【公報種別】特許法第１７条の２の規定による補正の掲載
【部門区分】第６部門第３区分
【発行日】令和5年4月6日(2023.4.6)

【公開番号】特開2021-60988(P2021-60988A)
【公開日】令和3年4月15日(2021.4.15)
【年通号数】公開・登録公報2021-018
【出願番号】特願2020-159841(P2020-159841)
【国際特許分類】
Ｇ０６Ｎ 20/00(2019.01)
Ｇ０６Ｎ 99/00(2019.01)
Ｇ０６Ｎ 3/08(2023.01)
Ｇ０５Ｂ 13/02(2006.01)
【ＦＩ】
Ｇ０６Ｎ 20/00
Ｇ０６Ｎ 99/00 １８０
Ｇ０６Ｎ 3/08
Ｇ０５Ｂ 13/02 Ｌ

【手続補正書】
【提出日】令和5年3月29日(2023.3.29)
【手続補正１】
【補正対象書類名】特許請求の範囲
【補正対象項目名】全文
【補正方法】変更
【補正の内容】
【特許請求の範囲】
【請求項１】
システムを制御する制御ポリシーを最適化するための、コンピュータで実現される学習方法であって、
前記システムに設けられたセンサに接続された入出力インターフェースを介して、ポリシー最適化方法を用いて特定のタスクが学習されるよう動作中のシステムの状態を受信することを備え、
前記システムの状態は、前記センサによって測定され、
前記方法は、
前記制御ポリシーを、ニューラルネットワークを含む関数近似器として初期化することと、
現在の制御ポリシーを用いて、状態、行動、および次の状態のタプルのデータを収集することと、
前記現在の制御ポリシーに基づいて、利点関数および状態訪問頻度を推定することと、
Kullback-Leiblerダイバージェンス制約（ＫＬダイバージェンス制約）および代理目的関数をポリシーパラメータの関数として推定することと、
準ニュートン信頼領域ポリシー最適化（ＱＮＴＰＲＯ）を用いて、前記推定された制約および前記代理目的関数に基づいて、前記現在の制御ポリシーを更新することと、
前記システムを制御するために、前記更新された現在の制御ポリシーを用いて蓄積された予想される平均報酬に基づいて、最適な制御ポリシーを決定することと、
前記最適な制御ポリシーに基づいて制御コマンドを生成することと、
前記制御コマンドの制御信号を前記システムへ送信することによって、前記最適な制御ポリシーに従って前記システムを動作させることとをさらに備える、方法。
【請求項２】

前記収集すること、前記推定すること、および前記更新することは、前記ポリシーの異なるエピソードからの前記平均報酬の値が定常状態に達し、未知の値に収束するまで、反復的に実行される、請求項１に記載の方法。
【請求項３】
利点関数Ａπは、状態－行動価値関数Ｑπおよび状態価値関数Ｖπによって表される、請求項１に記載の方法。
【請求項４】
前記利点関数は、
【数１】

$$A_\pi(s,a) = Q_\pi(s,a) - V_\pi(s)$$

によって表され、式中、sは前記システムの状態であり、aは行動である、請求項３に記載の方法。
【請求項５】
目的関数のヘッシアンを推定するためにＢＦＧＳ準ニュートン法が用いられる、請求項１に記載の方法。
【請求項６】
大規模問題に対して目的関数のヘッシアンの推定値を近似的に保つためにＬ－ＢＦＧＳ準ニュートン法が用いられる、請求項１に記載の方法。
【請求項７】
ＱＮＴＰＲＯは、エピソードのための目的関数を最大化するよう、ポリシーパラメータθｉを取得する、請求項１に記載の方法。
【請求項８】
ＱＮＴＰＲＯは、最適なステップ方向およびサイズを計算するためにDogleg法を用いる、請求項１に記載の方法。
【請求項９】
ＱＮＴＲＰＯは、信頼領域法を用いて、前記目的関数の二次近似を用いて反復態様で前記Dogleg法により計算されたステップを受け入れるかまたは拒否する、請求項１に記載の方法。
【請求項１０】
制御ポリシーを最適化することによってシステムを制御するためのコントローラであって、
前記システムの設けられたセンサを介して前記システムの行動および状態を受信するように構成されたインターフェースと、
ポリシー初期化器、ポリシー収集器または記憶部、推定器、エージェントおよびポリシー更新プログラム、目的関数のヘッシアンのための準ニュートン近似プログラム、最適化ステップを計算するためのDogleg法、ならびに前記目的関数のヘッシアン近似を用いてポリシーパラメータの次の推定を見つけるための信頼領域法を含むコンピュータ実行可能プログラムを記憶するメモリと、
プロセッサとを備え、前記プロセッサは、前記メモリに関連して、
前記制御ポリシーを、ニューラルネットワークを含む関数近似器として初期化し、
現在の制御ポリシーを用いて、前記状態に関してデータを収集し、
前記現在の制御ポリシーに基づいて、利点関数および状態訪問頻度を推定し、
準ニュートン信頼領域ポリシー最適化（ＱＮＴＰＲＯ）を用いて、前記収集されたデータに基づいて、前記現在の制御ポリシーを更新し、
前記システムを制御するために、最適な制御ポリシーを、前記更新された現在の制御ポリシーを用いて蓄積された平均報酬の値に基づいて決定し、
前記最適な制御ポリシーに基づいて制御コマンドを生成し、
前記制御コマンドの制御信号を前記システムへ送信することによって、前記最適な制御ポ

リシーに従って前記システムを動作させるよう構成される、コントローラ。
【請求項１１】
前記データ収集、前記推定、および前記更新は、前記ポリシーのエピソードについての前記平均報酬の値が未知の値において定常状態に達するまで反復的に実行される、請求項１０に記載のコントローラ。
【請求項１２】
利点関数Ａπは、状態-行動価値関数Ｑπおよび状態価値関数Ｖπによって表される、請求項１０に記載のコントローラ。
【請求項１３】
前記利点関数は、
【数２】

$$A_\pi(s,a) = Q_\pi(s,a) - V_\pi(s)$$

によって表され、式中、sは前記システムの状態であり、aは行動（または制御信号）である、請求項１０に記載のコントローラ。
【請求項１４】
ポリシー勾配最適化の目的関数のヘッシアンを推定するために、ＢＦＧＳ準ニュートン法を用いる、請求項１０に記載のコントローラ。
【請求項１５】
大規模問題に対して目的関数のヘッシアンの推定値を近似的に保つためにＬ－ＢＦＧＳ準ニュートン法が用いられる、請求項１０に記載のコントローラ。
【請求項１６】
ＱＮＴＰＲＯは、エピソードのための目的関数を最大化するよう、ポリシーパラメータθｉを取得する、請求項１０に記載のコントローラ。
【請求項１７】
ＱＮＴＰＲＯは、最適なステップ方向およびサイズを計算するためにDogleg法を用いる、請求項１０に記載のコントローラ。
【請求項１８】
ＱＮＴＰＲＯは、信頼領域法を用いて、前記目的関数の二次近似を用いて反復態様で前記Dogleg法により計算されたステップを受け入れるかまたは拒否する、請求項１０に記載のコントローラ。
【請求項１９】
制御ポリシーを最適化することによってシステムを制御するためのコントローラであって、
前記システムに設けられたセンサを介して前記システムの行動および状態を受信するように構成されたインターフェースと、
ポリシー初期化器、ポリシー収集器または記憶部、推定器、エージェントおよびポリシー更新プログラム、目的関数のヘッシアンのための制限付きメモリ準ニュートン近似プログラム、最適化ステップを計算するためのDogleg法、ならびに前記目的関数のヘッシアン近似を用いてポリシーパラメータの次の推定を見つけるための信頼領域法を含むコンピュータ実行可能プログラムを記憶するメモリと、
プロセッサとを備え、前記プロセッサは、前記メモリに関連して、
前記制御ポリシーを、ニューラルネットワークを含む関数近似器として初期化し、
現在の制御ポリシーを用いて、前記状態に関してデータを収集し、
前記現在の制御ポリシーに基づいて、利点関数および状態訪問頻度を推定し、
準ニュートン信頼領域ポリシー最適化（ＱＮＴＰＲＯ）を用いて、前記収集されたデータに基づいて、前記現在の制御ポリシーを更新し、
前記システムを制御するために、最適な制御ポリシーを、前記更新された現在の制御ポリシーを用いて蓄積された平均報酬の値に基づいて決定し、

前記最適な制御ポリシーに基づいて制御コマンドを生成し、
前記制御コマンドの制御信号を前記システムへ送信することによって、前記最適な制御ポリシーに従って前記システムを動作させるよう構成される、コントローラ。
【請求項２０】
前記データ収集、前記推定、および前記更新は、前記ポリシーのエピソードについての前記平均報酬の値が未知の値において定常状態に達するまで反復的に実行される、請求項１９に記載のコントローラ。
【請求項２１】
利点関数Aπは、状態-行動価値関数Qπおよび状態価値関数Vπによって表される、請求項１９に記載のコントローラ。
【請求項２２】
前記利点関数は、
【数３】

$$A_\pi(s,a) = Q_\pi(s,a) - V_\pi(s)$$

によって表され、式中、sは前記システムの状態であり、aは行動(または制御信号)である、請求項１９に記載のコントローラ。
【請求項２３】
ポリシー勾配最適化の目的関数のヘッシアンを推定するために、ＢＦＧＳ準ニュートン法を用いる、請求項１９に記載のコントローラ。
【請求項２４】
大規模問題に対して目的関数のヘッシアンの推定値を近似的に保つためにＬ－ＢＦＧＳ準ニュートン法が用いられる、請求項１９に記載のコントローラ。
【請求項２５】
ＱＮＴＰＲＯは、エピソードのために目的関数を最大化するよう、ポリシーパラメータθｉを取得する、請求項１９に記載のコントローラ。
【請求項２６】
ＱＮＴＰＲＯは、最適なステップ方向およびサイズを計算するためにDogleg法を用いる、請求項２５に記載のコントローラ。
【請求項２７】
ＱＮＴＰＲＯは、信頼領域法を用いて、前記目的関数の二次近似を用いて反復態様で前記Dogleg法により計算されたステップを受け入れるかまたは拒否する、請求項１９に記載のコントローラ。