The 36th Annual Conference of the Japanese Society for Artificial Intelligence, 2022

強化学習が可能にするオペレーション最適化の世界

Frontiers of operation optimization enabled by reinforcement learning

窪澤 駿平 *1*2 Shumpei Kubosawa
大西 貴士 *1*2 Takashi Onishi
鶴岡 慶雅 *1*3 Yoshimasa Tsuruoka

*1産業技術総合研究所 NEC-産総研人工知能連携研究室
NEC-AIST AI Cooperative Research Laboratory, National Institute of Advanced Industrial Science and Technology

*2NEC データサイエンス研究所
Data Science Research Laboratories, NEC Corporation

*3東京大学 大学院情報理工学系研究科
Gradiate School of Information Science and Technology, The University of Tokyo

The automation and optimization of planning tasks, such as resource allocation and operational planning of various systems such as transportation systems and production facilities have been addressed mainly in operations research and each specific field. Conventionally, planning problems i.e. scheduling problems are reduced to combinatorial optimization problems and addressed using their solvers. In such cases, scheduling complex systems for a long period might incur combinatorial explosions and would be difficult to obtain the solution. Several scheduling problems can also be regarded as optimal control problems. Optimal control problems include several problems concerning sequential decision making such as board games. Reinforcement learning is a method to address them, and its recent advancement is significant. If the complex scheduling problems are reduced to optimal control problems i.e. deciding resource allocation at each time step, not as a whole, recent powerful reinforcement learning can be leveraged to obtain solutions in a short period after the training. In this paper, we introduce these perspectives and their practical applications including railway scheduling and chemical plant operation.

1. はじめに

AI が囲碁のチャンピオンに勝利するのは 2030 年代になるだろうとの予測があった中 [松原 03]，深層強化学習を組み込んだ AI 囲碁プレーヤーの AlphaGo[Silver 16] がプロの囲碁棋士を破ったのは 2015 年のことだった。もともと強化学習は，囲碁の様に操作や状態（局面）が離散的なシステムよりも，倒立振子やロボットなど，操作量が連続値である機械制御への適用事例が多かったが，AlphaGo と時をほぼ同じくして，連続制御に適した深層強化学習手法の TRPO[Schulman 15] も提案された。その後も手法の改良が続き，応用事例も少しずつ増えてきた。つい先日にも，「トカマク型核融合炉」という複雑な装置の自動制御に成功したとの報告があった [Degrave 22]。

この様に，さまざまなタスクで高い性能を示してきた強化学習だが，ゲームや特殊な実験装置だけでなく，もう少し人々の日々の暮らしの近くで利用できないのだろうか。そもそも強化学習とは，どの様な問題を，どの様に解く方法なのだろうか。そこで本稿では，こうした観点に立ち戻り，強化学習で解決可能な問題を概観し，強化学習を日常の業務に適用する方法論と，筆者らの応用事例について述べる。

2. 関連研究

2.1 最適制御問題と解法の例

囲碁も倒立振子も，「直前の入力値だけでなく，入力値の過去の時系列によって出力が決まるシステム」すなわち時間の概念を含むダイナミカル（時間で変化する）システムである。囲碁の場合，現時刻に自分が石を打つと，直前の局面（状態）と，囲碁のルールと相手の手により次の時刻の局面（状態）が決まる。直前の状態が異なれば，同じ入力でも次の状態は異なる。倒立振子の場合も，各時刻の入力（例：トルク）を受けて内部状態が時間変化し，内部状態によって出力値（例：振り子の先端位置）が決まる。一般に，離散時間ダイナミカルシステムは

$$f : \mathcal{S} \times \mathcal{A} \longrightarrow \mathcal{S} \tag{1}$$

ただし $\mathcal{S}$ は状態空間，$\mathcal{A}$ は行動（入力）空間として定義できる。時刻 t における状態 $s_t \in \mathcal{S}$ と入力 $a_t \in \mathcal{A}$ について，次の時刻の状態は $s_{t+1} = f(s_t, a_t)$ である。システムの出力として，状態の一部または全部が観測できる。

ダイナミカルシステムを対象に，現在の状態 s_0（例：最初の盤面）から時々刻々と操作してゆき，目標状態（例：勝利）を実現する方法のことを最適制御と呼ぶ。最適制御とは，目的関数を満たすため，次の時刻の最適操作を選択する最適化問題

$$\hat{a}_1 = \mathrm{argmin}_{a_1} \sum_{t=0}^{\infty} \ell(s_t, a_t) \text{ s.t. } s_{t+1} = f(s_t, a_t) \tag{2}$$

である。ただし，$\ell : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ は状態と行動の望ましさを表す評価関数であり，小さな値ほど望ましい。

2.1.1 モデル予測制御

最適制御の方法のひとつに，式 (2) の総和の終端時刻を $\tau \in \mathbb{N}$ として，有限期間の最適解を毎時刻求め続けるモデル予測制御 (model predictive control; MPC) がある。MPC では，「現在の状態から，どの操作をすると，次はどの状態になるか」をモデル f を使用して予測し，評価関数で操作の良しあしを評価して最適操作を選択する方法であり，ゲーム AI なら τ 手先まで先読みして最善手を打つ方法とみなせる。MPC は化学プラントの運転制御などの導入事例があるが，次の操作時刻までに最適化計算を終えるため，予測モデルを線形に限定したり，終端時刻を τ で打ち切るなど，最適化問題を縮小することで実用化されてきた。なお，MPC も，最初の提案から 40 年以上に亘り研究されており，計算性能も向上した現在では，非線形 MPC や確率的 MPC なども盛んに研究されている。

連絡先: 窪澤駿平, 産業技術総合研究所 NEC-産総研人工知能連携研究室, kubosawa@nec.com

2.1.2 強化学習

強化学習は，データを利用して統計的に最適制御問題を解く方法のひとつである。強化学習では，現在の状態に応じて最適操作を出力する方策分布 $P_\pi(a_t|s_t)$ を，最適化問題

$$\hat{P}_\pi = \mathrm{argmax}_{P_\pi} \mathbb{E}_{a_t \sim P_\pi(a_t|s_t)} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \right] \quad \text{subject to } s_{t+1} \sim P_f(s_{t+1}|s_t, a_t) \tag{3}$$

を事前に解いて構築する。ここで，$\gamma \in (0, 1)$ は総和の発散を防ぐための割引率であり，$r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ は状態と行動の望ましさを表す報酬関数（大きな値ほど望ましい），$P_f(s_{t+1}|s_t, a_t)$ は現在の状態と行動から次の状態を確率的に出力するシステムである。MPC と異なり，強化学習では最適化する終端時刻は明示的に決める必要が無く，遠い未来までの合計報酬を最大化する。これを実現するため，強化学習では，現在の状態 s_k と各時刻の報酬に基づき，将来の全期間で得られると期待される報酬を表す価値関数

$$V_\pi(s_k) = \mathbb{E}_{a_{k:\infty} \sim P_\pi(a_t|s_t)} \left[\sum_{t=0}^{\infty} \gamma^t r(s_{k+t}, a_{k+t}) \right] \tag{4}$$

を定義し，各時刻のデータ (s_t, a_t, s_{t+1}) から価値関数を推定し，推定した価値関数を利用して方策分布を推定する。

強化学習では，事前にモデルを使用して方策分布を求めておくため，利用時の計算量が少なく済むメリットがある。また，データに基づく手法のため，モデルの入出力にさえアクセスできれば学習できる。このため，囲碁から制御システムまで，幅広いモデル，すなわちダイナミカルシステムに適用できる。

強化学習では，最適化途中の方策分布でシステムを操作して，学習データを自動で収集するのが一般的だが，あらかじめ収集しておいたデータを与えるだけのオフライン強化学習も可能である。データの収集方法よりも，強化学習の本質は，動的計画法の一種として，状態と行動および報酬からなる時系列データから，解いている制御問題（タスク）やシステムの特徴（コスト）を表す最適な価値関数や方策関数を近似してゆくこと [Bertsekas 95]，すなわち，状態同士の関係を整理し，「タスクについて類似した状態の価値関数値」という部分問題の解を，ひたすら記憶し，改善してゆくことにある。さらに深層強化学習では，価値関数や方策分布のメモリとして，大容量の深層学習を導入したことで，大量の場合の数が生じる囲碁や複雑な制御システムにおいて高性能を実現した。なお，冒頭の核融合炉の事例では，価値関数には大きなネットワークを使用し，方策分布には小さなネットワークを使用することで，利用時の計算量が抑制されている [Degrave 22]。

さて，この様に捉えてみると，深層強化学習は，次の様な特徴を持つ問題に適していると考えられる。すなわち (1) 時間軸を扱う問題（特に，長い時間を扱う問題），(2) 目標状態が定量的に記述できる問題，(3) システムの入出力しかアクセスできない問題，(4) 取れる行動（選択肢）や状態が沢山ある問題，(5) 状態空間や行動空間に一定の構造がある（表現学習が可能な）問題，そして (6) 大量のデータに高速にアクセスできる（十分な量の良質なデータベースがある，またはシミュレータで高速にデータを生成できる）問題である。

2.2 スケジューリング問題と典型的な解法

最適制御問題とは別に，時間を順序として扱う問題として，スケジューリング問題がある。スケジューリング問題とは，与えられた目的関数と期間について，資源を最適に割り当てる問題であり，製造業やサービス業の多くで取り組まれている [Pinedo 12]。例えば，作業員の勤務表作成や配送計画の立案などがある。また，巡回セールスマン問題として知られる，すべての都市を 1 回だけ訪問しつつ最短の経路を求める問題もスケジューリング問題に含まれる。スケジューリング問題は，線形計画法，動的計画法，および制約プログラミングなどの組み合わせ最適化問題に還元して解かれることが多い [Pinedo 12]。

2.2.1 混合整数計画法

混合整数計画法（mixed linear programming; MILP）は，線形計画法における変数の一部に整数である制約を加えたものである。例えば作業員が作業 $w \in W$ を時刻 $t \in T$ に割り当てられた場合は $x_{w,t} = 1$ そうでない場合は $x_{w,t} = 0$ として状態（決定変数）を表し，作業 w を時刻 t に行ったときのコストを $c_{w,t} \in \mathbb{R}$ で定義し，さらに守らなければならない作業の順序などを不等式で定義し，目的関数を minimize $\sum_{w,t \in W \times T} c_{w,t} x_{w,t}$ と定義してソルバに入力すると，コストを最小化する作業スケジュールとして $x_{w,t}$ からなる行列が得られる。MILP では，スケジューリング対象の問題（システム）を線形の不等式で表現し，目的関数も線形で表現する必要がある。

2.2.2 動的計画法

動的計画法 (dynamic programming; DP) は，解きたい主問題を粒度の小さい副問題に分割し，主問題の解が得られるまで副問題を解き続ける方法である。副問題を解くときに，既に解いた副問題の解を再帰的に利用することで効率的に解ける特徴がある。DP は状態の評価結果（目的関数値）を列挙してゆく方法のため，対象のシステムや目的関数は非線形でも構わない。なお，強化学習も，方策分布を求める主問題を解く上で，価値関数を求めるという副問題を解いている。加えて，ニューラルネットワークに代表される関数近似法を利用するため，強化学習はニューロ動的計画法とも呼ばれてきた [Bertsekas 95]。

2.2.3 制約プログラミング

制約充足問題を含む最適化問題を解く方法に，制約プログラミングがある。制約プログラミングでは，例えば「工程 A には 1 時間かかる」という状況を ‘工程 A(1)’ と表したり，順序の制約「工程 B は工程 A の後でなければならない」を ‘工程 B.startsAfterEnd(工程 A)’ など，述語論理の様な形式で表す。制約プログラミングでは，システムを制約の集合により表現し，さらに所要時間の最小化などの目的関数を設定してソルバで解く。線形計画法と比較して，制約プログラミングは最適性よりも解の実行可能性を重視した方法である [Pinedo 12]。

2.3 オペレーションズ・リサーチ

スケジューリング問題を含む，日常の様々な業務，すなわちオペレーションにおける問題を，科学的に解決する方法を研究する分野にオペレーションズ・リサーチ（operations research; OR）がある。OR では，これまで述べてきた方法などにより，幅広い分野の問題解決に取り組まれてきた。日本オペレーションズ・リサーチ学会が運営する OR Wiki[OR 学会 12] の事例集には，2 月 22 日時点で 939 本（重複無し）の論文が登録されている。そこから応用に関する単語を含む論文本数を集計し，「単語 (論文本数)」と表記していくつか示すと，生産 (36)，交通 (29)，金融 (18)，百貨店，道路，電力 (ここまで 17)，医療 (14)，ロジスティクス，需要，購買 (13)，輸送 (12)，電車 (11)，POS，教育，SCM[*1]，観光 (10)，サービス，小売，投資，石油 (9)，工場，マーケティング，ナース，配送，旅行，取引 (8)，野球，在庫，地震 (7) ，試合，流通，化学，供給，店

*1 サプライ・チェーン・マネジメント：原材料の調達から製品やサービスを顧客に届けるまで，全ての業務フローを管理・改善すること

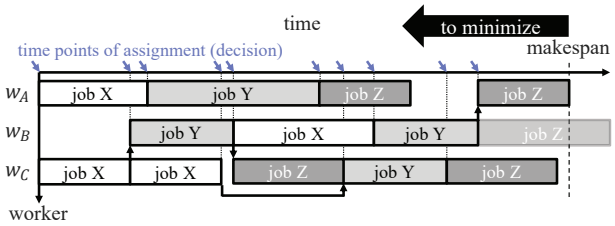

図 1: ジョブショップスケジューリング問題の例

舗，携帯，物流，販売，組立，チェーン (6)，ガス，年金，電話，動産，広告，救急，建築，郵便，鉄鋼 (5) などと続く。製造業，運輸業，金融業，小売業，インフラ事業などの事例が多くみられる。技術に関する単語は，モデル (125)，最適 (82)，計画 (74)，データ (69)，ネットワーク (53)，スケジューリング (43)，予測 (37)，シミュレーション (33)，マイニング (14)，ニューラル (12)，微分 (9)，マルコフ，ファジィ (8)，整数，指数，線形 (7)，モデリング，データベース (6)，GA，カオス，動的，プログラミング，クラスター (5) などと続いており，スケジューリング問題としての定式化が多く，シミュレーションの事例も豊富にあることがわかる。

3. 強化学習によるオペレーション最適化

OR の一角を占めるスケジューリング問題は，資源（例：作業員）を，各時刻において，配分したい対象（例：作業 X, Y, Z）に割り当てる問題である。つまり，時間軸と対象（作業）の軸それぞれの有限の区間から構成される領域に，資源をどういう組み合わせで詰め込むか，という問題である。作業員を作業（ジョブ）に割り当てるジョブショップスケジューリング問題の例を図 1 に示す。このとき，作業員 × 時間の領域が小さければ，組み合わせの場合の数も限られるため解きやすいが，いずれかの軸の区間が伸びるほど，場合の数が指数関数的に増加し，解くのが困難になることがある。例えば，作業時間が多様であり，作業や資源間の関係も複雑な場合に，長時間に亘るスケジュールを組もうとすると，場合の数が爆発的に増えてしまう。この例として，OR にて取り組まれてきた鉄道ダイヤの事例が挙げられる。鉄道ダイヤの自動作成において，近年では MILP を使用した事例が多いが，この場合，10 駅程度かつ 10 列車程度の規模であれば 10 分もあれば計算できるが，それ以上の規模になると実用的な計算時間では解けない場合が多い。しかし，大手私鉄の主要路線では少なくとも 20 駅程度は存在するため，全区間かつ 1 日分のダイヤを MILP で作成するのは困難であり，ダイヤ自動作成の実用化は進んでいない [窪澤 21]。同様の状況は他の分野でも多く存在すると考えられ，実態としては，本来解きたい問題を縮小・簡略化して自動化している，あるいは自動化されていないと考えられる。

一方で,「いつ，どこに，どれを割り当てるか」というスケジューリング問題を，時間軸で分割して「いつ」を省き,「どこに，どれを割り当てるか」を，毎時刻，その時の状態に応じて最適に判断する問題とすると，スケジューリング問題は最適制御問題に置き換わる。例えば図 1 の問題は，各時刻において，作業員の最適な割り当てを決定することで，合計作業時間を最小化する最適制御問題に置き換えられる。この解法に強化学習を利用できる。強化学習を利用した AlphaGo が囲碁で勝利できたのは，長い時間を隔てた状態と行動の関係を価値関数として学習できたためであり，長期間のスケジューリングが強化学習で可能であると示した例である。また，強化学習では，利用時の各時刻の計算量は均等であるため，問題の時間軸を伸ばしても，計算量は時間について線形にしか増えない。つまり，スケジューリング問題を最適制御問題とみなせば，従来は実用的に解けなかったような規模のスケジューリング問題についても，強化学習を利用することで，最適とは限らないが，現実的に解を得られる可能性が出てきている (§4.1)。

実際，従来 OR で取り組まれてきたような問題に強化学習で取り組む試みが広がってきている。例えば，強化学習で一般的な OpenAI Gym の API に準拠したシミュレーション環境として，OR の分野での利用が多いナップサック問題やサプライチェーンマネジメントにおける在庫管理問題，さらに巡回セールスマン問題などが実装された OR-Gym[Hubbs 20] が開発されている。また，Gym API ではないものの，現実的な資源配分問題として，海運網におけるコンテナ在庫管理問題や，自転車シェアサービスにおける自転車再配置問題，およびデータセンターにおける仮想マシンスケジューリング問題が実装された強化学習環境が提案されている [Jiang 20]。他にも，自動プランニングとスケジューリングに関する国際会議 ICAPS 2021 では，スマートグリッドの配電問題や，列車スケジューリング問題，ジョブショップスケジューリング問題，動的配送問題を実装した Gym 形式の環境が用意され，それらの問題を強化学習で解くコンペティション [ICAPS 21] が開催された。さらに，OR で多用される組み合わせ最適化問題を，強化学習で解く方法をレビューした論文もある [Mazyavkina 21]。OR に関する主要ジャーナル *Operations Research* 誌においても，'reinforcement learning' を含む論文数は，2018 年には 2 本，2019 年 3 本，2020 年 6 本，2021 年 13 本，2022 年も 3 月 1 日現在で既に 9 本掲載されるなど，急速に増加している。このように，強化学習は制御問題の解法としてだけでなく，OR の分野での利用と実用化も広がってゆくと考えられる。

3.1 Sim-to-Real ギャップ

強化学習の課題のひとつに，Sim-to-Real ギャップ (simulation to reality gap) がある。シミュレータで学習した方策分布を用いて，実世界のシステムを操作させたとき，実世界のシステムとシミュレータとの応答（挙動）の違いにより，シミュレーション通りの性能が出ない問題である。強化学習というより予測モデルの問題ともみなせるが，強化学習にはモデルが必要である限り見過ごせない課題である。筆者らは，この問題にも強化学習を利用するアプローチを提案し，化学プラントの運転に成功している (§4.2)。スケジューリング問題の場合でも同様に，システムを詳細に表現し，複雑な問題を解こうとするほど，モデルの再現性向上が重大な課題になってゆく。

3.2 オペレーションズ・モデリング

強化学習には，シミュレータ（予測モデル）か，モデルの代わりに大量かつ良質なデータベースが必要である。過去の上手いオペレーションの再現が重要であれば，データベースを利用したオフライン強化学習が可能である。また，昨今のパンデミックにおける社会情勢の様に，未経験の状況にも上手く対応したい場合には，現象を表す物理モデルや業務モデルを組み込むことで，再現性（外挿性）の高いダイナミックシミュレータを構築し，これを利用したオンライン強化学習が有効である。

近年，OR に関する強化学習用シミュレータの開発が相次いでいる様に，シミュレータを作る作業，すなわちオペレーションのモデリングが重要度を増している。もともと MILP や制約プログラミングで解かれてきた様な問題であれば，そのモデルをダイナミックシミュレータに簡単に変換する方法が望まれる。しかし新たに取り組む問題の場合には，新たにモデルを作

る必要がある。モデリングコストを下げつつ，未経験の状況も再現できるモデルを作るためには，データに基いており，かつ外挿性も向上させやすい手法が必要である。例えば，制約や予測式としてモデルの挙動についての知識を組み込むことができ，さらにデータを併用するシンボル回帰などが望まれる。

4. アプリケーション事例

本節では,「強化学習によるオペレーション最適化」として筆者らが取り組んできた応用事例について紹介する。

4.1 鉄道運行計画

大都市圏の鉄道路線では，多くの列車が過密なダイヤで運行されている。このため事故や自然災害等による遅延や運休などの輸送障害が発生すると，その影響は少数の列車にとどまらず，多くの列車に連鎖的に波及し，広い範囲で時刻表ダイヤ通りの運行が不可能になる。この場合，ダイヤの修正や再作成を迅速に行う必要があるが，現状では熟練者が手作業で行っている。ダイヤ作成の自動化は長年研究されており，近年ではMILP による手法が多く提案されているが，計算時間等の課題により実用化は進んでいない。そこで筆者らは，鉄道路線を再現するシミュレータを開発し，強化学習によるダイヤ作成手法を開発している。首都圏の複線路線を想定した仮想路線で検証したところ，実用的な時間で全区間のダイヤが自動で作成可能と示されている [窪澤 21]。

4.2 化学プラントの運転

化学プラントとは，石油等の原材料から，化学反応や蒸留等の操作により，所望の化学製品を合成・分離する工場である。一定条件での生産中は多くの操作が自動化されているが，生産量や銘柄（生産する製品）の変更や外乱への対応など，非定常状態には熟練者が手動操作している。こうした操作のためには，(1) 現在のプラントの状態の正確な把握と，(2) 目標状態に移行するために最適な操作手順の生成，および (3) 外乱等により目標とのズレが生じた際の迅速な補正，という 3 要素が重要である。そこで筆者らは，これら 3 点のそれぞれに強化学習を適用することで Sim-to-Real ギャップを克服し，現実のプラントにおける運転支援 AI を実現した [Kubosawa 22]。

4.3 水力発電用ダムの操作

水力発電用のダムでは，平常時には川の水を貯水して発電に利用している。しかし，発電用ダムには治水能力がないため，大雨により洪水の危険性が高まると，事前に放流を行い，発電も停止する。また，放流を行う際にも，下流の安全を確保するため，少しずつ放流してゆく必要がある。つまり，放流には時間がかかる。このため，気象や河川の状況に合わせて，いつから放流を開始し，いつ発電を停止するかという判断が求められる。そこで筆者らは，河川シミュレータと強化学習によって，放流判断を最適化する取り組みを行っている [田中 18]。

4.4 航空管制

航空管制では，航空機同士の接近を避けるため，管制官が高度や速度などを各機に指示している。近年の航空需要の増大により，航空管制の負担も増加しており，接近回避指示の支援が求められている。状況が変化し続ける中，短時間で適切な指示を行うためには，強化学習による最適化が有効と見込まれる。

5. おわりに

本稿では，オペレーション最適化における強化学習のポテンシャルについて述べた。機械学習の普及と並行して急速に性能が向上している強化学習は，従来適用されてきたよりも遥かに広い分野での利用が可能である。筆者らは，こうした観点に立ち，さらに人間と AI が協調する方法論と技術を提案してゆくことで，強化学習をはじめとする AI の実用化を，各オペレーションを担う実務者と共に進めてゆきたい。

参考文献

[Bertsekas 95] Bertsekas, D. & Tsitsiklis, J. Neuro-dynamic programming: an overview. In *Proc. 1995 34th IEEE Conf. Decis. Control*. 1995, **1**, p.560–564.

[Degrave 22] Degrave, J., Felici, F., Buchli, J. et al. Magnetic control of tokamak plasmas through deep reinforcement learning. *Nature*. 2022, **602**, p.414–419.

[Hubbs 20] Hubbs, C. et al. OR-Gym: A reinforcement learning library for operations research problems. *arXiv preprint*. 2020, arXiv:2008.06319.

[ICAPS 21] The 31st International Conference on Automated Planning and Scheduling. “ICAPS Competitions”. https://icaps21.icaps-conference.org/Competitions/ (参照 2022-02-22)

[Jiang 20] Jiang. A. et al. “MARO: A Multi-Agent Resource Optimization Platform.” 2020, https://github.com/microsoft/maro (参照 2022-02-22)

[窪澤 21] 窪澤 駿平ほか. ダイナミックシミュレーションと強化学習による運転整理ダイヤ生成システム. 日本機械学会 第 *28* 回鉄道技術連合シンポジウム講演論文集 *(J-RAIL2021)*. 2021, p.SS5-2-1, (arXiv:2201.06276).

[Kubosawa 22] Kubosawa, S. et al. Sim-to-real transfer in reinforcement learning-based non-steady-state control for chemical plants. *SICE J. Control Meas. Syst. Integr*. 2022, **15**(1), p.10–23.

[松原 03] 松原 仁. ゲームのプログラムについて. 計測と制御. 2003, **42**(6), p.512–515

[Mazyavkina 21] Mazyavkina, N. et al. Reinforcement learning for combinatorial optimization: A survey. *Comput. Oper. Res.*. 2021, **134**, p.105400

[OR 学会 12] 日本オペレーションズ・リサーチ学会 OR 事典編集委員会. “OR Wiki”. https://orsj-ml.org/orwiki/wiki/ (参照 2022-02-22).

[Pinedo 12] Pinedo, M. *Scheduling: Theory, Algorithms, and Systems*. 5th ed., Springer, 2016, 670p.

[Schulman 15] Schulman, J. et al. Trust region policy optimization. In *Proc. 32nd Int. Conf. Mach. Learn. (ICML 15)*, 2015, p.1889–1897.

[Silver 16] Silver, D. et al. Mastering the game of Go with deep neural networks and tree search. *Nature*. 2016, **529**, p.484–489.

[田中 18] 田中 友紀子ほか. 電力ダム操作における強化学習型シンボルグラウンディングによる意思決定支援に関する検討. 人工知能学会全国大会論文集 *(JSAI2018)*. 2018, p.1M203.