(12) 特許協力条約に基づいて公開された国際出願

(19) 世界知的所有権機関
国際事務局

(43) 国際公開日
2022年11月3日(03.11.2022)

WIPO | PCT

(10) 国際公開番号
WO 2022/230019 A1

(51) 国際特許分類:
G06N 20/00 (2019.01)

(21) 国際出願番号： PCT/JP2021/016630

(22) 国際出願日： 2021年4月26日(26.04.2021)

(25) 国際出願の言語： 日本語

(26) 国際公開の言語： 日本語

(71) 出願人: 日本電気株式会社 (**NEC CORPORATION**) [JP/JP]; 〒1088001 東京都港区芝五丁目７番１号 Tokyo (JP).

(72) 発明者: 江藤 力 (**ETO Riki**); 〒1088001 東京都港区芝五丁目７番１号 日本電気株式会社内 Tokyo (JP).

(74) 代理人: 岩壁 冬樹, 外(**IWAKABE Fuyuki** et al.); 〒1040031 東京都中央区京橋二丁目８番７号 読売八重洲ビル６階 サンライズ国際特許事務所 Tokyo (JP).

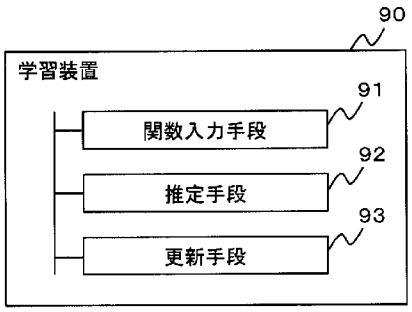
(81) 指定国(表示のない限り、全ての種類の国内保護が可能): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL,

WO 2022/230019 A1

(54) **Title:** LEARNING DEVICE, LEARNING METHOD, AND LEARNING PROGRAM

(54) 発明の名称： 学習装置、学習方法および学習プログラム

[図4]



90 Learning device
91 Function input means
92 Estimation means
93 Updating means

(57) **Abstract:** In the present invention, a function input means 91 receives an input of a reward function in which a feature quantity is set so as to satisfy the Lipschitz continuity condition. An estimation means 92 estimates a trajectory that minimizes a Wasserstein distance that represents the distance between the probability distribution of a trajectory of an expert and the probability distribution of a trajectory that is determined on the basis of a parameter of the reward function. An updating means 93 updates, on the basis of the estimated trajectory, the parameter of the reward function so as to maximize the log-likelihood of the Boltzmann distribution derived on the basis of the principle of maximum entropy.

(57) 要約：関数入力手段９１は、リプシッツ連続条件を満たすように特徴量が設定された報酬関数の入力を受け付ける。推定手段９２は、熟練者の軌跡の確率分布と、報酬関数のパラメータに基づいて決定される軌跡の確率分布との距離を表わすワッサースタイン距離を最小にする軌跡を推定する。更新手段９３は、推定された軌跡に基づいて、最大エントロピーの原理で導かれるボルツマン分布の対数尤度を最大にするように報酬関数のパラメータを更新する。

[続葉有]

ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) 指定国(表示のない限り、全ての種類の広域保護が可能): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), ユーラシア (AM, AZ, BY, KG, KZ, RU, TJ, TM), ヨーロッパ (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

添付公開書類:

一　国際調査報告（条約第21条(3)）

明　細　書

発明の名称 ： 学習装置、学習方法および学習プログラム

技術分野

[0001] 本発明は、逆強化学習を行う学習装置、学習方法および学習プログラムに関する。

背景技術

[0002] 機械学習の手法の一つに強化学習（ＲＬ：Reinforcement Learning）が知られている。強化学習は、様々な行動を試行錯誤しながら価値を最大化するような行動を学習する手法である。強化学習では、この価値を評価するための報酬関数が設定され、この報酬関数を最大にするような行動が探索される。しかし、報酬関数の設定は、一般には困難である。

[0003] この報酬関数の設定を容易にする方法として、逆強化学習（ＩＲＬ：Inverse Reinforcement Learning）が知られている。逆強化学習では、熟練者の意思決定履歴データを利用して、報酬関数を用いた最適化と、報酬関数のパラメータの更新とを繰り返すことで、熟練者の意図を反映する報酬関数を生成する。

[0004] 非特許文献１には、逆強化学習の一つである最大エントロピー逆強化学習（ＭＥ－ＩＲＬ：Maximum Entropy-IRL ）について記載されている。ＭＥ－ＩＲＬでは、最大エントロピー原理を用いて軌跡の分布を指定し、真の分布へ近づけること（すなわち、最尤推定）により報酬関数を学習する。これにより、熟練者の軌跡（行動履歴）を再現する報酬関数が複数存在するという不定性を解決している。

[0005] また、非特許文献２には、最大エントロピー逆強化学習を改良した逆強化学習の手法の一つであるＧＣＬ（Guided Cost Learning）について記載されている。非特許文献２に記載された手法では、重点サンプリングを用いて報酬関数の重みを更新する。

先行技術文献

非特許文献

[0006] 非特許文献1：B. D. Ziebart, A. Maas, J. A. Bagnell, and A. K. Dey, “Maximum entropy inverse reinforcement learning”, In AAAI, AAAI’08, 2008.

非特許文献2：Chelsea Finn, Sergey Levine, Pieter Abbeel, ”Guided Cost Learning: Deep Inverse Optimal Control via Policy Optimization”, Proceedings of The 33rd International Conference on Machine Learning, PMLR 48, pp.49-58, 2016.

発明の概要

発明が解決しようとする課題

[0007] 一方、非特許文献１に記載されたＭＥ－ＩＲＬでは、学習の際、可能な全ての軌跡に対する報酬の総和を計算する必要がある。しかし、現実的には、全ての軌跡に対する報酬の総和を計算することは困難である。

[0008] この問題に対し、非特許文献２に記載されたＧＣＬでは、重点サンプリングで近似的にこの値を算出する。ここで、ＧＣＬによる重点サンプリングを用いる場合、サンプリング自体の分布を想定する必要がある。しかし、組み合わせ最適化問題のように、サンプリング分布をどのように設定してよいか分からない問題も存在するため、非特許文献２に記載された方法では、様々な数理最適化に適用できないという問題がある。

[0009] そこで、本発明は、逆強化学習における不定性の問題を解決しつつ、組み合わせ最適化のような数理最適化問題に対しても適用可能な逆強化学習を行うことができる学習装置、学習方法および学習プログラムを提供することを目的とする。

課題を解決するための手段

[0010] 本発明による学習装置は、リプシッツ連続条件を満たすように特徴量が設定された報酬関数の入力を受け付ける関数入力手段と、熟練者の軌跡の確率分布と、報酬関数のパラメータに基づいて決定される軌跡の確率分布との距

離を表わすワッサースタイン距離を最小にする軌跡を推定する推定手段と、推定された軌跡に基づいて、最大エントロピーの原理で導かれるボルツマン分布の対数尤度を最大にするように報酬関数のパラメータを更新する更新手段とを備え、更新手段が、対数尤度の下限として、ワッサースタイン距離から、パラメータに対する最大の報酬値からそのパラメータに対する報酬の平均値を減じた式で定義されるエントロピー正則化項を減じる式を導出し、導出された対数尤度の下限を最大にするように報酬関数のパラメータを更新することを特徴とする。

[0011] 本発明による学習方法は、コンピュータが、リプシッツ連続条件を満たすように特徴量が設定された報酬関数の入力を受け付け、コンピュータが、熟練者の軌跡の確率分布と、報酬関数のパラメータに基づいて決定される軌跡の確率分布との距離を表わすワッサースタイン距離を最小にする軌跡を推定し、コンピュータが、推定された軌跡に基づいて、最大エントロピーの原理で導かれるボルツマン分布の対数尤度を最大にするように報酬関数のパラメータを更新し、パラメータの更新の際、コンピュータが、対数尤度の下限として、ワッサースタイン距離から、パラメータに対する最大の報酬値からそのパラメータに対する報酬の平均値を減じた式で定義されるエントロピー正則化項を減じる式を導出し、導出された対数尤度の下限を最大にするように報酬関数のパラメータを更新することを特徴とする。

[0012] 本発明による学習プログラムは、コンピュータに、リプシッツ連続条件を満たすように特徴量が設定された報酬関数の入力を受け付ける関数入力処理、熟練者の軌跡の確率分布と、報酬関数のパラメータに基づいて決定される軌跡の確率分布との距離を表わすワッサースタイン距離を最小にする軌跡を推定する推定処理、および、推定された軌跡に基づいて、最大エントロピーの原理で導かれるボルツマン分布の対数尤度を最大にするように報酬関数のパラメータを更新する更新処理を実行させ、更新処理で、対数尤度の下限として、ワッサースタイン距離から、パラメータに対する最大の報酬値からそのパラメータに対する報酬の平均値を減じた式で定義されるエントロピー正

則化項を減じる式を導出させ、導出された対数尤度の下限を最大にするように報酬関数のパラメータを更新させることを特徴とする。

発明の効果

[0013] 本発明によれば、逆強化学習における不定性の問題を解決しつつ、組み合わせ最適化のような数理最適化問題に対しても適用可能な逆強化学習を行うことができる。

図面の簡単な説明

[0014] [図1]本開示による学習装置の一実施形態の構成例を示すブロック図である。

[図2]学習装置の動作例を示すフローチャートである。

[図3]ロボット制御システムの一実施形態の構成例を示すブロック図である。

[図4]本開示による学習装置の概要を示すブロック図である。

[図5]少なくとも１つの実施形態に係るコンピュータの構成を示す概略ブロック図である。

発明を実施するための形態

[0015] 理解を容易にするため、初めに、本実施形態で前提としている最大エントロピー逆強化学習の問題設定、手法および課題について説明する。ＭＥ－ＩＲＬでは、以下の問題設定を想定している。すなわち、熟練者のデータＤ＝｛τ_1，τ_2，…τ_N｝（ただし、τ_i＝（（s_1，a_1），（s_2，a_2），…，（s_N，a_N））））から、ただ１つの報酬関数Ｒ（ｓ，ａ）＝θ・ｆ（ｓ，ａ）を推定するという設定である。ＭＥ－ＩＲＬでは、θを推定することで、熟練者の意思決定を再現できる。

[0016] 次に、ＭＥ－ＩＲＬの手法を説明する。ＭＥ－ＩＲＬでは、軌跡τは、以下に例示する式１で表わされ、軌跡の分布$p_\theta(\tau)$を表わす確率モデルは、以下に例示する式２で表わされる。式２における$\theta^T f_\tau$が報酬関数を表わす（式３参照）。また、Ｚは、全ての軌跡に対する報酬の総和を表わす（式４参照）。

[0017]

[数1]

$$\tau = \{(s_t, a_t)|t = 0, \cdots, T\} \qquad (式1)$$

$$p_\theta(\tau) := \frac{1}{Z}\exp(\theta^\top f_\tau) \qquad (式2)$$

ただし、

$$f_\tau := \sum_{t=0}^{T} \gamma^t f_{s_t,a_t} \qquad (式3)$$

$$Z := \sum_{\tau} \exp(\theta^\top f_\tau) \qquad (式4)$$

[0018] そして、最尤推定による報酬関数の重みの更新則（具体的には、勾配上昇法）は、以下に例示する式５および式６で表わされる。式５におけるαはステップ幅であり、Ｌ（θ）は、ＭＥ－ＩＲＬで用いられる分布間の距離尺度である。

[0019] [数2]

$$\theta \leftarrow \theta + \alpha\nabla_\theta \mathcal{L}(\theta) \qquad (式5)$$

$$\mathcal{L}(\theta) := \frac{1}{N}\sum_{n=1}^{N} \log p_\theta(\tau) = \frac{1}{N}\sum_{n=1}^{N}(\theta^\top f_{\tau^{(n)}})) - \log \sum_{\tau} \exp(\theta^\top f_\tau) \qquad (式6)$$

[0020] 式６における第２項は、全ての軌跡に対する報酬の総和である。ＭＥ－ＩＲＬでは、この第２項の値を厳密に計算できることを前提としている。しかし、現実的には、全ての軌跡に対する報酬の総和を計算することは困難である。以上が、ＭＥ－ＩＲＬの問題設定、手法および課題である。

[0021] 以下、実施形態を図面を参照して説明する。

[0022] 図１は、本開示による学習装置の一実施形態の構成例を示すブロック図である。本実施形態の学習装置１００は、機械学習により、対象者（熟練者）の行動から報酬関数を推定する逆強化学習を行う装置であり、熟練者の行動特性に基づく情報処理を具体的に行う装置である。学習装置１００は、記憶部１０と、入力部２０と、特徴量設定部３０と、重み初期値設定部４０と、数理最適化実行部５０と、重み更新部６０と、収束判定部７０と、出力部８

０とを備えている。

[0023] なお、数理最適化実行部５０、重み更新部６０および収束判定部７０により、後述する逆強化学習が行われることから、数理最適化実行部５０、重み更新部６０および収束判定部７０を含む装置を、逆強化学習装置と言うことができる。

[0024] 記憶部１０は、学習装置１００が各種処理を行うために必要な情報を記憶する。記憶部１０は、後述する入力部２０が受け付けた熟練者の意思決定履歴データ（軌跡）を記憶してもよい。また、記憶部１０は、後述する数理最適化実行部５０および重み更新部６０が学習に用いる報酬関数の特徴量の候補を記憶していてもよい。ただし、特徴量の候補は、必ずしも目的関数に使用される特徴量である必要はない。

[0025] また、記憶部１０は、後述する数理最適化実行部５０を実現するための数理最適化ソルバを記憶していてもよい。なお、数理最適化ソルバの内容は任意であり、実行する環境や装置に応じて決定されればよい。

[0026] 入力部２０は、学習装置１００が各種処理を行うために必要な情報の入力を受け付ける。入力部２０は、例えば、上述する熟練者の意思決定履歴データ（具体的には、状態と行動のペア）の入力を受け付けてもよい。また、入力部２０は、後述する逆強化学習装置が逆強化学習を行う際に用いる初期状態および制約条件の入力を受け付けてもよい。

[0027] 特徴量設定部３０は、状態および行動を含むデータから、報酬関数の特徴量を設定する。具体的には、特徴量設定部３０は、後述する逆強化学習装置が分布間の距離尺度としてワッサースタイン（Ｗａｓｓｅｒｓｔｅｉｎ）距離を利用できるように、関数全体で接線の勾配が有限になるように報酬関数の特徴量を設定する。特徴量設定部３０は、例えば、リプシッツ連続条件を満たすように報酬関数の特徴量を設定してもよい。

[0028] 例えば、f_τを軌跡τの特徴量ベクトルとする。報酬関数$\theta^\top f_\tau$の線形の場合、写像Ｆ：$\tau \to f_\tau$がリプシッツ連続であれば、$\theta^\top f_\tau$もリプシッツ連続である。そのため、特徴量設定部３０は、報酬関数が線形関数になるように特

徴量を設定してもよい。

[0029] なお、例えば、以下に例示する式７は、ａ$_0$において勾配が無限大になってしまうため、本開示において不適切な報酬関数と言える。

[0030] [数3]

$$f_\tau = \begin{cases} 1 & (a_0 \geq 0) \\ 0 & (otherwise) \end{cases} \quad \text{（式7）}$$

[0031] 特徴量設定部３０は、例えば、ユーザの指示に応じて特徴量が設定された報酬関数を決定してもよく、記憶部１０からリプシッツ連続条件を満たす報酬関数を取得してもよい。

[0032] 重み初期値設定部４０は、報酬関数の重みを初期化する。具体的には、重み初期値設定部４０は、報酬関数に含まれる個々の特徴量の重みを設定する。なお、重みを初期化する方法は特に限定されず、ユーザ等に応じて予め定められた任意の方法に基づいて重みが初期化されればよい。

[0033] 数理最適化実行部５０は、熟練者の軌跡（行動履歴）の確率分布と、最適化された（報酬関数の）パラメータに基づいて決定される軌跡の確率分布との間の距離を最小にする軌跡τ^（τ^は、τの上付き^）を導出する。具体的には、数理最適化実行部５０は、分布間の距離尺度として、ワッサースタイン距離を利用して、そのワッサースタイン距離を最小にするよう数理最適化を実行することにより、熟練者の軌跡τ^を推定する。

[0034] ワッサースタイン距離は、以下に例示する式８で定義される。すなわち、ワッサースタイン距離は、熟練者の軌跡の確率分布と、報酬関数のパラメータに基づいて決定される軌跡の確率分布との距離を表わす。なお、ワッサースタイン距離の制約から、報酬関数θ$^\top$ｆ$_\tau$は、リプシッツ連続条件を満たす関数である必要がある。一方、本実施形態では、特徴量設定部３０によってリプシッツ連続条件を満たすように報酬関数の特徴量が設定されているため、数理最適化実行部５０は、以下に例示するようなワッサースタイン距離を利用することが可能になる。

[0035] [数4]

$$W(\theta) := \frac{1}{N}\sum_{n=1}^{N}(\theta^{\top} f_{\tau^{(n)}}) - \frac{1}{N}\sum_{n=1}^{N}(\theta^{\top} f_{\tau_{\theta}^{(n)}}) \qquad \text{（式8）}$$

[0036] 上記に例示する式８で定義されるワッサースタイン距離は０以下の値をとり、この値を大きくすることが、分布同士を近づけることに対応する。また、式８の第２項において、$\tau_{\theta(n)}$は、パラメータθで最適化したｎ番目の軌跡を表わす。式８の第２項は、組み合わせ最適化問題でも算出可能な項である。そのため、式８に例示するワッサースタイン距離を分布間の距離尺度として用いることで、組み合わせ最適化問題のような数理最適化問題に対しても適用可能な逆強化学習を行うことができる。

[0037] 重み更新部６０は、推定された熟練者の軌跡$\tau\hat{}$に基づいて分布間の距離尺度を最大にするように報酬関数のパラメータθを更新する。ここで、最大エントロピー逆強化学習（すなわち、ＭＥ－ＩＲＬ）では、最大エントロピー原理により軌跡τがボルツマン分布に従うとされる。そこで、重み更新部６０は、ＭＥ－ＩＲＬと同様、上記式５および式６に示すように、推定された熟練者の軌跡$\tau\hat{}$に基づいて、最大エントロピーの原理で導かれるボルツマン分布の対数尤度を最大にするように報酬関数のパラメータθを更新する。

[0038] なお、更新の際、本実施形態の重み更新部６０は、式６における第２項（すなわち、全ての軌跡に対する報酬の総和）から、ｌｏｇ　ｓｕｍ　ｅｘｐｏｎｅｎｔｉａｌ（以下、ｌｏｇＳｕｍＥｘｐと記す。）の上限値を導出する。すなわち、重み更新部６０は、ＭＥ－ＩＲＬで用いられる分布間の距離尺度における下限Ｌ＿（θ）（Ｌ＿は、Ｌの下付きアンダーバーを示す。）を、以下の式９のように導出する。以下、導出された式を、単に対数尤度の下限と記すこともある。

[0039]

[数5]

$$\mathcal{L}(\theta) = \frac{1}{N}\sum_{n=1}^{N}\theta^{\top} f_{\tau^{(n)}} - \log\sum_{\tau}\exp(\theta^{\top} f_{\tau})$$
$$\geq \frac{1}{N}\sum_{n=1}^{N}\theta^{\top} f_{\tau^{(n)}} - \max_{n}\theta^{\top} f_{\tau_{\theta}^{(n)}} - \log N_{\tau} =: \underline{\mathcal{L}}(\theta) \qquad \text{（式9）}$$

[0040] 対数尤度の下限を示す式９における第２項は、現在のパラメータθに対する最大の報酬値であり、第３項は、取り得る軌跡の数（N_{τ}）のｌｏｇ値（対数値）である。このように、重み更新部６０は、ＭＥ－ＩＲＬの対数尤度をもとに、現在のパラメータθに対する最大の報酬値および取り得る軌跡の数（N_{τ}）のｌｏｇ値（対数値）を、軌跡の確率分布から減じて算出される対数尤度の下限を導出する。

[0041] さらに、重み更新部６０は、導出されたＭＥ－ＩＲＬの対数尤度の下限の式を、ワッサースタイン距離からエントロピー正則化項を減じる式に変形する。ＭＥ－ＩＲＬの対数尤度の下限の式を、ワッサースタイン距離とエントロピー正則化項とに分解した式は、以下に例示する式１０のように表される。

[0042] [数6]

$$\underline{\mathcal{L}}(\theta) = \left(\frac{1}{N}\sum_{n=1}^{N}\theta^{\top} f_{\tau^{(n)}} - \frac{1}{N}\sum_{n=1}^{N}\theta^{\top} f_{\tau_{\theta}^{(n)}}\right) - \left(\max_{n}\theta^{\top} f_{\tau_{\theta}^{(n)}} - \frac{1}{N}\sum_{n=1}^{N}\theta^{\top} f_{\tau_{\theta}^{(n)}}\right) + \text{const.}$$

（式10）

[0043] 式１０の１つ目の括弧内の式は、上記式８と同様、ワッサースタイン距離を表わす。また、式１０の２つ目の括弧内の式は、最大エントロピー原理より導かれたボルツマン分布の対数尤度上昇に寄与するエントロピー正則化項を表わす。具体的には、式１０に例示するエントロピー正則化項（すなわち、式１０の２つ目の括弧内の式）において、第１項は、現在のパラメータθに対する最大の報酬値を表わし、第２項は、現在のパラメータθに対する報

酬の平均値を表わす。

[0044] この第２項が、エントロピー正則化項として機能する理由を説明する。ＭＥ－ＩＲＬの対数尤度の下限を最大化するためには、第２項の値を小さくする必要があり、それは、最大の報酬値と平均値との差を小さくすることに相当する。最大の報酬値と平均値との差が小さくなることは、軌跡のばらつきが小さくなることを示す。

[0045] 言い換えると、最大の報酬値と平均値との差が小さくなることは、エントロピーが増加することを意味するため、エントロピー正則化が機能することになり、エントロピーの最大化に寄与することになる。これは、ボルツマン分布の対数尤度の最大化に寄与していることになり、結果として、逆強化学習における不定性の解決に寄与することになる。

[0046] 重み更新部６０は、上記に示す式１０に基づいて、例えば、推定された軌跡$\tau\hat{}$を固定して、勾配上昇法によりパラメータθを更新する。ただし、通常の勾配上昇法では、収束しない恐れがある。エントロピー正則化項において、最大の報酬値をとる軌跡の特徴量（$f_{\tau\theta max}$）は、他の軌跡の特徴量（$f_{\tau(n)}$）の平均値に一致することはない（すなわち、両者の差が０にはならない）。よって、通常の勾配上昇法では、対数尤度が振動して収束しないため安定せず、収束判定を適切に行うことが困難なためである（下記式１１参照）。

[0047] [数7]

$$\frac{\partial \mathcal{L}}{\partial \theta} = \frac{1}{N}\sum_{n=1}^{N} f_{\tau^{(n)}} - f_{\tau_\theta^{\max}} \neq 0 \qquad \text{（式11）}$$

[0048] そこで、本実施形態の重み更新部６０は、勾配法を用いる際、エントロピー正則化に寄与する部分（すなわち、エントロピー正則化項に対応する部分）を徐々に減衰させるようにパラメータθを更新してもよい。具体的には、重み更新部６０は、エントロピー正則化に寄与する部分に、減衰の程度を示す減衰係数β_tを設定した更新式を定義する。重み更新部６０は、例えば、上

記の式１０をθで微分し、ワッサースタイン距離を示す項に対応する部分（すなわち、ワッサースタイン距離を大きくする処理に寄与する部分）と、エントロピー正則化項に対応する部分のうち、エントロピー正則化項に対応する部分に減衰係数を設定した以下に例示する式１２を定義する。

[0049] [数8]

$$\frac{\partial \mathcal{L}}{\partial \theta} \approx \left(\frac{1}{N} \sum_{n=1}^{N} f_{\tau^{(n)}} - \frac{1}{N} \sum_{n=1}^{N} f_{\tau_\theta^{(n)}} \right) - \beta_t \left(f_{\tau_\theta^{\max}} - \frac{1}{N} \sum_{n=1}^{N} f_{\tau_\theta^{(n)}} \right)$$

（式12）

[0050] 減衰係数は、エントロピー正則化項に対応する部分を減衰させる方法に応じて予め定義される。例えば、滑らかに減衰させる場合、β_tは、以下に例示する式１３のように定義される。

[0051] [数9]

$$\beta_t = \beta_1 (t-1)^{-\beta_2}$$

（式13）

[0052] 式１３において、β_1は１に設定され、β_2は０以上に設定される。また、ｔは反復回数を示す。これにより、減衰係数β_tは、反復回数ｔが大きくなるほど、エントロピー正則化項に対応する部分を小さくする係数として機能する。

[0053] また、ワッサースタイン距離は、ＫＬダイバージェンスである対数尤度よりも弱位相なため、対数尤度を０に近づけることで、ワッサースタイン距離も０に近づけることが可能である。そこで、重み更新部６０は、更新の初期段階ではエントロピー正則化項に対応する部分を減衰させずにパラメータθを更新し、対数尤度が振動し始めたタイミングで、エントロピー正則化項に対応する部分の影響を低減させるようにパラメータθを更新してもよい。

[0054] 具体的には、重み更新部６０は、上記に示す式１２を用いて、初期段階では減衰係数β_t＝１としてパラメータθを更新する。その後、重み更新部６０

は、対数尤度が振動し始めたタイミングで減衰係数$\beta_t=0$に変更することで、エントロピー正則化項に対応する部分の影響をなくしてパラメータθを更新してもよい。

[0055] 重み更新部６０は、例えば、対数尤度の移動平均が一定になったときに、対数尤度が振動し始めたと判断してもよい。具体的には、重み更新部６０は、“対数尤度の下限”の時間窓（現在の値から過去の数点）における移動平均の変化が微小（例えば$1e^{-3}$以下）のときに、移動平均が一定になったと判断してもよい。

[0056] また、重み更新部６０は、対数尤度が振動し始めたタイミングで、いきなり減衰係数$\beta_t=0$とせず、まずは、振動係数を上記に示す式１３のように変更してもよい。そして、重み更新部６０は、変更後、対数尤度がさらに振動し始めたタイミングで、減衰係数$\beta_t=0$に変更してもよい。振動し始めたタイミングの判断方法は、上記方法と同様である。

[0057] さらに、重み更新部６０は、振動係数を上記に示す式１３のように変更した後、対数尤度がさらに振動し始めたタイミングでパラメータθの更新方法を変更してもよい。具体的には、重み更新部６０は、以下の式１４に例示するようなモーメンタム法を用いて、パラメータθを更新してもよい。式１４におけるγ_1およびαの値は予め定められる。例えば、$\gamma_1=0.9$、$\alpha=0.001$と定められていてもよい。

[0058] [数10]

$$m_t=\gamma_1 m_{t-1}+(1-\gamma_1)\frac{\partial \mathcal{L}}{\partial \theta}$$

$$\theta_{t+1}=\theta_t+\alpha_t m_t \qquad \text{（式14）}$$

[0059] 以降、後述する収束判定部７０によって、対数尤度の下限が収束されたと判定されるまで、数理最適化実行部５０による軌跡の推定処理、および、重み更新部６０によるパラメータθの更新処理が繰り返される。

[0060] 収束判定部７０は、分布間の距離尺度が収束したか否か判定する。具体的

には、収束判定部７０は、対数尤度の下限が収束したか否か判定する。判定方法は任意であり、収束判定部７０は、例えば、対数尤度の下限の値の絶対値が予め定めた閾値より小さくなったときに、分布間の距離尺度が収束したと判定してもよい。

[0061] 収束判定部７０は、分布間の距離尺度が収束していないと判断した場合、数理最適化実行部５０および重み更新部６０による処理を継続させる。一方、収束判定部７０は、分布間の距離尺度が収束したと判断した場合、数理最適化実行部５０および重み更新部６０による処理を終了させる。

[0062] 出力部８０は、学習された報酬関数を出力する。

[0063] 入力部２０と、特徴量設定部３０と、重み初期値設定部４０と、数理最適化実行部５０と、重み更新部６０と、収束判定部７０と、出力部８０とは、プログラム（学習プログラム）に従って動作するコンピュータのプロセッサ（例えば、ＣＰＵ（Central Processing Unit ））によって実現される。

[0064] 例えば、プログラムは、学習装置１００が備える記憶部１０に記憶され、プロセッサは、そのプログラムを読み込み、プログラムに従って、入力部２０、特徴量設定部３０、重み初期値設定部４０、数理最適化実行部５０、重み更新部６０、収束判定部７０および出力部８０として動作してもよい。また、学習装置１００の機能がＳａａＳ（Software as a Service ）形式で提供されてもよい。

[0065] また、入力部２０と、特徴量設定部３０と、重み初期値設定部４０と、数理最適化実行部５０と、重み更新部６０と、収束判定部７０と、出力部８０とは、それぞれが専用のハードウェアで実現されていてもよい。また、各装置の各構成要素の一部又は全部は、汎用または専用の回路（circuitry ）、プロセッサ等やこれらの組合せによって実現されもよい。これらは、単一のチップによって構成されてもよいし、バスを介して接続される複数のチップによって構成されてもよい。各装置の各構成要素の一部又は全部は、上述した回路等とプログラムとの組合せによって実現されてもよい。

[0066] また、学習装置１００の各構成要素の一部又は全部が複数の情報処理装置

や回路等により実現される場合には、複数の情報処理装置や回路等は、集中配置されてもよいし、分散配置されてもよい。例えば、情報処理装置や回路等は、クライアントサーバシステム、クラウドコンピューティングシステム等、各々が通信ネットワークを介して接続される形態として実現されてもよい。

[0067] 次に、本実施形態の学習装置１００の動作を説明する。図２は、本実施形態の学習装置１００の動作例を示すフローチャートである。入力部２０は、エキスパートデータ（すなわち、熟練者の軌跡／意思決定履歴データ）の入力を受け付ける（ステップＳ１１）。特徴量設定部３０は、状態および行動を含むデータから、リプシッツ連続条件を満たすように報酬関数の特徴量を設定する（ステップＳ１２）。また、重み初期値設定部４０は、報酬関数の重み（パラメータ）を初期化する（ステップＳ１３）。

[0068] 数理最適化実行部５０は、リプシッツ連続条件を満たすように特徴量が設定された報酬関数の入力を受け付ける（ステップＳ１４）。そして、数理最適化実行部５０は、ワッサースタイン距離を最小にするように数理最適化を実行する（ステップＳ１５）。具体的には、数理最適化実行部５０は、熟練者の軌跡の確率分布と、報酬関数のパラメータに基づいて決定される軌跡の確率分布との距離を表わすワッサースタイン距離を最小にする軌跡を推定する。

[0069] 重み更新部６０は、推定された軌跡に基づいてボルツマン分布の対数尤度を最大にするように報酬関数のパラメータを更新する（ステップＳ１６）。このとき、重み更新部６０は、対数尤度の下限を導出し、導出された対数尤度の下限を最大にするように報酬関数のパラメータを更新する。

[0070] 収束判定部７０は、対数尤度の下限が収束したか否か判定する（ステップＳ１７）。対数尤度の下限が収束していないと判定された場合（ステップＳ１７におけるＮｏ）、更新された軌跡を用いてステップＳ１５以降の処理が繰り返される。一方、対数尤度の下限が収束したと判定された場合（ステップＳ１７におけるＹｅｓ）、出力部８０は、学習された報酬関数を出力する

（ステップＳ１８）。

[0071] 以上のように、本実施形態では、数理最適化実行部５０が、リプシッツ連続条件を満たすように特徴量が設定された報酬関数の入力を受け付け、熟練者の軌跡の確率分布と、報酬関数のパラメータに基づいて決定される軌跡の確率分布との距離を表わすワッサースタイン距離を最小にする軌跡を推定する。そして、重み更新部６０が、推定された軌跡に基づいて、ボルツマン分布の対数尤度を最大にするように報酬関数のパラメータを更新する。具体的には、重み更新部６０は、ワッサースタイン距離からエントロピー正則化項を減じる式を対数尤度の下限として導出し、導出された対数尤度の下限を最大にするように報酬関数のパラメータを更新する。よって、逆強化学習における不定性の問題を解決しつつ、組み合わせ最適化のような数理最適化問題に対しても適用可能な逆強化学習を行うことができる。

[0072] 例えば、最大エントロピー逆強化学習では、報酬関数が複数存在するという不定性を解決しているが、全ての軌跡を計算可能な状況でなければ、適切な結果を得ることはできない。これに対し、軌跡をサンプリングする方法では、サンプリング分布を設定しなければならないという困難さが残る。最適化問題の一つである組み合わせ最適化は、離散的な値（言い換えると、連続的ではない値）をとるため、ある値を入力したときに、その値に対応する確率を返す確率分布の設定が難しい。組み合わせ最適化問題では、目的関数における値が少しでも変化すると、結果も大きく変化する可能性があるためである。

[0073] 一方、本実施形態の学習装置１００（重み更新部６０）は、最大エントロピー逆強化学習の対数尤度の下限を導出し、これをワッサースタイン距離とエントロピー正則化項に分解する。そして、学習装置１００は、導出された対数尤度の下限を最大にするように報酬関数のパラメータを更新する。よって、逆強化学習における不定性の解決できるとともに、サンプリング分布の設定が不要なため、様々な数理最適化、特に組み合わせ最適化へ適用することが可能になる。

[0074] 例えば、典型的な組み合わせ最適化問題の例として、経路問題やスケジューリング問題、切り出し・詰め込み問題や割り当て・マッチング問題などが挙げられる。具体的には、経路問題は、例えば、運搬経路問題や巡回セールスマン問題などであり、スケジューリング問題は、例えば、ジョブショップ問題や勤務スケジュール問題などである。また、切り出し・詰め込み問題は、例えば、ナップサック問題やビンパッキング問題などであり、割り当て・マッチング問題は、最大マッチング問題や一般化割当問題などである。

[0075] 次に、本実施形態の学習装置１００を利用したロボット制御システムの具体例を説明する。図３は、ロボット制御システムの一実施形態の構成例を示すブロック図である。図３に例示するロボット制御システム２０００は、学習装置１００と、学習データ記憶部２２００と、ロボット２３００とを備えている。

[0076] 図３に例示する学習装置１００は、上記実施形態における学習装置１００と同様である。学習装置１００は、学習の結果作成された報酬関数を後述するロボット２３００の記憶部２３１０に記憶する。

[0077] 学習データ記憶部２２００は、学習装置１００が学習に用いる学習データを記憶する。学習データ記憶部２２００は、例えば、熟練者の意思決定履歴データを記憶していてもよい。

[0078] ロボット２３００は、報酬関数に基づいて動作する機器である。なお、ここでのロボットには、人や動物を模した形状の機器に限られず、自動作業（自動運転、自動制御など）を行う機器なども含まれる。ロボット２３００は、記憶部２３１０と、入力部２３２０と、制御部２３３０とを含む。

[0079] 記憶部２３１０は、学習装置１００によって学習された報酬関数を記憶する。

[0080] 入力部２３２０は、ロボットを動作させる際の状態を示すデータの入力を受け付ける。

[0081] 制御部２３３０は、受け付けた（状態を示す）データ、および、記憶部２３１０に記憶された報酬関数に基づいて、ロボット２３００が行う動作を決

定する。なお、制御部２３３０が、報酬関数に基づいて制御動作を決定する方法は広く知られており、ここでは詳細な説明を省略する。本実施形態では、ロボット２３００のような、自動作業を行う機器を、熟練者の意図を反映した報酬関数に基づいて制御することができる。

[0082] 次に、本実施形態の概要を説明する。図４は、本開示による学習装置の概要を示すブロック図である。本開示による学習装置９０（例えば、学習装置１００）は、リプシッツ連続条件を満たすように特徴量が設定された報酬関数の入力を受け付ける関数入力手段９１（例えば、数理最適化実行部５０）と、熟練者の軌跡の確率分布と、報酬関数のパラメータに基づいて決定される軌跡の確率分布との距離を表わすワッサースタイン距離を最小にする軌跡を推定する推定手段９２（例えば、数理最適化実行部５０）と、推定された軌跡に基づいて、最大エントロピーの原理で導かれるボルツマン分布の対数尤度を最大にするように報酬関数のパラメータを更新する更新手段９３（例えば、重み更新部６０）とを備えている。

[0083] 更新手段９３は、対数尤度の下限として、ワッサースタイン距離から、パラメータに対する最大の報酬値からそのパラメータに対する報酬の平均値を減じた式で定義されるエントロピー正則化項を減じる式を導出し、導出された対数尤度の下限を最大にするように報酬関数のパラメータを更新する。

[0084] そのような構成により、逆強化学習における不定性の問題を解決しつつ、組み合わせ最適化のような数理最適化問題に対しても適用可能な逆強化学習を行うことができる。

[0085] また、更新手段９３は、エントロピー正則化項に、パラメータを更新する処理を繰り返すほど、エントロピー正則化項（例えば、式１０の２つ目の括弧内の式）に対応する部分（例えば、式１２の２つ目の括弧内の式）が対数尤度の下限の最大化に寄与する度合いを減衰させる減衰係数（例えば、β_t）を設定し、設定された対数尤度の下限を最大にするように報酬関数のパラメータを更新してもよい。

[0086] 一方、更新手段９３は、エントロピー正則化項が対数尤度の下限の最大化

に寄与する度合いを減衰させる減衰係数（例えば、β_t）を、そのエントロピー正則化項に対応する部分に設定し、パラメータを更新する処理を繰り返す途中で、エントロピー正則化項に対応する部分が対数尤度の下限の最大化に寄与する度合いを減衰させるように減衰係数を変更（例えば、$\beta_t=1$から$\beta_t=0$への変更、または、$\beta_t=1$から上記式１３に示すβ_tへの変更）してもよい。

[0087] 具体的には、更新手段９３は、対数尤度の移動平均が一定（例えば、移動平均の変化が微小）になったと判断したときに、減衰係数を変更してもよい。

[0088] また、更新手段９３は、ｌｏｇ　ｓｕｍ　ｅｘｐｏｎｅｎｔｉａｌの上限値に基づいて、対数尤度の下限を導出してもよい。

[0089] また、関数入力手段９１は、線形関数になるように特徴量が設定された報酬関数の入力を受け付けてもよい。

[0090] 図５は、少なくとも１つの実施形態に係るコンピュータの構成を示す概略ブロック図である。コンピュータ１０００は、プロセッサ１００１、主記憶装置１００２、補助記憶装置１００３、インタフェース１００４を備える。

[0091] 上述の学習装置９０は、コンピュータ１０００に実装される。そして、上述した各処理部の動作は、プログラム（学習プログラム）の形式で補助記憶装置１００３に記憶されている。プロセッサ１００１は、プログラムを補助記憶装置１００３から読み出して主記憶装置１００２に展開し、当該プログラムに従って上記処理を実行する。

[0092] なお、少なくとも１つの実施形態において、補助記憶装置１００３は、一時的でない有形の媒体の一例である。一時的でない有形の媒体の他の例としては、インタフェース１００４を介して接続される磁気ディスク、光磁気ディスク、ＣＤ－ＲＯＭ（Compact Disc Read-only memory ）、ＤＶＤ－ＲＯＭ（Read-only memory）、半導体メモリ等が挙げられる。また、このプログラムが通信回線によってコンピュータ１０００に配信される場合、配信を受けたコンピュータ１０００が当該プログラムを主記憶装置１００２に展開し

、上記処理を実行してもよい。

[0093] また、当該プログラムは、前述した機能の一部を実現するためのものであっても良い。さらに、当該プログラムは、前述した機能を補助記憶装置１００３に既に記憶されている他のプログラムとの組み合わせで実現するもの、いわゆる差分ファイル（差分プログラム）であってもよい。

[0094] 上記の実施形態の一部又は全部は、以下の付記のようにも記載されうるが、以下には限られない。

[0095] （付記１）リプシッツ連続条件を満たすように特徴量が設定された報酬関数の入力を受け付ける関数入力手段と、

熟練者の軌跡の確率分布と、前記報酬関数のパラメータに基づいて決定される軌跡の確率分布との距離を表わすワッサースタイン距離を最小にする軌跡を推定する推定手段と、

推定された軌跡に基づいて、最大エントロピーの原理で導かれるボルツマン分布の対数尤度を最大にするように前記報酬関数のパラメータを更新する更新手段とを備え、

前記更新手段は、前記対数尤度の下限として、前記ワッサースタイン距離から、前記パラメータに対する最大の報酬値から当該パラメータに対する報酬の平均値を減じた式で定義されるエントロピー正則化項を減じる式を導出し、導出された前記対数尤度の下限を最大にするように前記報酬関数のパラメータを更新する

ことを特徴とする学習装置。

[0096] （付記２）更新手段は、エントロピー正則化項に、パラメータを更新する処理を繰り返すほど、前記エントロピー正則化項に対応する部分が対数尤度の下限の最大化に寄与する度合いを減衰させる減衰係数を設定し、設定された対数尤度の下限を最大にするように報酬関数のパラメータを更新する

付記１記載の学習装置。

[0097] （付記３）更新手段は、エントロピー正則化項が対数尤度の下限の最大化に寄与する度合いを減衰させる減衰係数を当該エントロピー正則化項に対応す

る部分に設定し、パラメータを更新する処理を繰り返す途中で、前記エントロピー正則化項に対応する部分が対数尤度の下限の最大化に寄与する度合いを減衰させるように前記減衰係数を変更する

付記１記載の学習装置。

[0098] （付記４）更新手段は、対数尤度の移動平均が一定になったと判断したときに、減衰係数を変更する

付記３記載の学習装置。

[0099] （付記５）更新手段は、ｌｏｇ　ｓｕｍ　ｅｘｐｏｎｅｎｔｉａｌの上限値に基づいて、対数尤度の下限を導出する

付記１から付記４のうちのいずれか１つに記載の学習装置。

[0100] （付記６）関数入力手段は、線形関数になるように特徴量が設定された報酬関数の入力を受け付ける

付記１から付記５のうちのいずれか１つに記載の学習装置。

[0101] （付記７）コンピュータが、リプシッツ連続条件を満たすように特徴量が設定された報酬関数の入力を受け付け、

前記コンピュータが、熟練者の軌跡の確率分布と、前記報酬関数のパラメータに基づいて決定される軌跡の確率分布との距離を表わすワッサースタイン距離を最小にする軌跡を推定し、

前記コンピュータが、推定された軌跡に基づいて、最大エントロピーの原理で導かれるボルツマン分布の対数尤度を最大にするように前記報酬関数のパラメータを更新し、

前記パラメータの更新の際、前記コンピュータが、前記対数尤度の下限として、前記ワッサースタイン距離から、前記パラメータに対する最大の報酬値から当該パラメータに対する報酬の平均値を減じた式で定義されるエントロピー正則化項を減じる式を導出し、導出された前記対数尤度の下限を最大にするように前記報酬関数のパラメータを更新する

ことを特徴とする学習方法。

[0102] （付記８）コンピュータが、エントロピー正則化項に、パラメータを更新す

る処理を繰り返すほど、前記エントロピー正則化項に対応する部分が対数尤度の下限の最大化に寄与する度合いを減衰させる減衰係数を設定し、設定された対数尤度の下限を最大にするように報酬関数のパラメータを更新する

付記７記載の学習方法。

[0103] （付記９）コンピュータが、エントロピー正則化項が対数尤度の下限の最大化に寄与する度合いを減衰させる減衰係数を当該エントロピー正則化項に対応する部分に設定し、パラメータを更新する処理を繰り返す途中で、前記エントロピー正則化項に対応する部分が対数尤度の下限の最大化に寄与する度合いを減衰させるように前記減衰係数を変更する

付記７記載の学習方法。

[0104] （付記１０）コンピュータに、

リプシッツ連続条件を満たすように特徴量が設定された報酬関数の入力を受け付ける関数入力処理、

熟練者の軌跡の確率分布と、前記報酬関数のパラメータに基づいて決定される軌跡の確率分布との距離を表わすワッサースタイン距離を最小にする軌跡を推定する推定処理、および、

推定された軌跡に基づいて、最大エントロピーの原理で導かれるボルツマン分布の対数尤度を最大にするように前記報酬関数のパラメータを更新する更新処理を実行させ、

前記更新処理で、前記対数尤度の下限として、前記ワッサースタイン距離から、前記パラメータに対する最大の報酬値から当該パラメータに対する報酬の平均値を減じた式で定義されるエントロピー正則化項を減じる式を導出させ、導出された前記対数尤度の下限を最大にするように前記報酬関数のパラメータを更新させる

ための学習プログラムを記憶するプログラム記憶媒体。

[0105] （付記１１）コンピュータに、

更新処理で、エントロピー正則化項に、パラメータを更新する処理を繰り返すほど、前記エントロピー正則化項に対応する部分が対数尤度の下限の最

大化に寄与する度合いを減衰させる減衰係数を設定させ、設定された対数尤度の下限を最大にするように報酬関数のパラメータを更新させる学習プログラムを記憶する

付記１０記載のプログラム記憶媒体。

[0106] （付記１２）コンピュータに、

更新処理で、エントロピー正則化項が対数尤度の下限の最大化に寄与する度合いを減衰させる減衰係数を当該エントロピー正則化項に対応する部分に設定させ、パラメータを更新する処理を繰り返す途中で、前記エントロピー正則化項に対応する部分が対数尤度の下限の最大化に寄与する度合いを減衰させるように前記減衰係数を変更させる学習プログラムを記憶する

付記１０記載のプログラム記憶媒体。

[0107] （付記１３）コンピュータに、

リプシッツ連続条件を満たすように特徴量が設定された報酬関数の入力を受け付ける関数入力処理、

熟練者の軌跡の確率分布と、前記報酬関数のパラメータに基づいて決定される軌跡の確率分布との距離を表わすワッサースタイン距離を最小にする軌跡を推定する推定処理、および、

推定された軌跡に基づいて、最大エントロピーの原理で導かれるボルツマン分布の対数尤度を最大にするように前記報酬関数のパラメータを更新する更新処理を実行させ、

前記更新処理で、前記対数尤度の下限として、前記ワッサースタイン距離から、前記パラメータに対する最大の報酬値から当該パラメータに対する報酬の平均値を減じた式で定義されるエントロピー正則化項を減じる式を導出させ、導出された前記対数尤度の下限を最大にするように前記報酬関数のパラメータを更新させる

ための学習プログラム。

[0108] （付記１４）コンピュータに、

更新処理で、エントロピー正則化項に、パラメータを更新する処理を繰り

返すほど、前記エントロピー正則化項に対応する部分が対数尤度の下限の最大化に寄与する度合いを減衰させる減衰係数を設定させ、設定された対数尤度の下限を最大にするように報酬関数のパラメータを更新させる

付記１３記載の学習プログラム。

[0109] （付記１５）コンピュータに、

更新処理で、エントロピー正則化項が対数尤度の下限の最大化に寄与する度合いを減衰させる減衰係数を当該エントロピー正則化項に対応する部分に設定させ、パラメータを更新する処理を繰り返す途中で、前記エントロピー正則化項に対応する部分が対数尤度の下限の最大化に寄与する度合いを減衰させるように前記減衰係数を変更させる

付記１３記載の学習プログラム。

符号の説明

[0110] １０　記憶部

２０　入力部

３０　特徴量設定部

４０　重み初期値設定部

５０　数理最適化実行部

６０　重み更新部

７０　収束判定部

１００　学習装置

請求の範囲

[請求項1] リプシッツ連続条件を満たすように特徴量が設定された報酬関数の入力を受け付ける関数入力手段と、

熟練者の軌跡の確率分布と、前記報酬関数のパラメータに基づいて決定される軌跡の確率分布との距離を表わすワッサースタイン距離を最小にする軌跡を推定する推定手段と、

推定された軌跡に基づいて、最大エントロピーの原理で導かれるボルツマン分布の対数尤度を最大にするように前記報酬関数のパラメータを更新する更新手段とを備え、

前記更新手段は、前記対数尤度の下限として、前記ワッサースタイン距離から、前記パラメータに対する最大の報酬値から当該パラメータに対する報酬の平均値を減じた式で定義されるエントロピー正則化項を減じる式を導出し、導出された前記対数尤度の下限を最大にするように前記報酬関数のパラメータを更新する

ことを特徴とする学習装置。

[請求項2] 更新手段は、エントロピー正則化項に、パラメータを更新する処理を繰り返すほど、前記エントロピー正則化項に対応する部分が対数尤度の下限の最大化に寄与する度合いを減衰させる減衰係数を設定し、設定された対数尤度の下限を最大にするように報酬関数のパラメータを更新する

請求項１記載の学習装置。

[請求項3] 更新手段は、エントロピー正則化項が対数尤度の下限の最大化に寄与する度合いを減衰させる減衰係数を当該エントロピー正則化項に対応する部分に設定し、パラメータを更新する処理を繰り返す途中で、前記エントロピー正則化項に対応する部分が対数尤度の下限の最大化に寄与する度合いを減衰させるように前記減衰係数を変更する

請求項１記載の学習装置。

[請求項4] 更新手段は、対数尤度の移動平均が一定になったと判断したときに

、減衰係数を変更する

請求項３記載の学習装置。

[請求項5] 更新手段は、ｌｏｇ　ｓｕｍ　ｅｘｐｏｎｅｎｔｉａｌの上限値に基づいて、対数尤度の下限を導出する

請求項１から請求項４のうちのいずれか１項に記載の学習装置。

[請求項6] 関数入力手段は、線形関数になるように特徴量が設定された報酬関数の入力を受け付ける

請求項１から請求項５のうちのいずれか１項に記載の学習装置。

[請求項7] コンピュータが、リプシッツ連続条件を満たすように特徴量が設定された報酬関数の入力を受け付け、

前記コンピュータが、熟練者の軌跡の確率分布と、前記報酬関数のパラメータに基づいて決定される軌跡の確率分布との距離を表わすワッサースタイン距離を最小にする軌跡を推定し、

前記コンピュータが、推定された軌跡に基づいて、最大エントロピーの原理で導かれるボルツマン分布の対数尤度を最大にするように前記報酬関数のパラメータを更新し、

前記パラメータの更新の際、前記コンピュータが、前記対数尤度の下限として、前記ワッサースタイン距離から、前記パラメータに対する最大の報酬値から当該パラメータに対する報酬の平均値を減じた式で定義されるエントロピー正則化項を減じる式を導出し、導出された前記対数尤度の下限を最大にするように前記報酬関数のパラメータを更新する

ことを特徴とする学習方法。

[請求項8] コンピュータが、エントロピー正則化項に、パラメータを更新する処理を繰り返すほど、前記エントロピー正則化項に対応する部分が対数尤度の下限の最大化に寄与する度合いを減衰させる減衰係数を設定し、設定された対数尤度の下限を最大にするように報酬関数のパラメータを更新する

請求項７記載の学習方法。

[請求項9] コンピュータが、エントロピー正則化項が対数尤度の下限の最大化に寄与する度合いを減衰させる減衰係数を当該エントロピー正則化項に対応する部分に設定し、パラメータを更新する処理を繰り返す途中で、前記エントロピー正則化項に対応する部分が対数尤度の下限の最大化に寄与する度合いを減衰させるように前記減衰係数を変更する

請求項７記載の学習方法。

[請求項10] コンピュータに、

リプシッツ連続条件を満たすように特徴量が設定された報酬関数の入力を受け付ける関数入力処理、

熟練者の軌跡の確率分布と、前記報酬関数のパラメータに基づいて決定される軌跡の確率分布との距離を表わすワッサースタイン距離を最小にする軌跡を推定する推定処理、および、

推定された軌跡に基づいて、最大エントロピーの原理で導かれるボルツマン分布の対数尤度を最大にするように前記報酬関数のパラメータを更新する更新処理を実行させ、

前記更新処理で、前記対数尤度の下限として、前記ワッサースタイン距離から、前記パラメータに対する最大の報酬値から当該パラメータに対する報酬の平均値を減じた式で定義されるエントロピー正則化項を減じる式を導出させ、導出された前記対数尤度の下限を最大にするように前記報酬関数のパラメータを更新させる

ための学習プログラムを記憶するプログラム記憶媒体。

[請求項11] コンピュータに、

更新処理で、エントロピー正則化項に、パラメータを更新する処理を繰り返すほど、前記エントロピー正則化項に対応する部分が対数尤度の下限の最大化に寄与する度合いを減衰させる減衰係数を設定させ、設定された対数尤度の下限を最大にするように報酬関数のパラメータを更新させる学習プログラムを記憶する

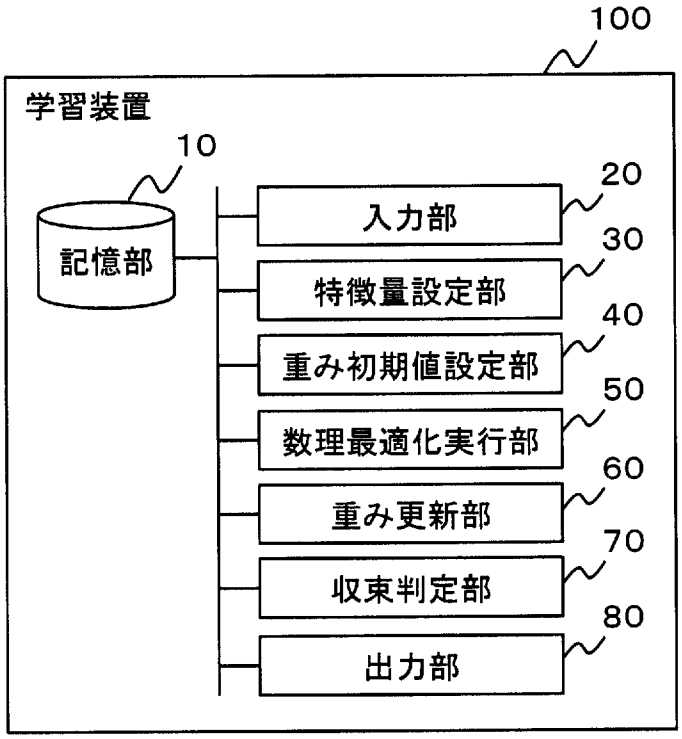
請求項１０記載のプログラム記憶媒体。

[請求項12] コンピュータに、

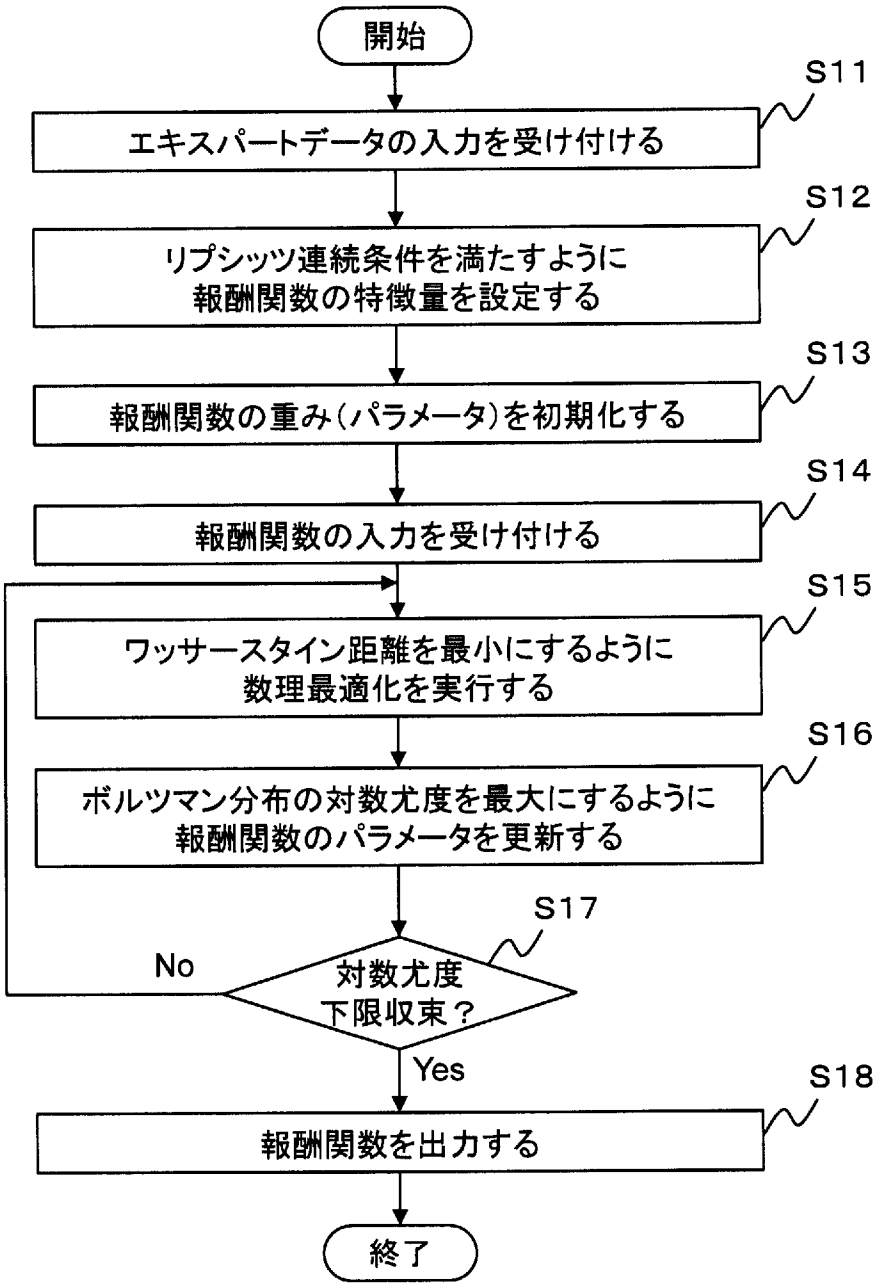
更新処理で、エントロピー正則化項が対数尤度の下限の最大化に寄与する度合いを減衰させる減衰係数を当該エントロピー正則化項に対応する部分に設定させ、パラメータを更新する処理を繰り返す途中で、前記エントロピー正則化項に対応する部分が対数尤度の下限の最大化に寄与する度合いを減衰させるように前記減衰係数を変更させる学習プログラムを記憶する

請求項１０記載のプログラム記憶媒体。

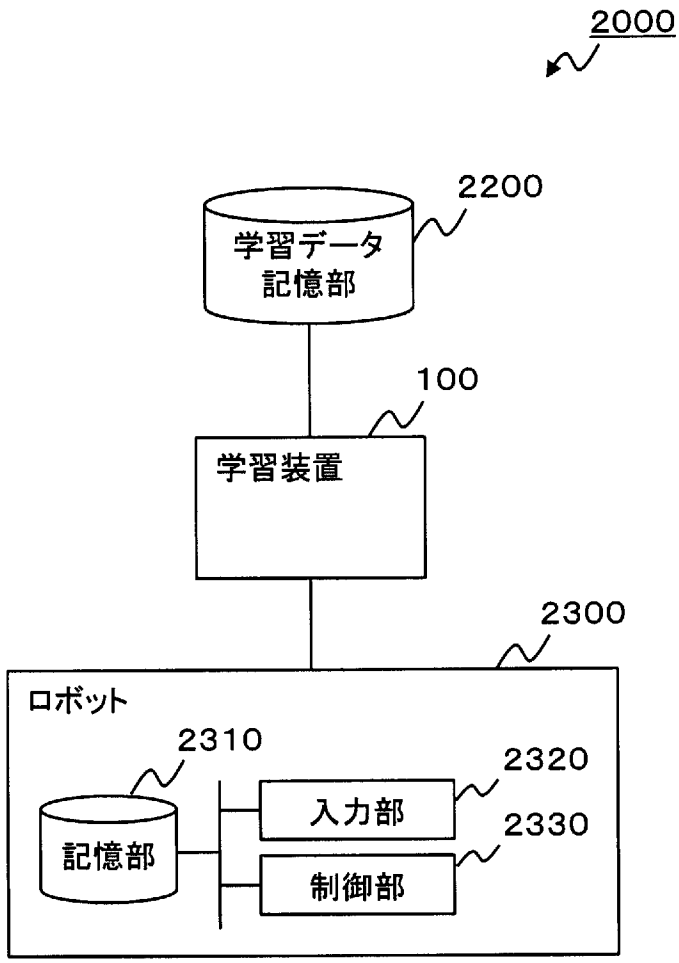
[図1]


100
学習装置
10
記憶部
20
入力部
30
特徴量設定部
40
重み初期値設定部
50
数理最適化実行部
60
重み更新部
70
収束判定部
80
出力部

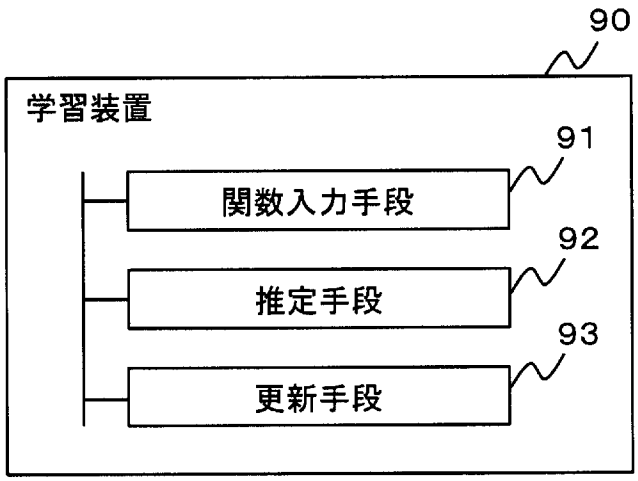
[図2]


開始
S11
エキスパートデータの入力を受け付ける
S12
リプシッツ連続条件を満たすように
報酬関数の特徴量を設定する
S13
報酬関数の重み(パラメータ)を初期化する
S14
報酬関数の入力を受け付ける
S15
ワッサースタイン距離を最小にするように
数理最適化を実行する
S16
ボルツマン分布の対数尤度を最大にするように
報酬関数のパラメータを更新する
S17
No
対数尤度
下限収束？
Yes
S18
報酬関数を出力する
終了

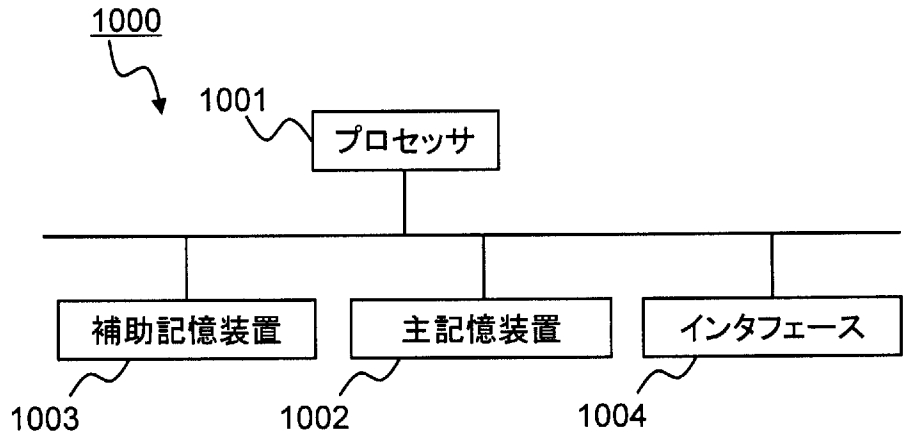
[図3]


2000
2200
学習データ
記憶部
100
学習装置
2300
ロボット
2310
記憶部
2320
入力部
2330
制御部

[図4]


90
学習装置
91
関数入力手段
92
推定手段
93
更新手段

[図5]



INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2021/016630

A. CLASSIFICATION OF SUBJECT MATTER

G06N 20/00(2019.01)i
FI: G06N20/00

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06N20/00

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Published examined utility model applications of Japan 1922-1996
Published unexamined utility model applications of Japan 1971-2021
Registered utility model specifications of Japan 1996-2021
Published registered utility model applications of Japan 1994-2021

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	XIAO, Huang et al. Wasserstein Adversarial Imitation Learning, [online]. arXiv. 19 June 2019, pp. 1-18, [retrieval date 24 June 2021], Internet: <URL:https://arxiv.org/pdf/1906.08113.pdf> entire text	1-12
A	GANGWANI, Tanmay et al. STATE-ONLY IMITATION WITH TRANSITION DYNAMICS MISMATCH, [online]. arXiv. 27 February 2020, pp. 1-17, [retrieval date 24 June 2021], Internet: <URL:https://arxiv.org/pdf/2002.11879.pdf> entire text	1-12
A	ZHANG, Ming et al. Wasserstein Distance guided Adversarial Imitation Learning with Reward Shape Exploration, [online]. arXiv. 08 December 2020, [retrieval date 24 June 2021], Internet: <URL:https://arxiv.org/pdf/2006.03503.pdf> entire text	1-12

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
24 June 2021	**06 July 2021**
Name and mailing address of the ISA/JP **Japan Patent Office (ISA/JP)** **3-4-3 Kasumigaseki, Chiyoda-ku, Tokyo 100-8915** **Japan**	Authorized officer Telephone No.

国際調査報告

国際出願番号
PCT/JP2021/016630

A. 発明の属する分野の分類（国際特許分類（ＩＰＣ））

G06N 20/00(2019.01)i
FI: G06N20/00

B. 調査を行った分野

調査を行った最小限資料（国際特許分類（ＩＰＣ））

G06N20/00

最小限資料以外の資料で調査を行った分野に含まれるもの

日本国実用新案公報	１９２２－１９９６年
日本国公開実用新案公報	１９７１－２０２１年
日本国実用新案登録公報	１９９６－２０２１年
日本国登録実用新案公報	１９９４－２０２１年

国際調査で使用した電子データベース（データベースの名称、調査に使用した用語）

C. 関連すると認められる文献

引用文献のカテゴリー*	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求項の番号
A	XIAO, Huang ほか, Wasserstein Adversarial Imitation Learning, [online], arXiv, 2019.06.19, pp.1-18, [検索日 2021.06.24], インターネット：＜ＵＲＬ：https://arxiv.org/pdf/1906.08113.pdf＞ 全文	1-12
A	GANGWANI, Tanmay ほか, STATE-ONLY IMITATION WITH TRANSITION DYNAMICS MISMATCH, [online], arXiv, 2020.02.27, pp.1-17, [検索日 2021.06.24], インターネット：＜ＵＲＬ：https://arxiv.org/pdf/2002.11879.pdf＞ 全文	1-12
A	ZHANG, Ming ほか, Wasserstein Distance guided Adversarial Imitation Learning with Reward Shape Exploration, [online], arXiv, 2020.12.08, [検索日 2021.06.24], インターネット：＜ＵＲＬ：https://arxiv.org/pdf/2006.03503.pdf＞ 全文	1-12

☐ Ｃ欄の続きにも文献が列挙されている。

☐ パテントファミリーに関する別紙を参照。

＊ 引用文献のカテゴリー
“A” 特に関連のある文献ではなく、一般的技術水準を示すもの
“E” 国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの
“L” 優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）
“O” 口頭による開示、使用、展示等に言及する文献
“P” 国際出願日前で、かつ優先権の主張の基礎となる出願の日の後に公表された文献
“T” 国際出願日又は優先日後に公表された文献であって出願と抵触するものではなく、発明の原理又は理論の理解のために引用するもの
“X” 特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの
“Y” 特に関連のある文献であって、当該文献と他の１以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの
“&” 同一パテントファミリー文献

国際調査を完了した日	国際調査報告の発送日
24.06.2021	06.07.2021

名称及びあて先

日本国特許庁(ISA/JP)
〒100-8915
日本国
東京都千代田区霞が関三丁目４番３号

権限のある職員（特許庁審査官）

多胡　滋 5B 3562

電話番号 03-3581-1101 内線 3545