復するのは困難
- ✓ 網羅性が破れているかを判定するの自体も非自明

Agenda

1. 強化学習：意思決定最適化のための機械学習

2. オフライン強化学習：限られたデータを用いた強化学習

3. オフライン強化学習の成功条件に関する理論的進展
   - M. (2024). Worst-Case Offline Reinforcement Learning with Arbitrary Data Support. *NeurIPS'24*.

新しい結果

	モデル	探索分布	アルゴリズム	誤差保証
従来結果 1 [US'21]	環境	網羅的	適度に保守的	$O\left(\frac{1}{\sqrt{N}}\right)$
従来結果 2 [Z+'22]	価値	網羅的	適度に保守的	$O\left(\frac{1}{\sqrt[6]{N}}\right)$
新結果 [M'24]	価値	-	-	$O\left(\frac{1}{\sqrt{N}}\right)^{\dagger}$

†: ベストエフォート保証

- ✓ 探索分布の網羅性がない場合はベストエフォート保証（後述）に自動切替
- ✓ 保守的正則化重みのオンライン調整は不要（オフラインで調整可能）
- ✓ 価値をモデル化する場合でも（N に関して）最適な誤差保証を達成

ベストエフォート保証の原理

■ 通常の方策価値

$$J(\pi|\mathcal{M}) \coloneqq \mathbb{E}^{\mathcal{M},\pi}\left[\sum_{t=0}^{\infty}\gamma^t r_t\right]$$

■ 最悪方策価値

$$\tilde{J}(\pi) \coloneqq \min_{\mathcal{M}'\in\mathfrak{U}} J(\pi|\mathcal{M}')$$

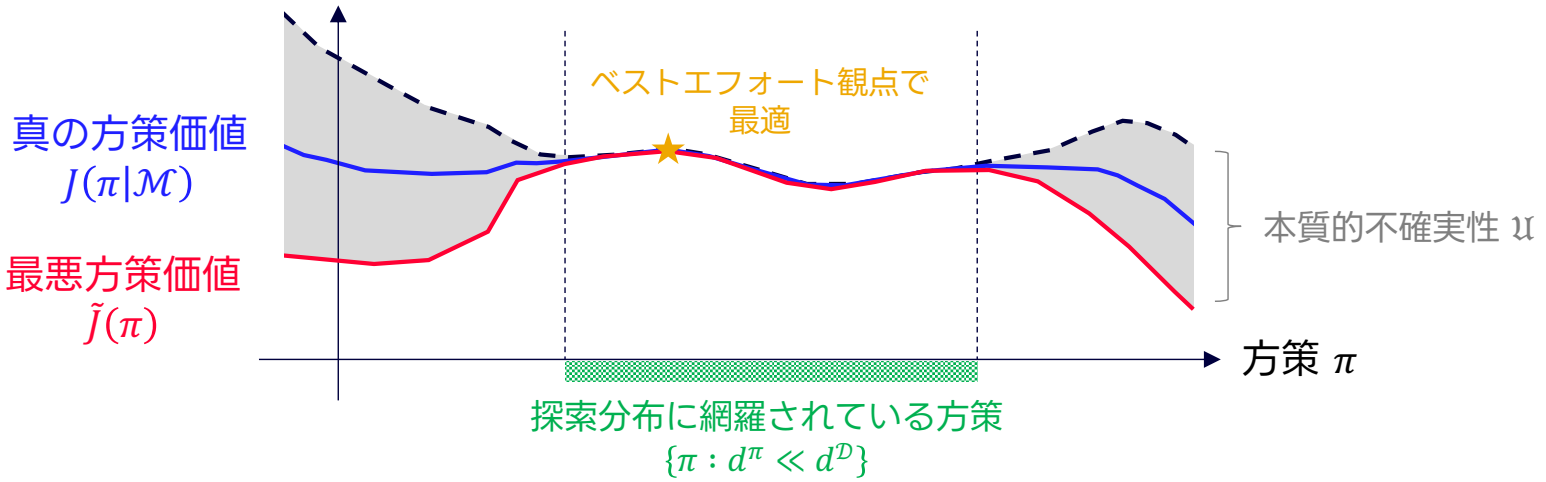
本質的不確実性集合

$$\mathfrak{U} \coloneqq \left\{\mathcal{M}' \in \mathfrak{M} : p^{\mathcal{M}}_{\text{data}} = p^{\mathcal{M}'}_{\text{data}}\right\}$$

※ $\mathfrak{M}$：報酬が規格化された環境全体

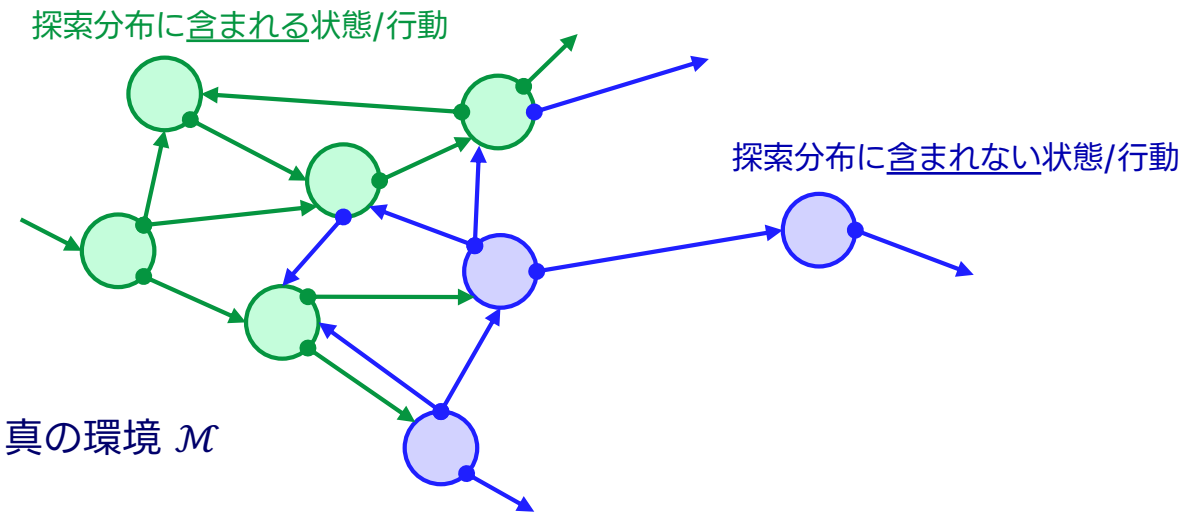
✓ 目的関数を通常の方策価値から<u>最悪方策価値</u>へと変更

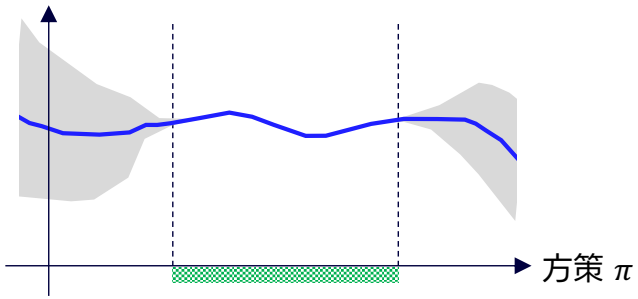
ベストエフォート保証の原理：図解



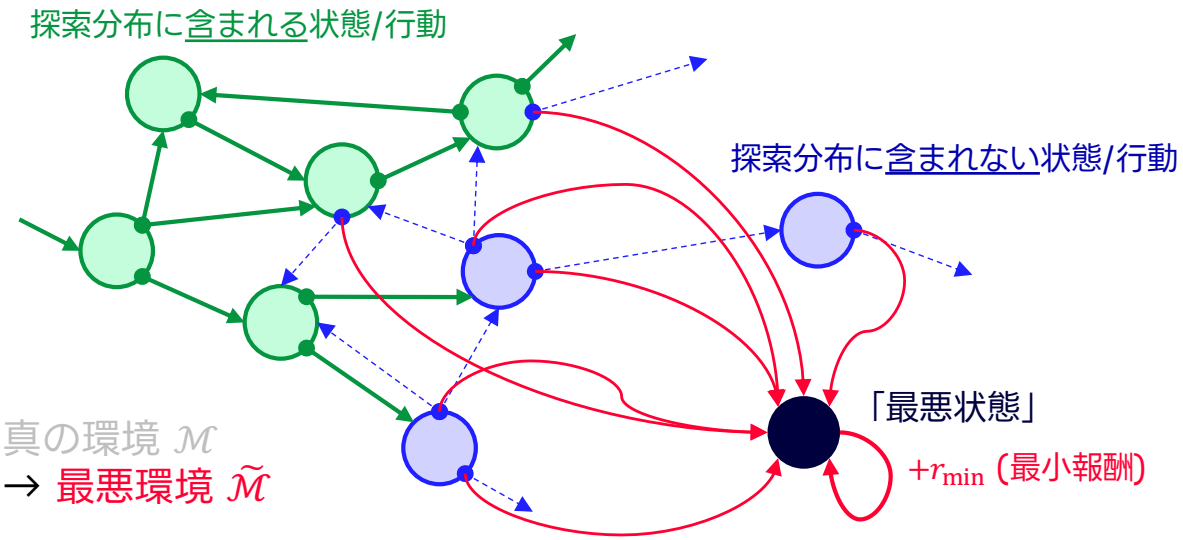
- ✓ 推定可能な方策価値の下界の中で最大（→「ベストエフォート」評価）
- ✓ 保守性 by design（→アルゴリズム側で保守性を考慮する必要なし）

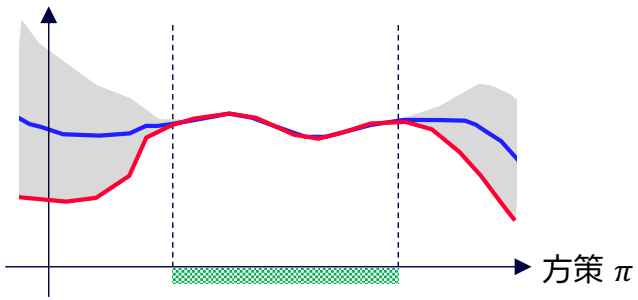
提案アルゴリズムの概要





提案アルゴリズムの概要

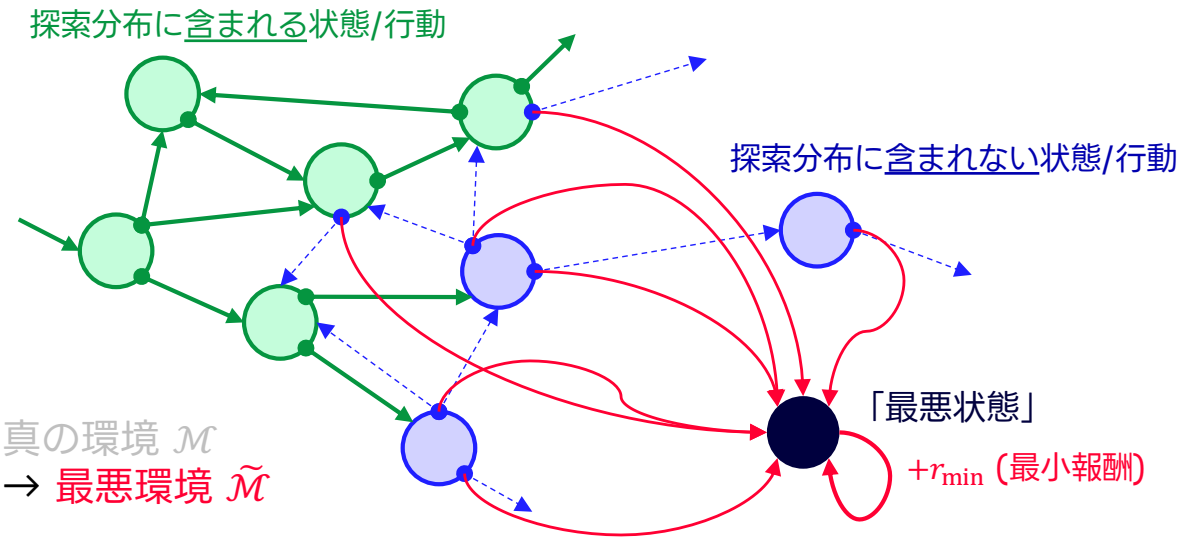


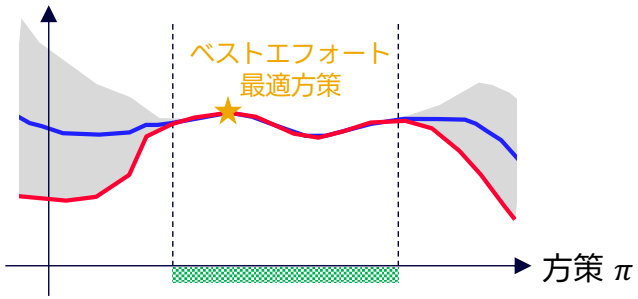


✓ 探索分布に含まれない状態行動対の遷移先を全て「最悪状態」に変更して最悪環境 $\widetilde{\mathcal{M}}$ を構成

→ $\widetilde{\mathcal{M}}$ での評価 $J(\pi|\widetilde{\mathcal{M}})$ とベストエフォート評価 $\tilde{J}(\pi)$ が一致

提案アルゴリズムの概要





✓ 本研究では対最悪環境 $\widetilde{\mathcal{M}}$ の強化学習を解くことを提案
（従来手法のマイナーチェンジでOK）

得られた結果（再掲）

	モデル	データ分布	アルゴリズム	誤差保証
従来結果 1 [US'21]	環境	網羅的	適度に保守的	$O\left(\frac{1}{\sqrt{N}}\right)$
従来結果 2 [Z+'22]	価値	網羅的	適度に保守的	$O\left(\frac{1}{\sqrt[6]{N}}\right)$
新結果 [M'24]	価値	-	-	$O\left(\frac{1}{\sqrt{N}}\right)^{\dagger}$

†: ベストエフォート保証

✓ 目的関数を最悪方策価値に変更することで、より実用的な成功条件が示せた

まとめ

Q. オフライン強化学習の成功条件とは？

~~A. 環境/価値モデルが正しく、データが網羅的（かつ価値をモデル化するなら大量）で、アルゴリズムが適度に保守的なとき。~~

A. 価値モデルが正しいとき。

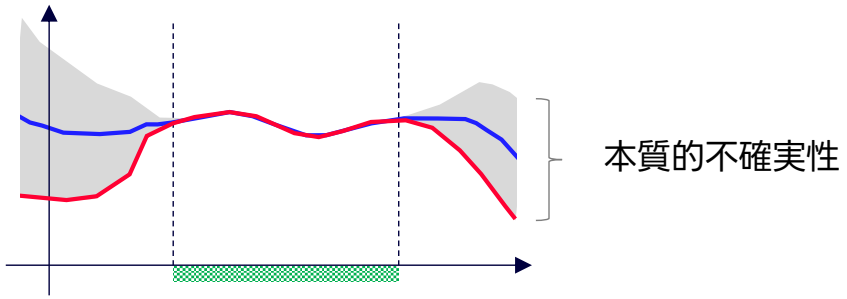
（※データが網羅的でない場合はベストエフォート評価を基準とする）

重要な含意）オフライン強化学習では最悪方策価値を評価すると良い

- 必要に応じて自動でベストエフォート保証に切り替わる
- 保守性のチューニング（=オンライン検証のコスト）が不要

今後の展望

- ベストエフォート評価の解釈：本質的不確実性の区間推定



- さらなる条件の緩和：正しい価値モデルの自動選択
- 実用的なアルゴリズムの開発・検証

参考文献

- Watkins, C. J. (1989). *Learning from delayed rewards*.
- Watkins, C. J., & Dayan, P. (1992). Q-learning. *Machine learning*.
- Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., ... & Hassabis, D. (2017). Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv*.
- Uehara, M., & Sun, W. (2021). Pessimistic model-based offline reinforcement learning under partial coverage. *ICLR'22*.
- Zhan, W., Huang, B., Huang, A., Jiang, N., & Lee, J. (2022). Offline reinforcement learning with realizability and single-policy concentrability. *COLT'22*.
- Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... & Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv.*
- Miyaguchi, K. (2024). Worst-Case Offline Reinforcement Learning with Arbitrary Data Support. *NeurIPS'24*.

画像引用元

1. Self-driving car, https://commons.wikimedia.org/w/index.php?curid=67227170
2. Smart manufacturing, https://commons.wikimedia.org/w/index.php?curid=11928438
3. Market crash, https://commons.wikimedia.org/w/index.php?curid=153036498
4. Diffusion model, https://arxiv.org/abs/2209.00796
5. Sepsis, https://commons.wikimedia.org/w/index.php?curid=87720795
6. いらすとや, https://www.irasutoya.com/