論　文

分散強化学習による下水道送水系の制御

非会員　青木　　圭*　正　員　木村　　元*
正　員　長岩　明弘**　非会員　小林　重信*

Adaptive control of Sewerage Systems using Distributed Reinforcement Learning

Kei Aoki*, Non-member, Hajime Kimura*, Member, Akihiro Nagaiwa**, Member,
Shigenobu Kobayashi*, Non-member

In this paper, we propose a new simulation-based Distributed Reinforcement Learning approach that solves large planning problems under uncertain environment. The proposed method is a distributed state-action representation for softening an interaction and reward design for making agents cooperate. We apply it to real sewerage control systems, as the problem with uncertainty. Simulation results show it finds good control rules , which can cope with various situations, by dealing with the uncertainty included in real data directly on a simulator based on real systems.

キーワード：下水道送水系, 分散強化学習, プランニング, 制御規則の獲得
Keywords: Sewerage System,Distributed Reinforcement Learning,Planning,Making Control Rules

1. はじめに

実世界の問題を考える場合に不確実性のある環境をどのように扱うかという問題がしばしば生ずる．このような環境において適切な政策を獲得するために，マルコフ決定過程（MDP）によるモデル化とダイナミックプログラミング（DP）等の解法が有望であることが知られている．しかし，これらの手法は比較的大規模な問題では計算量的に実行不可能になる．そこで，対象問題を MDP に記述し，強化学習を用いて状態遷移などをサンプリングすることで DP の計算過程を確率近似する．これにより計算量を削減し，実用的な時間で適切な政策を獲得することができる．

多くの実問題においてオンラインの強化学習は試行錯誤の困難さや運用限界外の挙動の危険性等から困難であるため，実施設を基にシミュレータを作成し，その上でオフラインで強化学習を行うことで制御規則を獲得する．その際，実行不可能な規模の状態行動空間を問題の構造を利用して分割し，メモリ計算量や学習速度を向上させる．

*東京工業大学大学院総合理工学研究科
〒 226-8502　神奈川県横浜市緑区長津田町 4259
Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology
4259,Nagatsudacho,midori-ku,Yokohana,Kanagawa
**（株）東芝　電力・産業システム技術開発センター
〒 183-8511　東京都府中市東芝町 1
TOSHIBA CORPORATION 1,TOSHIBA-CHO,FUCHU-SHI,TOKYO

本論文では対象問題として下水道送水系の制御問題を扱い，従来法の問題点を回避して，実運用に十分適用できる有効な制御規則が獲得できることを示す．

2. 対象問題

下水道はライフラインのひとつとして現代生活において極めて重要な役割を担い，必要不可欠なものとなってきている．近年では下水道施設の普及やコスト削減の要請に伴い，効率的な施設運用・運用支援，運転の自動化などが要望されている．また，本論文で扱うシステムは電力・ガス・物流等と有用な実問題として幅広く，特に上水道送水系では浄水場からの水の流れを逆に考えれば同様の定式化とアプローチが適用できる．

〈2･1〉　下水道送水系の概要　系はひとつの処理場と複数のポンプ場からなり，それぞれが配管によって連結す

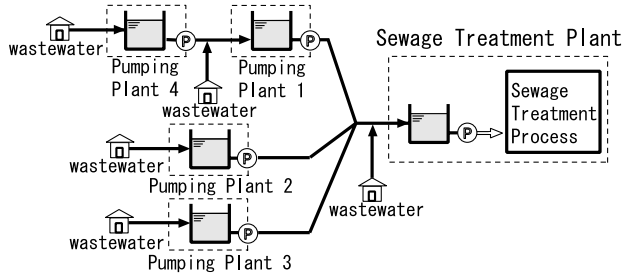

図 1　下水道送水系（5 施設）
Fig. 1.　5 plants Sewerage System.

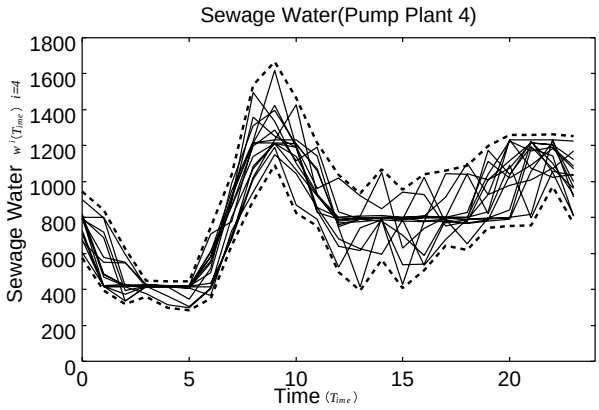

図 2　ポンプ場 4 の流入汚水量の時間系列データ（実線）．各時間の最大最小範囲（点線）
Fig. 2.　Sewage water data of plant 4.

る図 1 のような構造をしている．各施設は沈砂池やポンプ井といった貯水池を持ち，連接する施設に流入してきた汚水をポンプによって送水する．下水道には，家庭排水などの汚水のみが流れる分流式と汚水と雨水が流れる合流式があり，本論文では分流式，あるいは合流式の晴天時の汚水送水運用を考える．この際，以下の点に留意する．

（1） 汚水の集積
　各施設は担当地域の汚水を受け入れる．汚水は生活のリズムに連動して朝や夜に増え，深夜・早朝等は少なくなる時間的なパターン（図 2）を持つが，日々の変動は不確実性が高いものである．

（2） 貯水池の上下限制約
　各施設の貯水池は運用上下限制約があり，制約違反は汚水の流出や施設へのダメージにつながるため，その範囲内で運用されなければならない．

（3） 揚水量の調整
　各施設は設定されている複数のポンプの起動停止のみを考えた場合，揚水量を離散的に調整できる．

（4） 処理量の平滑化
　下水の処理過程では沈殿や生物化学処理を施す．これは通常長い時間を必要とし，薬剤や有機物の設定変更には時間とコストがかかる．そのため，設定の変更をなるべく少なくするために下水処理量を一定に保ちたい．つまり，処理場の揚水量（図 1 の末端の白矢印）を平滑化したいという要請がある．

（5） 運用コスト
　ポンプの消費電力と起動停止による消耗のコストがあり，消費電力は安価な夜間電力の利用で低減させうる．ただし，処理水質の保全面から考えるとコストの削減よりも処理量の平滑化の優先度が高い．

〈2･2〉 問題の所在　現状の運用では専門家が天候情報などから算出した予測値に基づいて，一日分の運用計画を作成し，それに従って当日の運用を行っている．運用計画は水位制約範囲内でできるだけ処理量が平滑化するように決定的に立案する．この際，以下のような問題が生じる．

（1） 予測の不確実性
　完全な予測は不可能であるため，実際には予測した状況から次第に誤差が生じる．この誤差は現場の運転員の経験や勘で補正操作され，運用計画によって期待される平滑化が十分には行われていない．

（2） トレードオフ
　現状の予測を前提とした計画手法では制約違反と平滑化性能のトレードオフを考慮することが難しい．

（3） 大規模な探索空間
　各施設の揚水量の組み合わせは施設数が増加するに従って指数的に増加し，プランニングが困難になる．

（4） 実時間性
　予測誤差によるずれを再計画で修正するために計算時間はなるべく短くなければならない．

〈2･3〉 本研究の立場　従来，上下水道の送水プランニングの研究は予測を前提に行われてきた．特に上水道に関しては，供給不足を避けるために GMDH [1] やニューロなどの手法を用いて高精度な需要予測が行われ，線形計画法 [2] や GA [3] で解く方法などが研究されている．下水道においては予測の精度が低いため，計画の補正をファジィ応用によって行う平滑化制御の研究 [4] などがあるが，これらの手法は上記に述べた問題点から必ずしも適切ではない．

そこで本研究では予測値に基づく従来法の問題点を回避するために，不確実性に柔軟に対処できる制御規則を実データから学習により獲得することを目指し，効率的な計画・運用法を提案する．この手法では学習に予測を用いないため，予測誤差や精度を考える必要がなく，一度，制御規則が獲得されれば状況に応じた運用計画は計算量なしで抜き出すことができる．また，運用支援や自動補正などにも獲得した制御規則を流用できるという利点がある．この際，施設の増加による計算量の爆発に対しては，次節の定式化で述べるように問題の特徴を利用して分割することで対処する．

獲得されるべき制御規則は，各時間ステップの各施設の揚水量を適切に決定し，貯水池水位の上下限制約を逸脱せずに処理場の揚水量をなるべく一定に保ち，急激な変化を避けるもの（平滑化）とする．また，本論文では平滑化に重点を置き電力コストについては扱わない．

3. 強化学習による解法の提案

〈3･1〉 対象のモデリング　下水道送水系の施設のモデルを図 3 に示す．N 施設からなる送水系の施設 $i(0,\ldots,\text{N-1})$ は，池底面積 $B^i[m^2]$，上限運用水位 $h^i_{MAX}[m]$，下限 $h^i_{MIN}[m]$ の貯水池と M 台のポンプ $P^i_m[m^3/h](\text{m}=1,\ldots,\text{M})$ を設備し，上流施設 i' と下流施設 i'' に接続している．

時刻 T_{ime} は現在時刻を表し，時間間隔 $Tw[h]$ で遷移する．ステップ t の始めに水位 $h^i(t)[m]$ と時刻 T_{ime} を観測し，揚水量 $u^i(t)[m^3/h]$ を決定する．揚水量 $u^i(t)$ はポンプ P^i_m の起動停止の組み合わせ $C^i_m(t) \in \{0,1\}$ によって

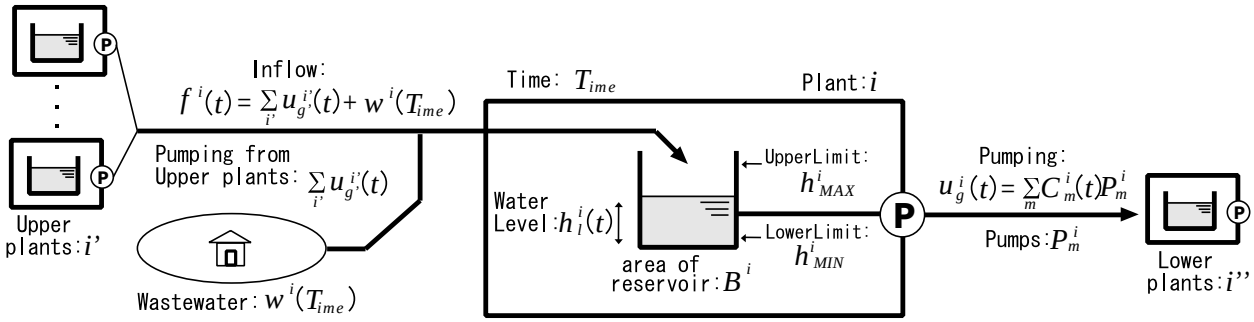

図 3　施設のモデル
Fig. 3.　Model of plants.

$$u_g^i(t)=\sum_m C_m^i(t)P_m^i Tw \cdots\cdots\cdots\cdots\cdots\cdots (1)$$

で求められ，貯水池から揚水される．ただし，同揚水量のポンプが複数あるため，$C_m^i(t)$ による組み合わせ数の全てではなく，g 段階で調節できる．この間に上流施設 i' からの合計揚水量 $\sum_{i'} u^{i'}(t)[m^3/h]$ と未知の確率関数で表現される汚水量 $w^i(T_{ime})[m^3/h]$ で決まる流入量

$$f^i(t)=\sum_{i'} u^{i'}(t)+w^i(T_{ime}) \cdots\cdots\cdots\cdots\cdots (2)$$

が流入する．以上より，次のステップ $t+1$ の貯水池水位は

$$h^i(t+1)=h^i(t)+\frac{(f^i(t)-u^i(t))}{B^i} \cdots\cdots\cdots (3)$$

で求められ, 下式の制約条件を満たさなければならない．

$$h^i_{MIN}<h^i(t+1)<h^i_{MAX} \cdots\cdots\cdots\cdots\cdots\cdots (4)$$

このような施設が木構造に近いネットワーク状に有向に連結し，下水道送水系を構成する．特にこのネットワークの終端（根）にあたる施設が処理場 (i=0) であり，系の運用性能を左右する平滑化を行う施設である．下水道送水系は各ステップ t において，各施設の水位 $h^0(t),h^1(t),\cdots,h^{N-1}(t)$ に対して $w^0(T_{ime}),\cdots,w^{N-1}(T_{ime})$ を考慮して制約条件 (4) 式を逸脱せずに，$u^0(t)$ をなるべく変更しない適切な揚水量 $u^0(t),u^1(t),\cdots,u^{N-1}(t)$ を決定しなければならない．

〈3･2〉 強化学習問題への定式化　以上のようなモデル化の下で強化学習を行うには観測から環境の状態の評価値を同定して適切な行動を学習する TD 法や Q-learning 等の手法[(6) (7)] が有効であると考えられる．

強化学習を下水道送水系に適用するために，状態行動と報酬を定義する．行動 a_t は各施設の揚水量の組み合わせ，状態 s_t は制約違反と $w^i(T_{ime})$ を考慮して定義できる．

$$s_t(\ T_{ime},h^0(t),\cdots,h^{N-1}(t)\ ),\cdots\cdots\cdots\cdots (5)$$

$$a_t(\ u^0(t),\cdots,u^{N-1}(t)\ ) \quad \cdots\cdots\cdots\cdots (6)$$

報酬は制約違反罰 $Penalty(h^i(t))$ と処理場の揚水量切換コスト $Switch(u^0(t),u^0(t-1))$ を線形和して定義する．

$$R_t=\sum_i Penalty(h^i(t)) + \beta\ Switch(u^0(t),u^0(t-1)) \cdots\cdots\cdots (7)$$

この重み β は平滑化性能と制約違反のトレードオフを決める重要なパラメータで適切な値に決める必要がある．

また，揚水量の平滑化とは継続的に同じ行動を選択し続けるということから行動選択をそのステップだけで考えることは難しい．強化学習はそのような政策でも報酬の遅れを考慮して学習できるが，短期記憶を保持して状態観測に加えることで学習効率が上がると予想される．そこで式 (5) に処理場が直前に取った行動 $u^0(t-1)$ を加え，観測する．

$$s_t(\ T_{ime},u^0(t-1),h^0(t),\cdots,h^{N-1}(t)\ )\cdots\cdots (8)$$

以上の設定で学習可能となるが，実際には状態行動空間の爆発の問題に直面する．例えば，揚水量は約 $g=10$ 段階あり，5 施設では約 $g^N=10^5$ 通りの膨大な行動空間から良い行動を選択しなければならない．加えて，予備実験で求めた学習に十分な状態空間の離散化の下では，Q-learning アルゴリズムは約 1000 億の状態行動空間を遷移しながら学習する必要がある．これは試行錯誤しながら学習するには現実的ではないし，メモリ計算量的にも実装不可能である．

〈3･3〉 分散強化学習による接近と協調制御の実現　以上の問題点を回避するために，対象問題の構造が図 3 に示す通り，施設毎に設備を持ち，揚水量を決定できることを利用して，系全体で定義した状態行動及び報酬を施設毎に分割する．これにより施設数に対して指数的に爆発する状態空間及び探索空間を学習可能な大きさにすることができる．

しかし，分割により施設間に強い相互作用があるマルチエージェント系となり，目的達成のための協調動作の獲得などの問題が発生する．ただし，本問題のようにひとつのシステムを分割して扱う場合，各エージェントはシステム全体の状況を把握することができるため，その情報を基にエージェント同士が適切に協調して目的を達成できるような状態空間や報酬設計の再構築が可能である．

〈3･3･1〉 状態行動空間の分割　式（5），(6）に示した状態行動を施設毎に単純に分割すると i=0, ···, N-1 で

$$s_t(\ T_{ime},h^i(t)\ )\ ,\ a_t(\ u^i(t)\ )\cdots\cdots\cdots\cdots\cdots (9)$$

また，式（8）では下式となり，行動は式（9）と同じである．

$$\begin{cases} s_t(\ T_{ime},u^0(t-1),h^0(t)\ ) & \text{処理場} \\ s_t(\ T_{ime},h^i(t)\ ) & \text{ポンプ場} \end{cases} \cdots\cdots (10)$$

表 1　状態表現，報酬設計の例

Table 1.　States and Rewards of Plants.

下水道広域送水系（5 施設）				
施設	状態次元数	行動数	状態観測	報酬
処理場	4	8	時刻（24），貯水池水位（10），前行動（8），上流合計揚水量（10）	揚水量変更量，貯水池水位違反
ポンプ場 1	3	8	時刻（24），貯水池水位（10），上流合計揚水量（10）	処理場流入変化量，貯水池水位違反
ポンプ場 2，3	2	11,8	時刻（24），貯水池水位（10）	処理場流入変化量，貯水池水位違反
ポンプ場 4	2	11	時刻（24），貯水池水位（10）	ポンプ場 1 流入変化量，貯水池水位違反

かっこ内は状態分割数，状態表現は式（11），報酬設計は式（17）．

各施設毎に探索空間を持つことは自然であるから，状態行動 s_t, a_t の分割は容易である．しかし，状態観測が自らに限定されるため他の施設の挙動が不明になり，相互作用を考慮して協調しないと系の目的達成は困難となる．そこで情報共有を行うが系全体の情報は大きいため，相互作用がネットワーク的に離れると間接的になることを利用し，連接施設のみを扱う．この際，情報共有が多ければより相互作用の影響を緩和できるが，観測する状態量の増加による状態爆発や学習速度とのトレードオフが存在する．

共有する情報として上流施設 i' が存在する中継ポンプ場の場合，これらの揚水量の合計 $\sum_{i'} u^{i'}(t)$ は次のステップの貯水池水位の変化に大きく関係がある．この値を観測することで必要な揚水量をかなり絞り込むことができる．

$$\begin{cases} s_t(\ T_{ime}, u^0(t-1), \sum_{i'} u^{i'}(t), h^0(t)\ ) \\ s_t(\ T_{ime}, \sum_{i'} u^{i'}(t), h^i(t)\ ) & \text{中継ポンプ場} \\ s_t(\ T_{ime}, h^i(t)\ ) & \text{端ポンプ場} \end{cases} \cdot (11)$$

さらに，同様の考えから施設 i の揚水量が影響を与える下流施設 i'' の水位 $h^{i''}(t)$ を観測することも有効である．

$$\begin{cases} s_t(\ T_{ime}, u^0(t-1), \sum_{i'} u^{i'}(t), h^0(t)\ ) \\ s_t(\ T_{ime}, \sum_{i'} u^{i'}(t), h^{i''}(t), h^i(t)\ ) \\ s_t(\ T_{ime}, h^{i''}(t), h^i(t)\ ) \end{cases} \cdots\cdots (12)$$

〈3•3•2〉 報酬設計　全体の目的達成のために各個の報酬をどのようするかという報酬設計は分散強化学習の大きな問題のひとつである．相互作用が全く無いとすると個々は独立して報酬を獲得できるので単純に分割できる．本問題では系全体の報酬（式（7））を施設 i 毎に分割できる．

$$r_t^0 = Penalty(h^0(t)) + \beta Switch(u^0(t), u^0(t-1)) \cdots\cdots\cdots (13)$$

$$r_t^i = Penalty(h^i(t)) \cdots\cdots\cdots\cdots\cdots\cdots (14)$$

ここで制約違反罰 $Penalty(h^i(t))$ はほぼ独立に考慮できるが，揚水量切換コスト $Switch(u^0(t), u^0(t-1))$ は処理場だけでは不十分で，ポンプ場の協調が必要である．そこでポンプ場では何らかの方法で協調を誘導するような報酬設計を行う必要がある．以下では r_t^0 は式（13）と同じとする．

- 全体のタスクに関する報酬を共有する方法

$$r_t^i = Penalty(h^i(t)) + \beta Switch(u^0(t), u^0(t-1)) \cdots\cdots (15)$$

全エージェントが全体のタスクを考慮する．しかし，個々の貢献度とは関係なく報酬を獲得し得るため協調が不十分にしか行われない可能性がある．

- 関連エージェントと報酬を共有する方法

$$r_t^i = Penalty(h^i(t)) + \sum_j f(i,j) r_t^j \cdots\cdots (16)$$

ここで $f(i,j)$ はエージェント i, j 間の関連度の関数である．各エージェントは直接報酬 r_t^j を $f(i,j)$ に応じた重みで共有する．問題の構造に応じて報酬を共有できるため，連接施設の相互作用を考慮して協調を促進しうる．実験では $f(i,j)$ を以下のように定義した．

$$f(i,j) = \begin{cases} \frac{1}{\text{下流施設 } i''\text{の数}+1} & j = i''\ or\ i = j \\ 0 & \text{それ以外} \end{cases}$$

- 協調報酬を与える方法

相互作用の強い問題では協調すること自体が難しい．このような場合，全体のタスクのため報酬を共有するだけでは不十分で，協調により報酬を獲得する設定が有効である．このために報酬設計に問題の階層構造を利用するが，本問題のように相互作用による影響が大きい問題では知識を利用した分割は特に重要である．そこで次の報酬を考える．

処理場の平滑化を行うためには，各施設はできうる限りバッファの役割を果たし，1 日の汚水量の増減の波を打ち消すことが必要である．これは階層構造を利用し，下流施設 i'' の流入量 $f^{i''}(t)$ の安定，つまり，下流施設 i'' に影響を及ぼす自施設 i を含む上流側の並列施設 j からの合計揚水量 $\sum_j u^j(t)$ と i'' に流入する汚水量 $w^{i''}(T_{ime})$ の合計 $f^{i''}(t)$ を安定させるということである．そこでその変化量 $Variate(|\bar{f}^{i''} - f^{i''}(t)|)$ を報酬とする．これにより，下流施設 i'' に連接する同階層の並列施設は協調して揚水量を決定する必要が生じ，かつ，処理場の平滑化に貢献することができる．

$$r_t^i = Penalty(h^i(t)) + \beta'\ Variate(|\bar{f}^{i''} - f^{i''}(t)|) \cdot (17)$$

状態表現・報酬設計の例を表 1 に示す．

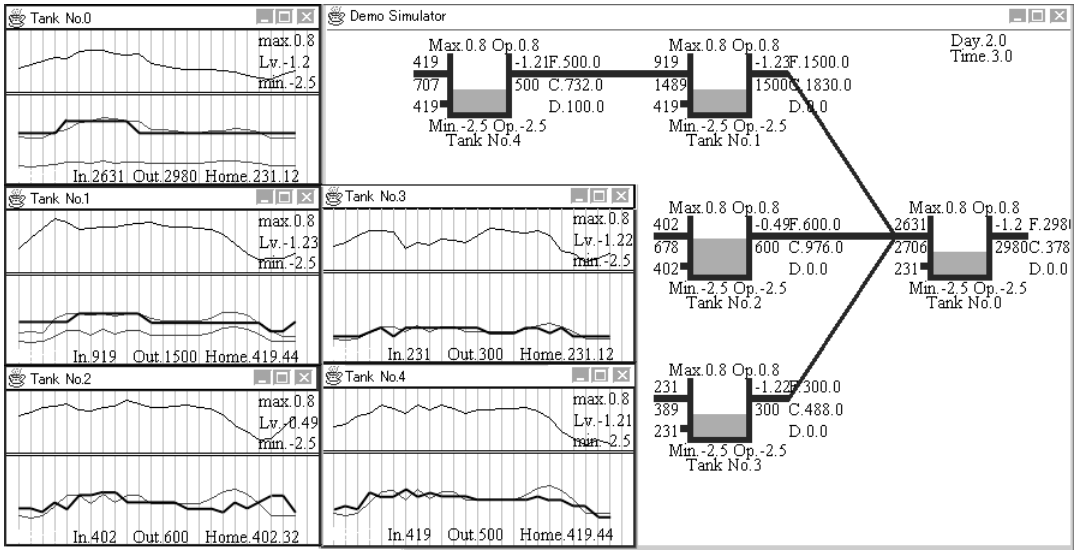

図 4　シミュレータの動作例：下水道送水系 5 施設の場合

Fig. 4.　Simulator, 5 plants Sewerage System.

4. 実験及び考察

〈4･1〉 実験の対象　大規模な処理人口を持つある都市の下水道送水系をモデル化した 5 施設の系（図 1）及び小規模系の 3 施設の系を対象とする．各施設は表 2 に示す諸元の貯水池と複数のポンプを持ち，全施設が可制御で，揚水量は 1 時間毎に決定できる．また，処理場も含めた全施設が集積担当地域を持ち，汚水が流入する．実験では 24 時間分の汚水量の実データ 15 日分を学習に使用する．

〈4･2〉 実験の目的　これまでに 2 章において対象問題及びこれに対する従来法とその問題点を述べ，本研究における立場を明確にした．そして，3 章で下水道送水系の定式化と分散強化学習による解決法を提案した．その妥当性と有用性について検証することが本章の目的である．

- 分散強化学習の有効性の確認
  - ○ 状態行動分割の妥当性
    〈3･3･1〉で述べた状態行動分割について式（9）から（12）の状態表現の下で性能比較を行い，相互作用を考慮した状態表現による性能向上を確認する．
  - ○ 報酬設計の妥当性
    〈3･3･2〉で述べた式（13）から（17）の報酬設計で性能比較を行い，協調獲得を考慮した報酬設計の有効性を確認する．また，平滑化性能と制約違反のトレードオフを決める重み β について検討する．

  以上の実験で得られた知見を基に，獲得される制御規則が処理場の平滑化を協調して行えることを示す．
- 提案手法の実用化へ向けての課題の検討
  コスト削減要求，リスク回避などの実問題からの要請に対して，性能獲得を確認する．また，学習データの生成と評価についても検討する．

表 2　施設の諸元

Table 2.　Parameters of Plants.

下水道送水系（$N=5,3$）				
施設	池底面積 $B_i[m^2]$	運用上下限 [m] h^i_{MAX}〜h^i_{MIN}	揚水ポンプ $P^i[m^3/h]\times$ 台数	対象系
下水処理場	2000	0.8〜-2.5	980×2,2000×2	5，3
ポンプ場 1	1000	0.8〜-2.5	500×2,1000×2	5，3
ポンプ場 2	1000	0.8〜-2.5	300×2, 500×3	5，3
ポンプ場 3	300	0.8〜-2.5	150×2, 400×2	5
ポンプ場 4	600	0.8〜-2.5	300×3, 500×2	5

〈4･3〉 実験の方法

〈4･3･1〉 シミュレータ　実施設とデータを基にシミュレータ（図 4）を作成した．学習アルゴリズムは各ステップで各施設の状態を観測し行動を出力する．シミュレータは 1 時間単位で学習アルゴリズムから行動を受け取り貯水池の水位の変動などを以下の要領でシミュレートする．

（1） $i=N-1$ に初期化
（2） 施設 i はステップ t の時刻 T_{ime}, 水位 $h^i(t)$ を観測
（3） 上流施設 i' の合計揚水量 $\sum_{i'} u^{i'}(t)$ を算出
（4） 学習アルゴリズムに従って揚水量 $u^i_g(t)$ を選択
（5） $i \neq 0$ の時は $i=i-1$ に更新して 2 に戻る
（6） 全施設の汚水量 $w^i(T_{ime})$ をデータより算出
（7） 貯水池水位を更新（式（3））
（8） 制約違反（式（4））や切換コストから報酬を算出
（9） $t=t+1$ に更新して 1 に戻る

〈4･3･2〉 実験の条件　以上の定式化の下で実験を行い，学習した制御規則の評価のために 2 つの指標を導入する．

- 制約違反率　1 日当たりの系の運用上下限逸脱確率
- 切換回数　1 日当たりの処理場の揚水量変更回数

以降に示す学習曲線及び性能の図では 1 万ステップ（1 万学習時間）で平均した値について，各試行のそのステップまでの最良値を求め，試行全体でその値の平均値，最良・

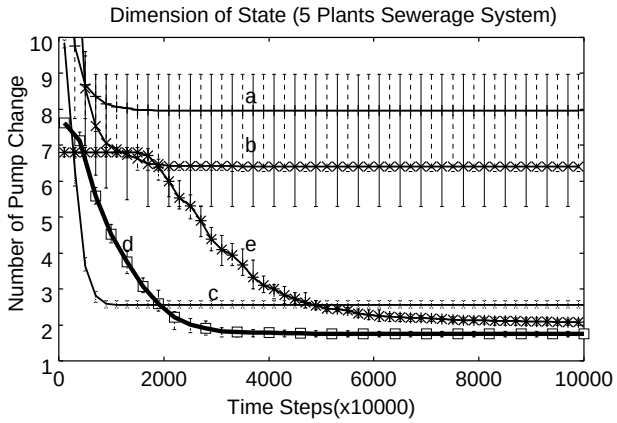

図 5　状態行動分割，性能の学習曲線
Fig. 5.　Performance on dimension of state.
a：eq 9，b：a + $\sum_{i'} u^{i'}(t)$，c,d,e：eq 10, 11,12

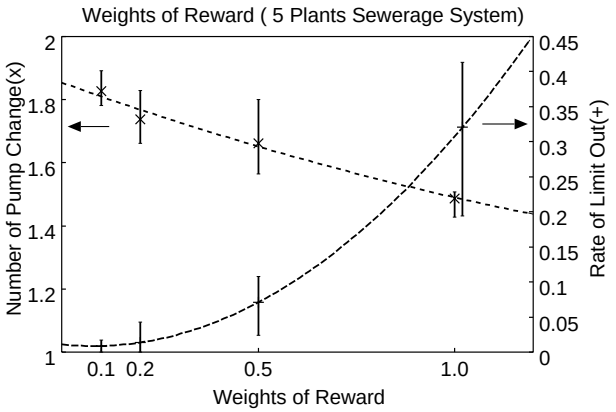

図 6　重み（β）と性能の比較
＋：制約違反率，×：切換回数
Fig. 6.　weights(β) and performance.

最悪値をプロットした．学習率及び割引率は予備実験から α=0.01，γ=0.96 で固定した値を適用した．

〈4･4〉 実験の結果と考察

〈4･4･1〉 状態行動分割の妥当性　下水道 5 施設系での学習性能の違いを図 5 に示す．報酬設計は式（17），重み β は 0.2 とした．グラフの a は式（9）の単純に分割した状態表現で学習した．b はこれに $\sum_{i'} u^{i'}(t)$ の観測を加えたもの，c，d，e は式（10）（11）（12）の状態表現とする．

まず，〈3･2〉において処理場の前行動 $u^0(t-1)$ の短期記憶が平滑化に貢献すると予測したが，a と c，b と d に明らかな差があり有効であることがわかる．

相互作用を考慮した場合，a と b，c と d のように性能が向上する．しかし，e では豊かな状態観測でより性能は向上しうるが，学習速度は遅くなることがわかる．

このように分割した系では，強い相互作用を情報共有して状態として観測することで緩和できる．しかし，観測次元数に対して指数的な状態数の増加や，政策の複雑化のため学習時間が増大してしまう．従って適切な状態表現や分割は極めて重要である．以降の実験では式（11）を用いる．

〈4･4･2〉 報酬設計の妥当性　学習性能の違いを表 3 に示す．重み β は 0.2 とした．式（14）で定義した単純分割報酬ではポンプ場は制約違反は回避できるが平滑化に全く貢献しない．その環境下で処理場はできる限り平滑化を行うが 1 日に約 4 回程度の切換が必要である．

全体のタスクを共有する報酬設計（15）式では，処理場で発生する切換罰に対して報酬の遅れを考慮して学習できるが，必ずしも全施設が協調しなくても報酬が得られてしまうために，あまり協調行動の学習が進まず，性能はある程度向上しても停滞してしまう．

構造を考慮した報酬設計（16）式は比較的良い性能を獲得することもあるが安定して性能は得られない．これは直接報酬の受渡では非隣接施設との協調は困難であり，本問題では同じ施設に連接した並列施設間の協調が特に必要なためと考えられる．また，報酬共有には平滑化性能と制約違反率のトレードオフが重み β によって調節困難になり，制約違反率が高くなってしまう問題があることもわかる．

式（17）で示した協調報酬では比較的早い学習速度で安定した性能を獲得した．局所的な階層毎に協調による報酬を獲得でき，学習を進めることができる．このように相互作用の強い問題においては協調を誘導する報酬分割と，局所的に協調動作を獲得できる設定を行うことが有効である．以降の実験では式（17）の報酬を用いて行う．

従来研究[5]では報酬分割に関する研究があり，式（16）は分散報酬関数と呼ばれるものと同じである．また，分散 value 関数（DVF）という Q 値を更新する際に共有する手法が提案されており，$f(i,j)$=0 にした重みゼロのエージェントも考慮できることを主張しているが，この方法でも十分な協調動作の獲得は難しいことがわかる．

〈4･4･3〉 制約違反率と平滑化のトレードオフ　貯水池水位の制約違反率と処理量平滑化のトレードオフは報酬の重み β によってほぼ決定され，重要なパラメータであることを述べた．図 6 のように制約違反率に対する平滑化の重みの比率を上げていくと平滑化の性能は向上し制約違反率が高まっていく．本来，この重みは制約違反と平滑化性能を金額などのコストに換算して決定すべきものだが，事実上困難なため実験で決定した．重み 0.2 で多くの試行が違反をほぼ 0 にできることからこの値を用いた．

表 3　報酬設計の違いによる性能比較
Table 3.　Performance on Rewards.

下水道 5 施設系（N=5）			
	切り換え回数		平均制約
報酬設計	平均値	最良値～最悪値	違反率（%）
式（14）	3.97	3.94～4.05	4.03
式（15）	3.10	2.48～3.45	7.87
式（16）	2.53	2.07～3.53	14.6
式（17）	**1.73**	**1.66～1.81**	**1.44**
DVF	2.48	2.33～2.71	1.92

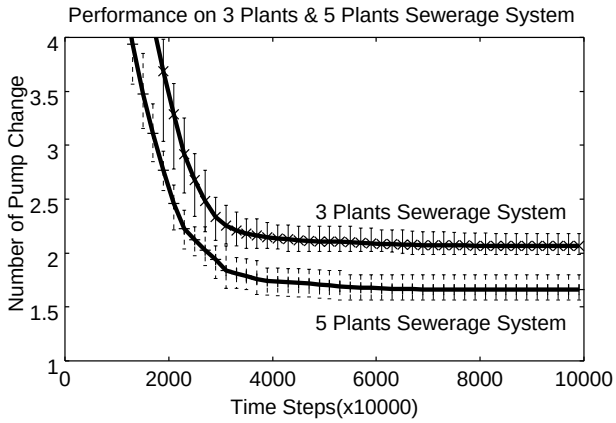

図 7　3 施設と 5 施設下水道系の学習曲線
Fig. 7.　performance on 3 and 5 plants system.

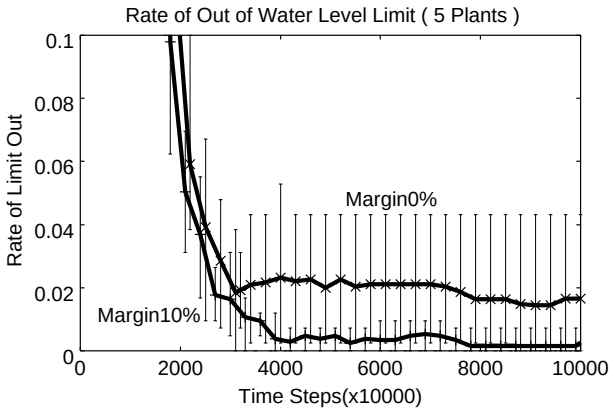

図 9　マージンに対する制約違反率の学習曲線
Fig. 9.　performance with margin on tank level.

〈4･4･4〉 獲得された制御規則　図 7 に示すように下水道 3 施設系で学習により獲得された制御の例を図 8 に示す．制約違反率を数%に抑えたまま，1 日の切換回数をほぼ 2 回で制御し，処理場の揚水量平滑化を実現している．この 1 日約 2 回という制御規則は実用上十分に優秀な性能であると評価されている．5 施設系でもほぼ同様の協調が得られる．

獲得された挙動を見ると，比較的余裕がないポンプ場 1 が水位制約違反を回避するために汚水量の増減に追従する制御を獲得し，比較的余裕のあるポンプ場 2 が汚水量の変化をなるべく処理場に伝えないように逆の政策を獲得している．このような協調動作の結果，処理場に流入する汚水量が安定し平滑化を可能にしている．以上ように分割により独立に学習しているにもかかわらず協調し，施設の能力に対して適応的に政策を獲得できることがわかる．

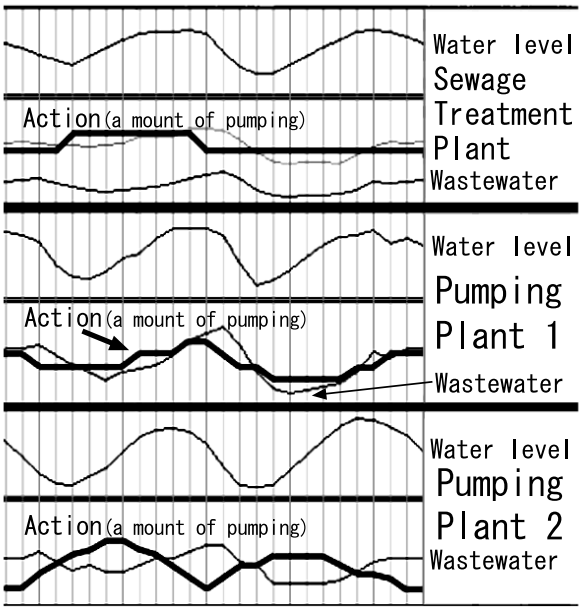

図 8　3 施設下水道系の獲得挙動．24 時間分．
Fig. 8.　Control of 3 plants sewerage system for 24hours.

〈4･4･5〉 実用化へ向けての検討

– **切換間隔を広げることによるコスト低減**　実運用上，処理場で一度切換えたら次の切換をなるべく遅くしたい．つまり水処理に時間がかかるため，一定の状態を保持したいということである．これを切換コストの一部に加えた設定で実験を行った．これにより平均 1.72（最悪 1.81，最良 1.62）回とほぼ性能を維持したまま，切換間隔の長い有効な制御規則を獲得できることを確認した．

– **マージンによるリスク回避**　運用上下限は物理上下限に対して安全マージンを考慮して設定されているため，実運用に照らしても数%の制約違反率は許容される．しかし，計画段階では制約違反がない方が好ましい．そこで運用上下限にさらにマージンを設けることで，貯水池水位の運用領域を狭くして制約違反率を低くすることが期待できる．

図 9 に制約違反率の学習曲線を示す．マージンが 5%〜10%で制約違反はほとんど起こらなくなり，切換回数は平均 2.10（最悪 2.30，最良 2.02）回とそれほど悪化しない．このマージンと重み β によるリスク回避は同時に考慮できる．

– **学習データの生成と評価**　実用性を考える上では，多くの実データを用いて学習するべきであるが，得られない場合に作成したデータで学習し，適用できることが必要である．そこで図 2 の点線で示す範囲で一様乱数によりデータを作成した．この際，実状況に近くなるように終日で比較的流量の多い日や少ない日を混入し，獲得された制御規則に対して，実データを使用して評価を行った．

作成データは実データの幅より約 10%程度大きい幅でランダムに決定し，水位のマージンを 5%で実験した．これにより平均切換 2.25 回（0.036%）の制御規則が獲得され，ランダムに作成したデータを用いて，実データに十分対応できることが確認された．本実験では一様乱数を用いたが，より適切な分布モデルにより改善が期待できる．

5. おわりに

不確実性を有する大規模な実問題として下水道送水系の制御プランニング問題を取りあげ，分散強化学習による解法により接近を試みた．予測を用いた従来のプランニング手法は不確実性のために再計画や補正が必須である．シミュレータを用いた強化学習により実データに含まれる不確実性を直接扱い，予測を用いず，かつ，再計画や補正をすることなく様々な状況に対処できる制御規則を自動的に獲得できることを示した．複数の下水道制御の専門家より，獲得した１日当たり約２回切換の制御規則は実運用の立場から十分評価できるとのコメントを得ている．以上より，提案手法は実用化に向けて極めて有望な枠組みであると考えられる．

分散強化学習は，従来手法では困難であった大規模な状態行動空間を取り扱う上で有用な枠組みである．本論文では対象の構造を利用して相互作用を緩和するための分散状態表現及び協調を誘導するための報酬設計の方法を提案し，有用性を確認した．

本手法の汎用性を示すために上水道送水系の制御プランニング問題への適用を予定している．

今後の課題として，安価な夜間電力を考慮した場合への対応，雨水の流入がある合流式への対応などを検討したい．さらに，予測を利用した学習手法，リスクを考慮した学習手法[(8)(9)]に取り組みたい．

（平成 14 年 5 月 2 日受付，同 14 年 9 月 13 日再受付）

文　献

（1） I.Hayashi: ”GMDH”, Journal of Japan Society for Fuzzy Theory and systems, vol.7 No.2 pp.270–274 (1995)(in Japanese)
林 勲：ＧＭＤＨ，日本ファジィ学会誌，vol.7 No.2 pp.270–274 (1995).

（2） K.Matsumoto,S.Miyata: ”A Space-Time Hierarchical Operation Scheduling Method for Large Scale Water Supply System”, IEE Japan,Vol.102-C No.5 pp.109–116 (1982)(in Japanese)
松本 邦顕，宮岡 伸一郎: 大規模上水道のための時空間階層型運用計画手法，電気学会論文誌 C，102 巻 5 号 pp.109–116 (1982).

（3） Y.Sakamoto, F.Kurokawa, M.Sano, T.Yamada, T.Ashiki, H.Yuki: ”Quasi-Optimization of Water Distribution Scheduling Based on GA”, IEE Japan,Vol.120-B No.8/9 pp.987–999 (2000)(in Japanese)
坂本 義行，黒川 太，佐野 方俊，山田 毅，芦木 達雄，結城博司：ＧＡによる送水計画の近似的最適化手法，電気学会論文誌Ｄ，120 巻 8/9 号 pp.987–999 (2000).

（4） 本田 和広，小林 主一郎，奥 満男，國見 正樹，近藤 真一：AI・ファジィ応用による水処理負荷変動平滑化制御，平成３年電気・情報関連学会連合大会，S4-3 pp.83–86 (1991).

（5） Schneider, J. G., Wong, W.K., Moore, A. & Riedmiller, M.: Distributed Value Functions, *Proceedings of the 16th International Conference on Machine Learning*, pp.371–378 (1999).

（6） Sutton, R. S. & Barto, A.: Reinforcement Learning: An Introduction, *A Bradford Book*, The MIT Press (1998).

（7） Watkins, C. J. C. H. & Dayan, P.: Technical Note: *Q*-Learning, *Machine Learning 8*, pp.279–292 (1992).

（8） Geibel, P.: Reinforcement Learning with Bounded Risk, *Proceedings of the 18th International Conference on Machine Learning*, pp.162–169 (2001).

（9） M.Sato,H.Kimura,S.Kobayashi: ”TD Algorithm for the Variance of Return and Mean-Variance Reinforcement Learning”, Journal of Japanese Society for Artificial Intelligence,Vol.16,No.3 pp.353–362 (2001)(in Japanese)
佐藤 誠, 木村 元, 小林 重信: 報酬の分散を推定する TD アルゴリズムと Mean-Variance 強化学習法の提案, 人工知能学会誌,Vol.16,No.3 pp.353–362 (2001).

青　木　　圭　（非会員） 2002 年東京工業大学大学院知能システム科学専攻修士課程修了．同年４月博士課程在籍中．主に強化学習に関する研究に従事．

木　村　　元　（正員） 1997 年東京工業大学大学院知能科学専攻博士課程修了．同年４月日本学術振興会ＰＤ研究員．1998 年４月，東京工業大学大学院総合理工学研究科助手．現在に至る．人工知能，特に強化学習に関する研究に従事．

長　岩　明　弘　（正員） 1989 年九州大学工学部大学院電気工学修士課程修了。同年（株）東芝に入社。主として公共・社会システムの研究開発に従事。1999 年計測自動制御学会論文賞受賞。現在，同社電力・産業システム技術開発センター社会システム開発部公共制御システム担当主務。計測自動制御学会会員。

小　林　重　信　（非会員） 1974 年東京工業大学大学院経営工学専攻博士課程終了．同年４月，同大学工学部制御工学科助手．1981 年８月，同大学大学院総合理工学研究科助教授．1990 年８月，教授．現在に至る．問題解決と推論制御，知識獲得と学習などの研究に従事．