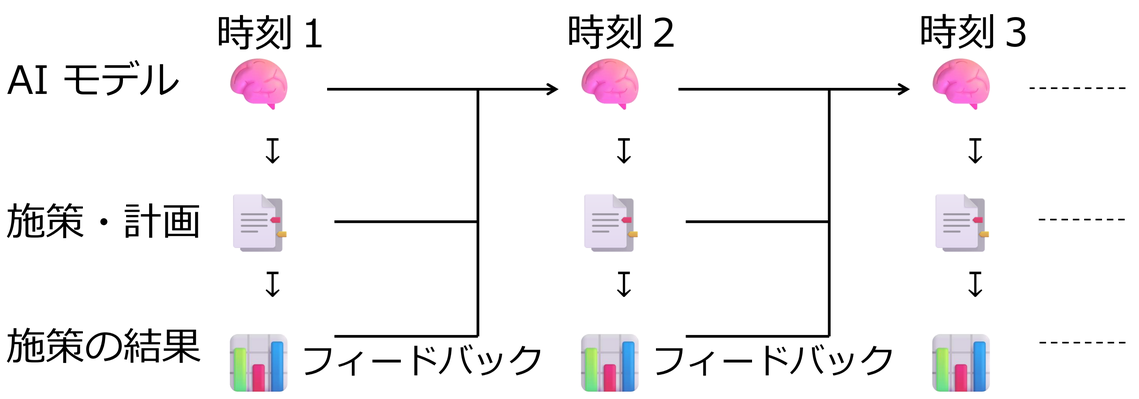
FY2024/2024年度

Sequential Decision Making Team
Shinji Ito
逐次的意思決定チーム　伊藤 伸志

チームリーダー：伊藤 伸志
客員研究員：本多 淳也，土屋 平，
小宮山 純平，筒井 和詩，
坂上 晋作，相馬 輔

活動内容：不確実性や環境変化の中で逐次的に合理的な判断を下すための方法論・理論の開発
FY24は，主にオンライン学習の理論的成果を創出

オンライン学習・オンライン最適化による意思決定プロセス

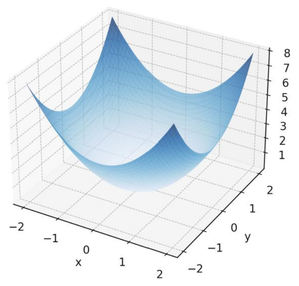


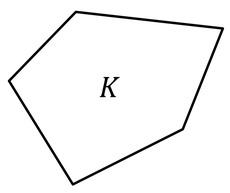
フィードバック情報をもとにモデルの更新を常に繰り返す

利点：
☺過去データが限られている状況でも導入可能
☺環境の変化に柔軟に適応・追従が可能

オンライン凸最適化における高速な収束レートの理論：[Tsuchiya and Ito, NeurIPS'24]

目的：オンライン凸最適化において，どのような条件で収束レートが改善するかを解明
既存研究：目的関数が強凸なケースや，実行可能領域が多面体的なケースで収束レートが改善

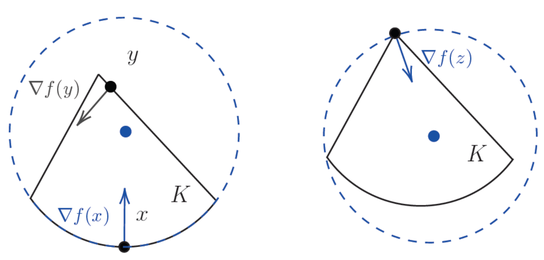




貢献：目的関数が強凸でなくとも，実行可能領域の（収束先における）局所的な曲がり方に応じて収束レートが改善することを証明

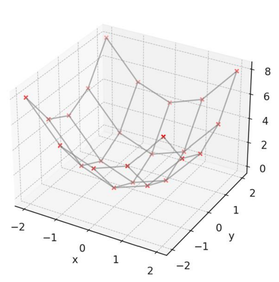
1. $u \in \mathrm{bd}(\mathbb{B}(c,\rho))$
2. $K \subseteq \mathbb{B}(c,\rho)$
3. there exists $k>0$ such that $u + k\nabla f(u) = c$

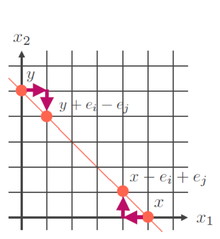
$$\mathrm{R}_T = O\left(\frac{G^2\rho}{\|\nabla f^\circ(x_\star)\|_2}\log T\right)$$

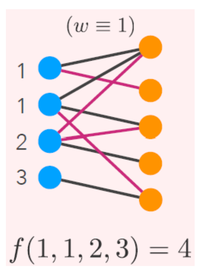


離散凸解析とオンライン学習の融合 [Oki and Sakaue, NeurIPS'24]

目的：現実の計画・意思決定問題（スケジューリング・マッチング・経路計画 etc.）に多く現れる**離散変数**のモデルを扱えるオンライン学習の枠組みを確立
アプローチ：**離散凸解析** (cf. [Murota, 2013]) とリグレット解析の枠組みを融合







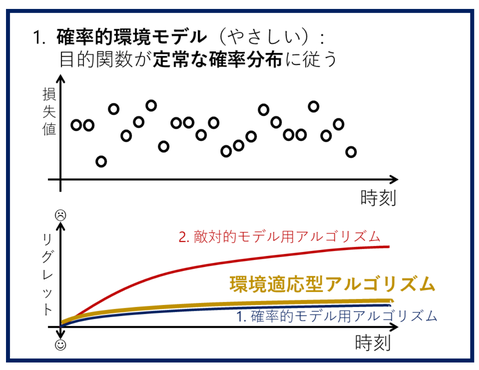
貢献：オンラインM♮凸最適化の枠組みを導入し，
- 確率的（定常的）環境で有効なアルゴリズムを提案

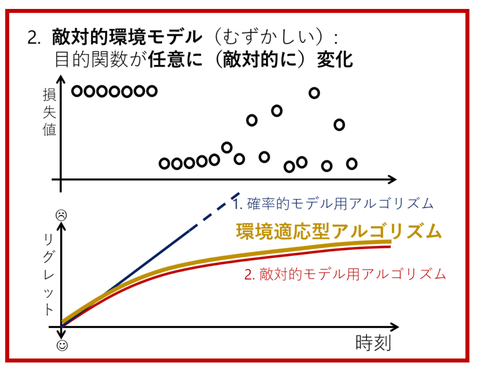
$$\mathrm{sReg}_T = O(K\sqrt{KN/T}) \qquad \mathrm{Reg}_T = O(KN^{1/3}T^{2/3})$$

- 敵対的（非定常的）環境におけるリグレット最小化は計算量的に困難であることを証明

多様な環境に適応するアルゴリズムの理論
[Tsuchiya, Ito and Honda, ICML'24][Ito, Tsuchiya and Honda, COLT'24] [Tsuchiya and Ito, NeurIPS'24]

目的：環境の性質に応じて性能が改善する**両環境最適**アルゴリズムを設計する汎用的原理の構築





アプローチ：FTRL (follow-the-regularized-leader) の枠組みにおいて，新たな**適応的学習率**調整の仕組みを開発

FTRL: $p_t \in \arg\min_{p\in\Delta^N}\left\{\left\langle\sum_{s=1}^{t-1}\hat{\ell}_s, p\right\rangle + \frac{1}{\eta_t}\psi(p)\right\}$ $\Longrightarrow$ $R_T \le \sum_{t=1}^{T}\eta_t z_t + \sum_{t=1}^{T}\left(\frac{1}{\eta_t}-\frac{1}{\eta_{t-1}}\right)h_t$

$\hat{\ell}_t$: 損失 ℓ_t の不偏推定量，$\psi(p)$: 正則化関数，η_t: 学習率

stability 項　　penalty 項

提案手法： $\eta_{t-1}z_{t-1} = \left(\frac{1}{\eta_t}-\frac{1}{\eta_{t-1}}\right)h_t$ となるよう学習率 η_t を更新

貢献：FTRLの適応的学習率の更新規則を提案し，
- 競合比解析の枠組みにおいて学習率の最適性を証明
- 複数の異なる問題設定で両環境最適性を達成

アルゴリズム: Tsallis エントロピーを用いた FTRL　$p_t \in \arg\min_{p\in\Delta(K)}\left\{\sum_{s=1}^{t-1}\hat{\ell}_s^T p - \frac{1}{\eta_t}\sum_{i=1}^{K}p_i^\alpha - \frac{1}{\bar{\eta}}\sum_{t=1}^{T}p_i^{1-\alpha}\right\}$

Table: Examples of BOBW regret bounds

	Parameters	α	Stochastic	Adversarial
Multi-armed bandit	K: #arms	½	$\sum_{i\neq i^*}\frac{1}{\Delta_i}\log T$	$\sqrt{KT}$
Multi-armed bandit	K: #arms	$(0,1)$	$\frac{K\log T}{\alpha(1-\alpha)\Delta_{\min}}$	$\sqrt{\frac{KT}{\alpha(1-\alpha)}}$
Graph bandit (strongly observable)	K: #arms, ζ: ind. number	$1-\frac{1}{2\log(Ke/\zeta)}$	$\zeta\log\left(\frac{Ke}{\zeta}\right)\frac{\log T}{\Delta_{\min}}$	$\sqrt{\zeta\log\left(\frac{Ke}{\zeta}\right)T}$
Linear bandit	K: #arms, d: dimensinoality	$1-\frac{1}{2\log(K/d)}$	$d\log(K)\frac{\log T}{\Delta_{\min}}$	$\sqrt{d\log(K)\,T}$
Contextual bandit	M: #arms, K: #experts	$1-\frac{1}{2\log(K/M)}$	$M\log(K)\frac{\log T}{\Delta_{\min}}$	$\sqrt{M\log(K)\,T}$

ミニマックスリグレットの解析 [Ito, NeurIPS'24]

目的：エキスパート選択バンディット問題の困難性の解析

推薦システムにおけるモデル選択への応用例：

For $t = 1,2,\dots,T$:
- ユーザ t (👵)がサービスにアクセス
- 各モデル $j \in [N]$ (🤖🧠)の推薦 $e_t(j) \in [K]$ を確認
- ユーザ t への推薦コンテンツ $i_t \in [K]$ (🎞️)を決定
- ユーザ t からのフィードバック$r_t(i_t)$を確認

💻←👵
🤖: 🎞️　　🧠:🎧
💻: 🎞️ → 👵
👵👍 or 👵👎

貢献：上界に一致する下界を証明

問題設定	リグレット上界（アルゴリズムの性能）	リグレット下界（達成可能限界）
エキスパート選択多腕バンディット	$O\left(\sqrt{KT\log N}\right)$ [Auer+, 2002]	$\Omega\left(\sqrt{KT}\right)$ [Auer+, 2002]
	$O\left(\sqrt{KT\log\frac{N}{K}}\right)$ [Kale, 2014]	$\Omega\left(\sqrt{KT\frac{\log N}{\log K}}\right)$ [SL2016]
		$\Omega\left(\sqrt{KT\log\frac{N}{K}}\right)$ **[I2024]**

- T: ラウンド数
- K: アクション数
- N: エキスパート数

文献情報 (selected publication)
- Oki and Sakaue. No-Regret M♮-Concave Function Maximization: Stochastic Bandit Algorithms and NP-Hardness of Adversarial Full-Information Setting, NeurIPS'24.
- Tsuchiya and Ito. Fast Rates in Online Convex Optimization by Exploiting the Curvature of Feasible Sets, NeurIPS'24.
- Ito, Tsuchiya and Honda. Adaptive Learning Rate for Follow-the-Regularized-Leader: Competitive Ratio Analysis and Best-of-Both- Worlds, COLT'24.
- Tsuchiya, Ito and Honda. Exploration by Optimization with Hybrid Regularizers: Logarithmic Regret with Adversarial Robustness in Partial Monitoring, ICML'24.
- Tsuchiya and Ito. A Simple and Adaptive Learning Rate for FTRL in Online Learning with Minimax Regret of $\Theta(T^{2/3})$ and its Application to Best-of-Both-Worlds, NeurIPS'24.
- Ito. On the minimax regret for contextual linear bandits and multi-armed bandits with expert advice, NeurIPS'24.