部分的フィードバックに基づくオンライン凸最適化 ー未知の環境に適応する意思決定技術ー

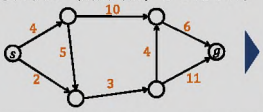
伊藤伸志（NEC）

オンライン最適化とは

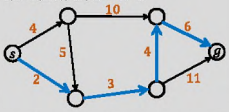
通常の最適化問題（**オフライン最適化**）：
事前に目的関数 f が与えられた状況で解 x を出力

f を入力 ▶ x を出力　Goal: Minimize $f(x)$

例：各枝の長さ（所要時間）が与えられ、



最短経路を見つける



最短経路長:
2 + 3 + 4 + 6 = 15

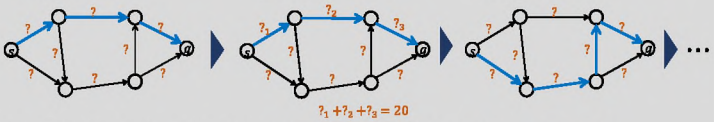
☹ 実用上の制約：評価指標を事前に知っていることが必要

評価指標が事前にわからない／時々刻々と変化する状況、
例えば道の混雑度が常に変化する状況に対応するには？
→ **オンライン最適化** の枠組みが有効

部分的（バンディット）フィードバックオンライン最適化：
解 x_t を出力した**後で**目的関数値 $f_t(x_t)$ だけわかる

x_1 を出力 ▶ f_1 を一部観測 ▶ x_2 を出力 ▶ f_2 を一部観測 ▶ …

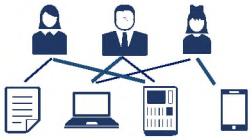
例：校長さがわからない状況で経路をきめる　あとになって、選んだ経路の長さがわかる　同様に繰り返し（枝重みは変化）



☺「試してみて初めて結果がわかる」状況にも対応

幅広い応用先

- 機械学習モデルの自動更新
- ユーザー傾向・需要変動に応じた商品推薦・広告表示
- 需要変動に応じた価格設定，販売・流通計画の最適化



オンライン最適化のゴール

リグレット R_T を小さくするアルゴリズムの開発

$$R_T := \sum_{t=1}^{T} f_t(x_t) - \min_{x^* \in A} \sum_{t=1}^{T} f_t(x^*)$$

アルゴリズムの実績値　　最良の固定戦略の実績値

- リグレット R_T が小さいことは，**最適に近い意思決定**ができていることを意味する．

本研究の位置づけ

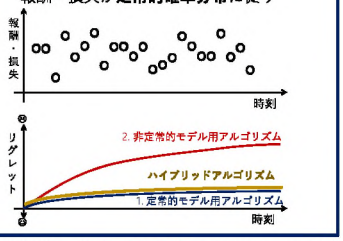
- **既存技術**：環境 $\{f_t\}$ のふるまい（定常的／非定常的など）の設定に応じて，様々なアルゴリズムが存在
  ☹ 活用するには，環境についての事前知識が不可欠
- **本研究**：様々なアルゴリズムの強みを両立する技術
  ☺ 事前知識なしで，未知の環境に自動で適応

成果１ ハイブリッド型オンライン最適化

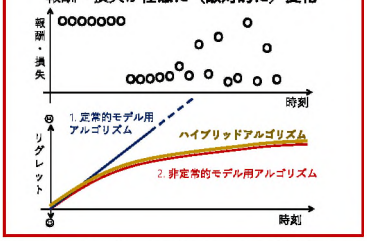
複数のアルゴリズムの特長を併せもつアルゴリズムを提案
NeurIPS2020, NeurIPS2021, COLT2021 などで発表
[Ito et al., 2020] [Ito, 2021b] [Ito, 2021a]

２つの環境モデルとアルゴリズム

1. **定常的（確率的）モデル**
報酬・損失が**定常的確率分布**に従う



2. **非定常的（敵対的）モデル**
報酬・損失が**任意に（敵対的に）変化**



従来技術：**定常**／**非定常** を見極め，適材適所でアルゴリズムを選択することが必要
ハイブリッドアルゴリズム：定常／非定常が未知の状況でも自動で最適な性能を達成

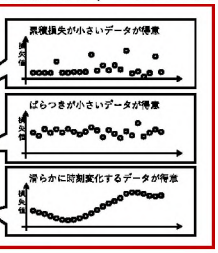
多腕バンディット問題における貢献

有限の選択肢から行動を決める最も基本的な問題設定　N: 選択肢の個数，T: ラウンド数

1. **定常的（確率的）モデル**
- UCB アルゴリズムなど:
$R_T = O\left(\frac{N \log T}{\Delta}\right)$
[Lai and Robbins, 1985; Auer et al., 2002a]

Δ: suboptimality gap parameter
（最適戦略の識別の難しさを司るパラメタ）

2. **非定常的（敵対的）モデル**
- EXP3 アルゴリズムなど:
$R_T = \tilde{O}\left(\sqrt{NT}\right)$ [Auer et al., 2002b]
- $R_T = \tilde{O}\left(\sqrt{N \sum_{t=1}^{T} \ell_{t i^*}}\right)$ [Allenberg et al., 2006]
- $R_T = \tilde{O}\left(\sqrt{N \sum_{t=1}^{T} \|\ell_t - \bar{\ell}\|_\infty^2}\right)$ [Wei and Luo, 2018]
- $R_T = \tilde{O}\left(\sqrt{N \sum_{t=1}^{T-1} \|\ell_t - \ell_{t+1}\|_\infty}\right)$ [Bubeck et al., 2019]



Best-of-both-worlds アルゴリズム：確率的／敵対的のどちらのモデルでもうまく働く [Bubeck and Slivkins, 2012]
- Tsallis-INF アルゴリズム: 確率的モデルに対し $R_T = O\left(\frac{N \log T}{\Delta}\right)$, 敵対的モデルに対し $R_T = O\left(\sqrt{NT}\right)$ [Zimmert and Seldin, 2021]
- **[本研究]**: 確率的モデルに対し $R_T = O\left(\frac{N \log T}{\Delta}\right)$, [Ito, 2021a]
敵対的モデルに対し $R_T = \tilde{O}\left(\sqrt{N \min\left\{T, \sum_{t=1}^{T} \ell_{t i^*}, \sum_{t=1}^{T} \|\ell_t - \bar{\ell}\|_\infty^2, \sum_{t=1}^{T-1} \|\ell_t - \ell_{t+1}\|_1\right\}}\right)$

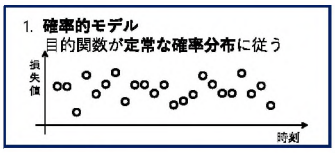
- 同様の成果を，組合せセミバンディット問題（経路最適化，推薦最適化等）や劣モジュラ最小化（価格組合せ最適化等）にも拡張

成果２ 環境変動にロバストな手法と解析

外れ値データ，敵対的ノイズに対しロバストなアルゴリズムを提案
達成可能なロバスト性の限界を解析　NeurIPS2021 などで発表
[Ito, 2021c]

２つの環境モデルの中間: 敵対的改変つき確率的モデル

1. **確率的モデル**
目的関数が**定常な確率分布**に従う

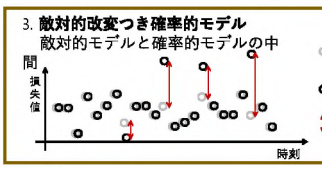


…

2. **敵対的モデル**
目的関数が**任意に（敵対的に）変化**

$C = 0$

3. **敵対的改変つき確率的モデル**
敵対的モデルと確率的モデルの中間



$C = O(T)$

- 改変の大きさの度合い： $C := \sum_{t=1}^{T} \|\ell_t - \ell'_t\|_\infty$ と定義
- C の大きさに応じて性能はどう変化するか？：敵対的学習，外れ値データへの頑健性などの問題意識と関連

（完全情報）オンライン意思決定問題における貢献

定理: 上記設定で，あるアルゴリズムが $R_T = O\left(\min\left\{\frac{\log N}{\Delta} + \sqrt{\frac{C \log N}{\Delta}}, \sqrt{T \log N}, T\right\}\right)$ を達成

定理: 任意の N, T, $\Delta_{\min} \in (0,1]$, $C \in [0,T]$ と任意のアルゴリズムに対し，上記設定において
最悪時には $R_T \geq \Omega\left(\min\left\{\frac{\log N}{\Delta} + \sqrt{\frac{C \log N}{\Delta}}, \sqrt{T \log N}, T\right\}\right)$ が成立

- 上限と下限が一致 ⇒ 定理の理論保証はこれ以上改善できない．
- $O(C)$ の敵対的改変に対する最適な頑健性は $O\left(\sqrt{\frac{C \log N}{\Delta_{\min}}}\right)$ で特徴づけられる．
- 同様の成果を多腕バンディット，組合せセミバンディットなどにも拡張

参考文献

- Chamy Allenberg, Peter Auer, László Györfi, and György Ottucsák. Hannan consistency in on-line learning in case of unbounded losses under partial monitoring. In International Conference on Algorithmic Learning Theory, pages 229–243. Springer, 2006.
- Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In Conference on Learning Theory, pages 217–226, 2009.
- Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. Machine learning, 47(2-3):235–256, 2002a.
- Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. SIAM Journal on Computing, 32(1):48–77, 2002b.
- Sébastien Bubeck and Aleksandrs Slivkins. The best of both worlds: Stochastic and adversarial bandits. In Conference on Learning Theory, pages 42–1, 2012.
- Sébastien Bubeck, Yuanzhi Li, Haipeng Luo, and Chen-Yu Wei. Improved path-length regret bounds for bandits. In Conference on Learning Theory, pages 508–528, 2019.
- Mark Herbster and Manfred K Warmuth. Tracking the best linear predictor. Journal of Machine Learning Research, 1(281-309):10–1162, 2001.
- Shinji Ito, Shuichi Hirahara, Tasuku Soma, and Yuichi Yoshida. Tight first-and second-order regret bounds for adversarial linear bandits. Advances in Neural Information Processing Systems, pages 2028–2038, 2020.
- Shinji Ito. Parameter-free multi-armed bandit algorithms with hybrid data-dependent regret bounds. In Conference on Learning Theory, pages 2552–2583, 2021a.
- Shinji Ito. Hybrid regret bounds for combinatorial semi-bandits and adversarial linear bandits. Advances in Neural Information Processing Systems, 34, 2021b.
- Shinji Ito. On optimal robustness to adversarial corruption in online decision problems. Advances in Neural Information Processing Systems, 34, 2021c.
- Chen-Yu Wei and Haipeng Luo. More adaptive algorithms for adversarial bandits. In Conference on Learning Theory, pages 1263–1291, 2018.
- Julian Zimmert and Yevgeny Seldin. Tsallis-inf: An optimal algorithm for stochastic and adversarial bandits. Journal of Machine Learning Research, 22(28):1–49, 2021.