(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第6856575号
(P6856575)
(45)発行日　令和3年4月7日(2021.4.7)　(24)登録日　令和3年3月22日(2021.3.22)

(51)Int.Cl.		F I	
B60W 50/00	(2006.01)	B60W	50/00
B60W 30/00	(2006.01)	B60W	30/00
B60W 30/10	(2006.01)	B60W	30/10
G06N 99/00	(2019.01)	G06N	99/00

請求項の数 13　外国語出願　(全 33 頁)

(21)出願番号　特願2018-91189(P2018-91189)
(22)出願日　平成30年5月10日(2018.5.10)
(65)公開番号　特開2019-31268(P2019-31268A)
(43)公開日　平成31年2月28日(2019.2.28)
審査請求日　令和2年9月3日(2020.9.3)
(31)優先権主張番号　15/594,020
(32)優先日　平成29年5月12日(2017.5.12)
(33)優先権主張国・地域又は機関
米国(US)

早期審査対象出願

(73)特許権者　507342261
トヨタ　モーター　エンジニアリング　アンド　マニュファクチャリング　ノース　アメリカ，インコーポレイティド
アメリカ合衆国、７５０２４　テキサス州、プレイノ、ダブリュ１－３シー・ヘッド　クォーターズ・ドライブ、６５６５
(74)代理人　100099759
弁理士　青木　篤
(74)代理人　100123582
弁理士　三橋　真二
(74)代理人　100092624
弁理士　鶴田　準一
(74)代理人　100147555
弁理士　伊藤　公一

最終頁に続く

(54)【発明の名称】能動的探索なしの強化学習に基づく制御ポリシー学習及び車両制御方法

(57)【特許請求の範囲】
【請求項１】
車両の操作を行なう目的で車両を自律的に制御するコンピュータ実装型の方法において、
最低予想累積コストで前記車両の操作を実施すべく制御ポリシーによって前記車両の制御が可能になるように既存の制御ポリシーを適応させるために、前記車両の操作に関連する受動的に収集されたデータに対して、ｐａｓｓｉｖｅ　ａｃｔｏｒ－ｃｒｉｔｉｃ（ｐＡＣ）強化学習方法を適用するステップと、
前記車両の操作を行なうべく前記制御ポリシーにしたがって前記車両を制御するステップと、を含み、
受動的に収集されたデータに対してｐａｓｓｉｖｅ　ａｃｔｏｒ－ｃｒｉｔｉｃ（ｐＡＣ）強化学習方法を適用する前記ステップが、
ａ）　ｃｒｉｔｉｃネットワークにおいて、前記受動的に収集されたデータのサンプルを用いて最適な制御ポリシーの下でＺ値及び平均コストを推定するステップと、
ｂ）　ｃｒｉｔｉｃネットワークに作用的に連結されたａｃｔｏｒネットワークにおいて、前記ｃｒｉｔｉｃネットワークからの最適な制御ポリシーの下で、前記受動的に収集されたデータのサンプル、前記推定されたＺ値、及び前記推定された平均コストを用いて前記制御ポリシーを修正するステップと、
ｃ）　前記推定された平均コストが収束するまで、ステップ（ａ）～（ｂ）を反復的に繰返すステップと、を含む、方法。

【請求項２】
前記車両の操作が、車線内を走行する第２の車両と第３の車両の間で前記車線内に前記車両を合流させる操作であり、前記制御ポリシーが、前記第２の車両と前記第３の車両の間の中間に前記車両を合流させるべく前記車両を制御するように構成されている、請求項１に記載の方法。
【請求項３】
前記Ｚ値がベルマン方程式の線形化版を用いて推定される、請求項１に記載の方法。
【請求項４】
最適なポリシーの下で前記平均コストを推定する前記ステップが、前記制御ポリシーを修正する前記ステップの前に、前記平均コストを更新するステップを含む、請求項１に記載の方法。
【請求項５】
Ｚ値を推定する前記ステップが、
重み付けされた放射基底関数の線形結合を用いてＺ値関数を近似するステップと、
近似されたＺ値関数及び前記受動的に収集されたデータのサンプルを用いてＺ値を近似するステップと、を含む、請求項１に記載の方法。
【請求項６】
重み付けされた放射基底関数の線形結合を用いてＺ値関数を近似する前記ステップが、前記重み付けされた放射基底関数内で使用される重みを最適化するステップを含む、請求項５に記載の方法。
【請求項７】
重み付けされた放射基底関数の線形結合を用いてＺ値関数を近似する前記ステップが、前記重みを最適化する前記ステップの前に、前記重み付けされた放射基底関数内で使用される重みを更新するステップを含む、請求項６に記載の方法。
【請求項８】
前記制御ポリシーを修正する前記ステップが、
制御ゲインを近似するステップと、
前記制御ゲインを最適化して、最適化された制御ゲインを提供するステップと、
前記最適化された制御ゲインを用いて前記制御ポリシーを修正するステップと、を含む、請求項１に記載の方法。
【請求項９】
前記制御ゲインを最適化する前に、
制御入力を決定するステップと、
前記制御入力、前記受動的に収集されたデータのサンプル及び前記近似された制御ゲインを用いて、行動価値関数の値を決定するステップと、をさらに含む、請求項８に記載の方法。
【請求項１０】
制御ゲインを近似する前記ステップが、重み付けされた放射基底関数の線形結合を用いて前記制御ゲインを近似するステップを含む、請求項８に記載の方法。
【請求項１１】
重み付けされた放射基底関数の線形結合を用いて前記制御ゲインを近似する前記ステップの前に、前記重み付けされた放射基底関数内で使用される重みを更新するステップをさらに含む、請求項１０に記載の方法。
【請求項１２】
操作を行なうようシステムを制御するのに使用可能な制御ポリシーを最適化するコンピュータ実装型方法であって、
前記システムを制御するのに使用可能な制御ポリシーを提供するステップと、
行なうべき操作に関する受動的に収集されたデータに対してｐａｓｓｉｖｅ　ａｃｔｏｒ－ｃｒｉｔｉｃ（ｐＡＣ）強化学習方法を適用して、最低予想累積コストで前記操作を行なうように前記システムを制御すべく前記制御ポリシーが操作可能になるように前記制

御ポリシーを修正するステップと、を含み、
受動的に収集されたデータに対してｐａｓｓｉｖｅ　ａｃｔｏｒ－ｃｒｉｔｉｃ（ｐＡＣ）強化学習方法を適用する前記ステップが、
ａ）ｃｒｉｔｉｃネットワークにおいて、前記受動的に収集されたデータのサンプルを用いてＺ値を推定し、前記受動的に収集されたデータのサンプルを用いて最適なポリシーの下で平均コストを推定するステップと、
ｂ）ａｃｔｏｒネットワークにおいて、前記受動的に収集されたデータのサンプル、前記システムについての制御ダイナミクス、到達コスト及び制御ゲインを用いて前記制御ポリシーを修正するステップと、
ｃ）前記制御ポリシーを修正するのに使用されるパラメータ及び最適なポリシーの下で前記Ｚ値及び前記平均コストを推定するのに使用されるパラメータを更新するステップと、
ｄ）前記推定された平均コストが収束するまで、ステップ（ａ）～（ｃ）を反復的に繰返すステップと、を含む、方法。
【請求項１３】
車両の操作を行なうべく車両を自律的に制御するのに使用可能な制御ポリシーを最適化するように構成されたコンピュータ処理システムであって、
当該コンピュータ処理システムが、前記コンピュータ処理システムの操作を制御するための１つ以上のプロセッサと、前記１つ以上のプロセッサにより使用可能なデータ及びプログラム命令を記憶するためのメモリとを含み、
前記メモリは、コンピュータコードを記憶するように構成され、該コンピュータコードは、前記１つ以上のプロセッサによって実行された時点で、前記１つ以上のプロセッサに、
ａ）前記車両の操作に関する受動的に収集されたデータを受信させ、
ｂ）前記車両についての到達コストを推定するのに使用可能なＺ値関数を決定させ、
ｃ）前記コンピュータ処理システム内のｃｒｉｔｉｃネットワークにおいて、
ｃ１）前記Ｚ値関数及び前記受動的に収集されたデータのサンプルを使用してＺ値を決定させ、
ｃ２）前記受動的に収集されたデータのサンプルを用いて最適なポリシーの下で平均コストを推定させ、
ｄ）前記コンピュータ処理システム内のａｃｔｏｒネットワークにおいて、前記受動的に収集されたデータ、前記車両についての制御ダイナミクス、到達コスト及び制御ゲインを用いて前記制御ポリシーを修正させ、
ｅ）前記推定された平均コストが収束するまで、ステップ（ｃ）及び（ｄ）を反復的に繰返させる、コンピュータ処理システム。
【発明の詳細な説明】
【技術分野】
【０００１】
関連出願の相互参照
本出願は、２０１６年７月８日出願の米国特許出願第１５／２０５，５５８号の一部継続出願であり、その利益を主張するものである。
【０００２】
本発明は、車両を自律的に制御する方法に関し、より詳細には車両の操作を自律的に制御するために使用可能な制御ポリシー（control policy）を修正及び／又は最適化するための強化学習方法に関する。
【背景技術】
【０００３】
一定のタイプのシステムにおいては、環境を能動的に探索することにより、最適なシステム制御ポリシーを決定するために、モデルフリー強化学習（ＲＬ）技術を利用することができる。しかしながら、車両が採用し得るあらゆる活動の広範な能動的探索に付随する

潜在的にマイナスの帰結に起因して、車両の自律的制御のために使用可能な制御ポリシーに対して従来のＲＬアプローチを適用することは困難であり得る。さらに、車両の安全性の確保を支援するのに必要とされる形で能動的探索を行なうことによって、高い計算コストが必要となる可能性がある。代替案としてのモデルベースのＲＬ技術の使用には、車両が作動する環境の正確なシステムダイナミクスモデルが必要になり得る。しかしながら、自律的車両が作動する複雑な環境は、正確にモデリングすることが非常に困難なものであり得る。

【発明の概要】

【０００４】

本明細書中に記載の実施形態の一態様においては、車両の操作を行なう目的で車両を自律的に制御するためのコンピュータ実装型方法が提供されている。該方法は、最低予想累積コストで車両の操作を実施する目的で車両を制御するように構成された制御ポリシーを学習するために、車両の操作に関連する受動的に収集されたデータに対して、受動的ａｃｔｏｒ－ｃｒｉｔｉｃ強化学習方法を適用するステップと；車両の操作を行なうために制御ポリシーにしたがって車両を制御するステップと；を含む。

【０００５】

本明細書中に記載の実施形態の別態様においては、操作を行なうようシステムを制御するために使用可能な制御ポリシーを最適化するためのコンピュータ実装型方法が提供されている。該方法は、システムを制御するために使用可能な制御ポリシーを提供するステップと、行なうべき操作に関する受動的に収集されたデータに対して受動的ａｃｔｏｒ－ｃｒｉｔｉｃ強化学習方法を適用して、最低予想累積コストで操作を行なうようにシステムを制御するために制御ポリシーが操作可能になるような形で制御ポリシーを修正するステップと、を含む。

【０００６】

本明細書中に記載の実施形態の別態様においては、車両の操作を行なう目的で車両を自律的に制御するために使用可能な制御ポリシーを最適化するように構成されたコンピュータ処理システムが提供されている。このコンピュータ処理システムは、コンピュータ処理システムの操作を制御するための１つ以上のプロセッサと、１つ以上のプロセッサにより使用可能なデータ及びプログラム命令を記憶するためのメモリとを含む。メモリは、１つ以上のプロセッサによって実行された時点で、１つ以上のプロセッサに、ａ）車両の操作に関わる受動的に収集されたデータを受信させ；ｂ）到達コストを推定するために使用可能なＺ値関数を決定させ；ｃ）コンピュータ処理システム内のｃｒｉｔｉｃネットワークにおいて：Ｚ値関数及び受動的に収集されたデータのサンプルを使用してＺ値を決定させ；受動的に収集されたデータのサンプルを用いて最適なポリシー下で平均コストを推定させ；ｄ）コンピュータ処理システム内のａｃｔｏｒネットワークにおいて、受動的に収集されたデータ、システムについての制御ダイナミクス、到達コスト及び制御ゲインを用いて制御ポリシーを修正させ；ｅ）推定平均コストが収束するまで、ステップ（ｃ）及び（ｄ）を反復的に繰返させる；コンピュータコードを記憶するように構成されている。

【図面の簡単な説明】

【０００７】

【図１】本明細書中に記述された実施形態に係る、（例えば自律車両などの）システムに対する制御入力を決定すべく且つシステム制御ポリシーを修正及び／又は最適化すべく構成されたコンピュータ処理システムのブロック図である。

【図２】本明細書中に記述された方法に係る、車両制御入力の決定、及び／又は、制御ポリシーの修正若しくは最適化の間における情報の流れを示す概略図である。

【図３】本明細書中に記述された実施形態に係る、一つ以上の制御入力と制御ポリシーとを使用する自律的制御に向けて構成された車両であって、当該車両に対する制御入力を決定すべく且つ自律車両操作制御ポリシーを修正及び／又は最適化すべく構成されたコンピュータ処理システムが組み込まれた車両の概略的ブロック図である。

【図４】本明細書中に記述された実施形態に係る方法を用いる、高速道路合流用の制御ポ

リシーの最適化の例において採用された車両の構成の概略図である。
【図５】図４に示された車両の構成に関して実施される最適化のグラフ表示である。
【図６】車両を制御するように構成された制御ポリシーを学習するために受動的ａｃｔｏｒ－ｃｒｉｔｉｃ強化学習方法を適用し、学習した制御ポリシーを用いて車両を制御するための方法の実装を例示するフローチャートである。
【図７】本明細書中に記載の実施形態に係る受動的ａｃｔｏｒ－ｃｒｉｔｉｃ（ＰＡＣ）強化学習方法の適用を例示するフローチャートである。
【図８】図７のブロック８２０に示されているように受動的に収集されたデータのサンプルを用いて最適な制御ポリシー下でＺ値及び平均コストを推定するための、ｃｒｉｔｉｃネットワークによる受動的ａｃｔｏｒ－ｃｒｉｔｉｃ（ＰＡＣ）強化学習方法のステップの適用を例示するフローチャートである。
【図９】図７のブロック８３０に示されているように受動的に収集されたデータのサンプルを用いて最適な制御ポリシー下でＺ値及び平均コストを推定するための、ａｃｔｏｒネットワークによる受動的ａｃｔｏｒ－ｃｒｉｔｉｃ（ＰＡＣ）強化学習方法のステップの適用を例示するフローチャートである。
【発明を実施するための形態】
【０００８】
本明細書中に記載の実施形態は、最低予想累積コストで車両の操作を行なうように車両を自律的に制御するために構成された制御ポリシーを学習することを目的として、車両の操作に関連する受動的に収集されたデータに対して受動的ａｃｔｏｒ－ｃｒｉｔｉｃ（ｐＡＣ）強化学習方法を適用するためのコンピュータ実装型の方法に関する。このとき、車両は、車両の操作を行なうための制御ポリシーにしたがってコンピュータ処理システムによって制御され得る。ｐＡＣ方法は、制御ポリシーを学習するために、合流操作中に車両が作動している環境の正確なシステムダイナミクスモデルを必要としない。ｐＡＣ方法は同様に、（例えば、行動を行ない且つその行動の結果を監視して制御ポリシーを決定し変更することを伴い得る）制御ポリシーを学習するための環境の能動的探索を使用しない。本明細書中に記載のｐＡＣ方法は、能動的探索の代りに、制御されている車両の、受動的に収集されたデータ、部分的に公知のシステムダイナミクスモデル及び公知の制御ダイナミクスモデルを使用する。特定の実施形態において、ｐＡＣ方法は、１車線内を走行する第２の車両と第３の車両の間においてこの車線内に車両を合流させるように車両を制御するために使用可能な制御ポリシーを学習する目的で使用され得る。
【０００９】
本開示に関連して、「オンライン」とは、コンピュータ処理システムが学習し得ると共に、actor及びcriticのネットワークパラメータが、上記システムが作動するにつれて（例えば車両が移動するなどにつれて）、コンピュータ処理され且つ更新され得ることを意味する。オンラインのソリューションを用いてactorパラメータ及びcriticパラメータを決定かつ更新すると、車両及びシステムのダイナミクス（dynamics）の変更が許容され得る。同様に、自律的操作とは、自律的に実施される操作である。
【００１０】
図１は、本明細書中に開示される種々の実施形態に係る方法を実現すべく構成されたコンピュータ処理システム１４のブロック図である。更に詳細には、少なくとも一つの実施形態において、コンピュータ処理システム１４は、本明細書中に記述された方法に従い、制御入力を決定すべく構成され得る。また、コンピュータ処理システムは、システム（例えば、自律車両）を制御して特定の操作若しくは機能を自律的に実施すべく使用可能な制御ポリシーを修正及び／又は最適化するようにも構成され得る。
【００１１】
最適な又は最適化された制御ポリシーは、最低予想累積コストで車両の操作を行なう目的で車両を制御するように構成された制御ポリシーであり得る。最適な制御ポリシーは、車両の操作に関連する受動的に収集されたデータに対して受動的ａｃｔｏｒ－ｃｒｉｔｉｃ（ｐＡＣ）強化学習方法を適用することにより初期制御ポリシーを修正することを通し

て学習され得る。ｐＡＣ強化学習方法は、制御ポリシーのパラメータ値を反復的に最適化するために初期制御ポリシーに対し適用され得る。初期制御ポリシーのパラメータはランダム値に初期化され得る。１つ以上の配置において、最適な制御ポリシーは、このポリシーに付随する平均コストが収束したときに学習されたものとみなされる。平均コストは、平均コストの値がｐＡＣ方法の予め定められた回数の反復について予め定められた範囲又は許容誤差ゾーン外へ変動しない場合に、収束したものとみなすことができる。例えば、図５に例示された実施形態において、平均コストは、２００００回の反復後、約０．３の値を達成している。２００００回の反復後の予め定められた回数の反復について平均コストがいずれの方向にも０．３から一定値を超えて変動しない場合、制御ポリシーは最適化されたものと考えられてよい。そのとき、車両は、車両の操作を行なうために、最適化された制御ポリシーにしたがって制御され得る。車両の操作を行なう目的で車両を制御するための最適化された制御ポリシーの使用は、このとき、最低予想累積コストでの車両の操作の実施を結果としてもたらすはずである。

【００１２】

少なくとも一つの実施形態において、コンピュータ処理システムは、車両に組み込まれ得ると共に、車両の操作の制御に向けられた制御ポリシーを修正及び最適化するように構成され得る。制御ポリシーを修正及び／又は最適化するためにコンピュータ処理システムにより必要とされる情報（例えば、データ、命令、及び／又は他の情報）は、任意の適切な手段から、例えば車両センサから又は無線接続を介して遠隔データベースのような車外情報源から、受信され且つ／又はそれにより収集され得る。幾つかの実施形態においては、制御ポリシーを修正及び／又は最適化するためにコンピュータ処理システムにより必要とされる情報（例えば、データ）の少なくとも幾つかは、車両の操作の前に（例えば、メモリ内に記憶されたデータ及び他の情報として）コンピュータ処理システムに提供され得る。また、コンピュータ処理システムは、修正若しくは最適化された制御ポリシーに従って車両を制御することで、関連する自律的操作を実施するようにも構成され得る。

【００１３】

少なくとも一つの実施形態において、コンピュータ処理システムは、（例えばスタンドアロンのコンピュータ処理システムとして）車両から遠隔的に配置され得ると共に、制御ポリシーを車両から遠隔的に修正及び／又は最適化するように構成され得る。遠隔的なコンピュータ処理システムによって生成された最適化又は修正された制御ポリシーは、その後、車両による展開のために車両のコンピュータ処理システムへロード又はインストールされて、実際の交通環境において車両を制御し得る。

【００１４】

図１を参照すると、コンピュータ処理システム１４は、コンピュータ処理システム１４及び関連する構成要素の全体的な操作を制御する（少なくとも一つのマイクロプロセッサを含み得る）一つ以上のプロセッサ５８であって、メモリ５４のような一時的でない（non-transitory）コンピュータ可読媒体内に記憶された命令を実行する、プロセッサ５８を含み得る。本開示に関連して、コンピュータ可読記憶媒体とは、命令を実行するシステム、装置若しくはデバイスによって使用されるか又はそれに関連して使用されるプログラムを含む又は記憶し得る任意の有形媒体であり得る。プロセッサ５８は、プログラムコード中に含まれた命令を実施すべく構成された少なくとも一つのハードウェア回路（例えば、集積回路）を含み得る。複数のプロセッサ５８が在る構成において、斯かるプロセッサは相互から独立して作動し得るか、又は、一つ以上のプロセッサが相互に協働して作動し得る。

【００１５】

幾つかの実施形態において、コンピュータ処理システム１４は、ＲＡＭ５０、ＲＯＭ５２、及び／又は他の任意で適切な形態のコンピュータ可読メモリを含み得る。メモリ５４は、一つ以上のコンピュータ可読メモリを備え得る。一つ又は複数のメモリ５４は、コンピュータ処理システム１４の構成要素であり得るか、又は、一つ又は複数のメモリは、コンピュータ処理システム１４に作用的に接続されてコンピュータ処理システム１４に使用

され得る。本説明を通して使用される「作用的に接続された」という語句は、直接的な物理接触のない接続を含め、直接的又は間接的な接続を含み得る。

【００１６】

一つ以上の構成において、本明細書中に記述されたコンピュータ処理システム１４は、人工的又はコンピュータ的な知能要素、例えば、ニューラルネットワーク、又は他の機械学習アルゴリズム、を組み込み得る。更に、一つ以上の構成において、本明細書中に記述された特定の機能又は操作を実施するように構成されたハードウェア及び／又はソフトウェア要素は、複数の要素及び／又は箇所に分散され得る。コンピュータ処理システム１４に加え、車両は、コンピュータ処理システム１４により実施される制御機能を増強若しくは支援するために、又は他の目的のために、付加的なコンピュータ処理システム及び／又はデバイス（図示せず）を組み込み得る。

【００１７】

メモリ５４は、さまざまな機能を実行するためにプロセッサ５８により実行可能であるデータ６０及び／又は命令５６（例えばプログラム論理）を格納し得る。データ６０は、制御ポリシーにより制御されるべき車両の操作に関連する受動的に収集されたデータを含み得る。さらに受動的に収集されたデータは、コンピュータ処理システム１４による使用のため他のソースに対して提供されてよい（あるいは、他のソース上に存在し得る）。受動的に収集されたデータとは、能動的探索から収集されないデータとして定義され得る。高速道路合流操作に関連する受動的に収集されたデータとしては例えば、入り口ランプの近くで高速道路の最も右側のレーン内を走行する車両の速度及び加速、ならびに入り口ランプに沿って走行し最も右側のレーンに進入するサンプル車両の速度及び加速が含まれ得る。受動的に収集されたデータの一例は、建物の上に組付けられたカメラを用いた高速道路の入口の周囲の車両の軌跡の獲得について説明するｈｔｔｐ：／／ｗｗｗ．ｆｈｗａ．ｄｏｔ．ｇｏｖ／ｐｕｂｌｉｃａｔｉｏｎｓ／ｒｅｓｅａｒｃｈ／ｏｐｅｒａｔｉｏｎｓ／０６１３７／中に記載のデータセットである。別の例において、受動的に収集されたデータは、人間のドライバが実行する操作に応答して車両センサが収集するデータを含み得る。人間のドライバにより実行された操作、その操作が実行された車両環境条件、及び操作に後続して及び／又は操作に応答して車両の周囲で発生する事象に関連するデータが収集され、コンピュータ処理システムに提供され得る。代替的には、コンピュータ処理システムが車両内に設置された場合、コンピュータ処理システム１４は、（制御ポリシー１０１などの）１つ以上の車両制御ポリシーのオンライン修正及び／又は最適化のために、このような受動的に収集されたデータを蓄積及び／又は受信するように構成され得る。

【００１８】

車両制御ダイナミクスモデル８７は、車両がさまざまな入力にどのように応答するかを記述する刺激－応答モデルであり得る。車両制御ダイナミクスモデル８７は、所与の車両状態ｘにおける車両についての車両制御ダイナミクスＢ（ｘ）を（受動的に収集されたデータを用いて）決定するように使用され得る。状態コスト関数ｑ（ｘ）は、状態ｘにある車両又はコストであり、逆強化学習などの公知の方法に基づいて学習され得る。状態コストｑ（ｘ）及び車両制御ダイナミクスＢ（ｘ）は、本明細書中に記載されているように制御ポリシー１０１の修正及び最適化の両方のために使用され得る。任意の所与の車両についての車両制御ダイナミクスモデル８７を決定し、メモリ５４などのメモリ内に記載することができる。

【００１９】

再び図１を参照すると、コンピュータ処理システムの実施形態は、２つの学習システム又は学習ネットワーク、並びに相互に作用するactorネットワーク（又は「actor」）８３及びcriticネットワーク（又は「critic」）８１も含み得る。これらネットワークは、例えば、人工ニューラルネットワーク（ＡＮＮ）を用いて実現され得る。本明細書中に記述された目的に対し、（変数πによっても表される）制御ポリシー１０１は、一群の車両の状態のうちの各状態ｘに応じて車両により取られるべき行動ｕを特定又は決定する関数又は他の関係として定義され得る。故に、自律的操作の実行中の車両の各状態ｘに対し、車

両は、関連する行動ｕ＝π（ｘ）を実施するように制御され得る。したがって、制御ポリシーは、車両の操作を制御して、例えば、高速道路合流などの関連する操作を自律的に実施する。actor８３は、制御ポリシーに関して作動し、criticから受信した情報及び他の情報を用いて、ポリシーを修正及び／又は最適化し得る。制御ポリシーにより自律的に制御された車両操作は、高速道路への合流、又は、車線の変更のような特定の目的を達成すべく実施される一つの運転操作又は一群の運転操作として定義され得る。

【００２０】

コンピュータ処理システム１４は、制御ポリシーの修正及び最適化に対して使用可能である新規な半モデルフリーＲＬ方法（semi-model-free RL method）（本明細書においては受動的actor／critic（ｐＡＣ）方法という）を実行するように構成され得る。この方法において、criticは、車両の種々の状態に対する評価関数を学習し、且つ、actorは、能動的探索なしで、代わりに受動的に収集されたデータと既知の車両制御ダイナミクスモデルとを用いて制御ポリシーを改善する。この方法は、部分的に既知であるシステムダイナミクスモデルを使用することにより、能動的探索に対する必要性を回避する。この方法は、車両環境の制御されていないダイナミクス又は過渡的なノイズレベルに関する知見を必要としない。この方法は、例えば、環境がノイズ的に如何に展開するかのサンプルは入手可能であるが車両センサにより能動的に探索することは困難であり得る自律車両に関して、実行可能である。

【００２１】

本明細書中に記載の特定の実施形態において、制御ポリシーにより制御されるべき車両の操作は、１車線内を走行する第２の車両と第３の車両の間でこの車線内に車両を合流させるための操作である。制御ポリシーは、第２の車両と第３の車両の間の中間に車両を合流させる目的で車両を制御するために構成され得る。

【００２２】

本明細書中に記載のコンピュータ処理システム１４の実施形態は、さまざまなタイプの入力及び出力情報を測定、受信及び／又はアクセスすることにより、システム（例えば車両）の状態ｘ（ｔ）を決定する。例えば、システムに連結された又は他の形でシステムと通信状態にあるセンサを用いて、データを測定することができる。コンピュータ処理システム１４は、方程式（１）により特徴付けされる車両の安定性及び所望の運動を達成するためと同時に、方程式（２）中に記載のエネルギーベースのコスト関数を最小化するために、制御入力ｕを決定し得る。

【００２３】

制御ポリシーを修正及び最適化する目的に対し、状態$x \in R^n$及び制御入力$u \in R^m$により、離散時間確率論的ダイナミクス系は以下のように定義され得る。

$$\Delta x = A(x_t)\Delta t + B(x_t)u_t\Delta t + C(x_t)d\omega \qquad (1)$$

式中、ω（ｔ）はブラウニアン運動であり、Ａ（x_t）、Ｂ（x_t）u_t、及びＣ（x_t）は、それぞれ、受動的ダイナミクス、車両制御用ダイナミクス、及び、過渡的ノイズレベルである。Δｔは、時間のステップサイズである。この種の系は、多くの状況において生ずる（例えば、ほとんどの機械系のモデルはこれらのダイナミクスに従う）。関数Ａ（ｘ）、Ｂ（ｘ）及びＣ（ｘ）は、理解されるべく、モデル化されている特定の系に依存する。受動的ダイナミクスは、車両の環境における変化であって、車両システムに対する制御入力の結果ではない変化を含む。

【００２４】

本明細書中に記述された方法及びシステムにおいて、離散時間ダイナミクス系に対するマルコフ決定過程（ＭＤＰ）は、タプル＜Ｘ、Ｕ、Ｐ、Ｒ＞であり、式中、$X \subseteq R^n$及び$U \subseteq R^m$は、状態空間及び行動空間である。Ｐ：＝｛ｐ（ｙ｜ｘ，ｕ）｜ｘ，ｙ∈Ｘ，ｕ∈Ｕ｝は、行動による状態遷移モデルであり、且つ、Ｒ：＝｛ｒ（ｘ，ｕ）｜ｘ∈Ｘ，ｕ∈Ｕ｝は、状態ｘ及び行動ｕに関する即時コスト関数である。先に記述されたように、制御ポリシーｕ＝π（ｘ）は、状態ｘから行動ｕへとマッピングする関数である。予期される累積コストである、ポリシーπの下での到達コスト関数（cost-to-go function）（又

は価値関数）V^{π}（ｘ）は、無限時間区間（infinite horizon）の平均コストの最適性判断基準の下で、以下のように定義される。

【数１】

$$V^{\pi}(x_t) := \sum_{k=1}^{\infty} p(\mathrm{x}_k, \pi(\mathrm{x}_k)) r(\mathrm{x}_k, \pi(\mathrm{x}_k)) \Delta t - V_{avg}^{\pi}$$

$$V_{avg}^{\pi} := \lim_{N\to\infty} \frac{1}{N\Delta t} \sum_{k=1}^{N} p(\mathrm{x}_k, \pi(\mathrm{x}_k)) r(\mathrm{x}_k, \pi(\mathrm{x}_k)) \Delta t$$

式中、

V_{avg}^{π}

は平均コストであり、ｋは時間インデックスであり、且つ、Δｔは時間ステップである。最適な到達コスト関数は、以下の離散時間ハミルトン－ヤコビ－ベルマン方程式を満足する。

【数２】

$$V_{avg}^{\pi} + V^{\pi}(\mathrm{x}_k) = \min_{u_k} Q^{\pi}(\mathrm{x}_k, \mathrm{u}_k), \qquad (2)$$

式中、 $Q^{\pi}(\mathrm{x}_k, \mathrm{u}_k) := r(\mathrm{x}_k, \mathrm{u}_k)\Delta t + \mathcal{G}[V^{\pi}](\mathrm{x}_k)$,であり、且つ、

$\mathcal{G}[V^{\pi}](\mathrm{x}_k) := \int_X p(\mathrm{y}|\mathrm{x}_k, \pi) V^{\pi}(\mathrm{y}) d\mathrm{y}$ である。

式中、Q^{π}（ｘ，ｕ）は行動価値関数（action-value function）であり、且つ、ｇ［・］は積分演算子である。ＭＤＰの目的は、以下の関係に従い、無限時間区間に亘り、平均コストを最小化する制御ポリシーを見出すことである。

【数３】

$$\pi^*(\mathrm{x}_k) = \arg\min_{\pi} \mathbb{E}\left[V_{avg}^{\pi}\right]$$

ここで、最適な制御ポリシーにおける値は、上付き文字*を以て表され得る（例えば、V^*、V^*_{avg}）。

【００２５】

離散時間ダイナミクス系に対する線形マルコフ決定過程（Ｌ－ＭＤＰ）は、連続的な状態空間及び行動空間に対して厳密な解が迅速に求められ得るという利点を備えた汎用マルコフ決定過程のサブクラスである。構築されたダイナミクス、及び、別体的な状態コスト及び制御コストの下で、ベルマン方程式は、組み合わされた状態コスト及び制御されていないダイナミクスの線形固有関数を見出すことに解が限定された線形微分方程式として再構築され得る。その後、Ｌ－ＭＤＰに対する到達コスト関数（又は、価値関数）は、正確なダイナミクスモデルが利用可能であるときに、二次プログラミング（ＱＰ）のような最適化方法により、効率的に求められ得る。

【００２６】
マルコフ決定過程の線形公式は、以下に示されるように、制御コストを定義すべく、且つ、車両ダイナミクスに関する条件を加えるべく使用され得る。
【数４】

$$r(\mathrm{x}_k, \mathrm{u}_k) := q(\mathrm{x}_k) + KL(p(\mathrm{x}_{k+1}|\mathrm{x}_k) \| p(\mathrm{x}_{k+1}|\mathrm{x}_k, \mathrm{u}_k)) \quad (3)$$

$$p(\mathrm{x}_{k+1}|\mathrm{x}_k) = 0 \Longrightarrow \forall \mathrm{u}_k, p(\mathrm{x}_{k+1}|\mathrm{x}_k, \mathrm{u}_k) = 0 \quad (4)$$

ここで、ｑ（ｘ）≧０は状態コスト関数であり、ｐ（ｘ）は行動による状態遷移モデルであり、且つ、ＫＬ（・||・）はクルバック－ライブラー（ＫＬ）偏差である。式（３）は、行動のコストを、それが系に対して有する確率論的効果の量に対して関連付け、且つ、それを状態コストに対して加算する。第２の条件は、何らの行動も、受動的ダイナミクスの下では達成され得ない新たな遷移を導入しないことを確実とする。式（１）により表された確率論的ダイナミクス系は、当然、上記仮定を満足する。
【００２７】
ハミルトン－ヤコビ－ベルマン方程式（式（２））は、Ｌ－ＭＤＰ形態において、指数的に変換された到達コスト関数に対する線形微分方程式（以下、線形化ベルマン方程式という）へと書き換えられ得る。
【数５】

$$Z_{avg} Z(\mathrm{x}_k) = \exp(-q(\mathrm{x}_k)\Delta t)\mathcal{G}[Z](\mathrm{x}_{k+1})$$

$$Z(\mathrm{x}) := \exp\left(-V^*(\mathrm{x})\right), \quad (5)$$

$$Z_{avg} := \exp\left(-V^*_{avg}\right)$$

$$p(\mathrm{x}_{k+1}|\mathrm{x}_k, \pi^*_k) = \frac{p(\mathrm{x}_{k+1}|\mathrm{x}_k) Z(\mathrm{x}_{k+1})}{\mathcal{G}[Z](\mathrm{x}_{k+1})}$$

式中、Ｚ（ｘ）及びZ_{avg}は、それぞれ、Ｚ値と称される指数的に変換された到達コスト関数、及び、最適ポリシーの下での平均コストである。Ｚ値は入力パラメータｘの対応する値に対するＺ値関数Ｚ（ｘ）の特定の値であってもよい。（式（１））における状態遷移はガウス性であることから、制御されたダイナミクスと受動的なダイナミクスとの間のＫＬ偏差は、
【数６】

$$KL\left(p(\mathrm{x}_{k+1}|\mathrm{x}_k) \,\|\, p(\mathrm{x}_{k+1}|\mathrm{x}_k, \mathrm{u}_k)\right) = \frac{1}{2\rho(\mathrm{x}_k)} \mathrm{u}_k^{\mathsf{T}} \mathrm{u}_k$$

$$\frac{1}{\rho(\mathrm{x}_k)} := B(\mathrm{x}_k)^{\mathsf{T}} \left(C(\mathrm{x}_k)^{\mathsf{T}} C(\mathrm{x}_k)\right)^{-1} B(\mathrm{x}_k) \quad (6)$$

として表され得る。
【００２８】
その後、Ｌ－ＭＤＰ系に対する最適な制御ポリシーは、

【数７】

$$\pi^*(\mathrm{x}_k) = -\rho(\mathrm{x}_k)B(\mathrm{x}_k)^{\mathrm{T}} V_{\mathrm{x}_k} \qquad (7)$$

として表され、式中、

$$V_{\mathrm{x}_k}$$

は、x_kにおけるｘに関する到達コスト関数Ｖの偏微分値であり、パラメータρ（x_k）はＢ（x_k）$^{T}V_k$で表されるベクトルが乗算する回数を表す制御ゲインである。Ｚ値及び平均コストは、系のダイナミクスが完全に入手可能であるとき、固有値又は固有関数を解くことにより、線形化ベルマン方程式から導かれ得る。

【００２９】

固有値問題の解決法については、全体が参照により本明細書に組込まれているAdvances in Neural Information Processing Systems, ２００６, ｐ１３６９～１３７６, Ｖｏｌ．１９中で公開された「Linearly-solvable Markov Decision Problems」内でTodorovにより論述されている。固有関数問題の解決法については、同様に全体が参照により本明細書に組込まれているConference:In Adaptive Dynamic Programming and Reinforcement Learning,IEEE Symposium, ２００９, ｐ１６１～１６８中で公開された「Eigenfunction Approximation Methods for Linearly-solvable Optimal Control Problems」内でTodorovにより論述されている。

【００３０】

本明細書中に記載されているコンピュータ処理システム１４の実施形態には、互いに相互作用する２つの学習システム又は学習ネットワーク、ａｃｔｏｒネットワーク（又は「ａｃｔｏｒ」）８３及びｃｒｉｔｉｃネットワーク（又は「ｃｒｉｔｉｃ」）８１が含まれる。これらのネットワークは、人工神経ネットワークを用いて実装され得る。

【００３１】

１つ以上の配置において、ａｃｔｏｒ８３は内部ループフィードバックコントローラとして実装され、ｃｒｉｔｉｃ８１は外部ループフィードバックコントローラとして実装される。両方共、車両起動型メカニズム又は制御指令をもたらすために操作可能である制御機構との関係においてフィードフォワード経路内に位置設定されてよい。

【００３２】

反復とは、（ｃｒｉｔｉｃについては重みω、ａｃｔｏｒについてはμなどの）ｃｒｉｔｉｃ及びａｃｔｏｒパラメータの更新として定義され得る。さらに、ｃｒｉｔｉｃネットワークパラメータの更新は、車両が動いているときに行なうことができる。本明細書中に記載の方法において、ここで、ｃｒｉｔｉｃネットワーク及びａｃｔｏｒネットワークパラメータの更新中に使用される唯一のデータは、受動的に収集されたデータである。

【００３３】

ｃｒｉｔｉｃ８１は、受動的に収集されたデータのサンプル内に反映されている状態及び状態コストを用いて、推定平均コスト及び、ａｃｔｏｒネットワークによって適用された場合に車両の到達コストについての最小値を生成する近似された到達コスト関数を決定する。受動的に収集されたデータのサンプル内に反映された車両状態及び車両制御ダイナミクスモデル８７から受信した状態コストｑ（ｘ）を用いて、ｃｒｉｔｉｃ８１は、車両の現在の状態x_k及び推定された次の状態x_{k+1}、及び受動的に収集されたデータのサンプルを用いる最適ポリシー下の状態コストq_kを評価する。ｃｒｉｔｉｃ８１は同様に、近似された到達コスト関数

$\hat{Z}(x)$

（Ｚ値関数）及び現在の状態についての付随する推定Ｚ値を決定し、ａｃｔｏｒ８３による使用のための推定された平均コスト

$\hat{Z}_{avg}$

を生成するために、前述のベルマン方程式（方程式（５））の線形化版を使用する。推定された次の状態x_{k+1}は、受動的に収集されたデータ及び車両ダイナミクスモデル８７を用いて計算可能である。

【００３４】

Ｚ値の推定を目的として、重み付けされた放射基底関数（ＲＢＦ）の線形結合がＺ値関数を近似するために使用され得る：

【数８】

$$Z(x) \approx \hat{Z}(x) := \sum_{j=0}^{N} \omega_j f_j(x)$$

ここでωは重み、f_jはｊ番目のＲＢＦそしてＮはＲＢＦの数である。基底関数は、車両システムの非線形ダイナミクスに応じて好適に選択され得る。Ｚ値は、近似Ｚ値関数及び受動的に収集されたデータのサンプルを用いて近似され得る。

【００３５】

重み付けされた放射基底関数の線形結合を用いてＺ値関数を近似する前に、重み付けされた放射基底関数内で使用される重みを最適化することができる。放射基底関数内で使用するため、累乗された真の到達コスト（又はＺ値）と推定到達コストとの間の最小二乗誤差を最小化することにより、重みωを最適化することができる。Ｚ（x_k）及びZ_{avg}を真のＺ値とし、

$\hat{Z}_k$

、

$\hat{Z}_{avg}$

を推定Ｚ値とすると、

【数９】

$$\min_{\omega,\hat{Z}_{avg}} \frac{1}{2} \sum_{D} \left(\hat{Z}_{avg}\hat{Z}_k - Z_{avg}Z_k\right)^2, \qquad (8)$$

【数１０】

$$s.t\sum_{i=0}^{N}\omega_i = C, \forall i\ \omega_i \geq 0, \forall x\ \hat{Z}_{avg}\hat{Z}(x) \leq 1, \qquad (9)$$

ここで、Ｃは、自明な解ω＝０への収束を回避するために使用される一定値である。方程式（５）に由来する∀ｘ、０＜Ｚ$_{avg}$Ｚ（ｘ）≦１，＼及び∀ｘ、ｑ（ｘ）≧０を満たすために、第２及び第３の制約が必要とされる。

【００３６】

重みを最適化する前に、重み付けされた放射基底関数内で使用される重みを更新することができる。重みωは最適化に先立ち更新され得、ｃｒｉｔｉｃネットワークにより使用される重み及び平均コスト

$\hat{Z}_{avg}$

は、ｐＡＣ方法のために使用される情報を用いて真の到達コスト及び真の平均コストを決定することができないことを理由として、線形化されたベルマン方程式（ＬＢＥ）（方程式（５））から以下のように決定される近似された時間差誤差ｅ$_k$で真の到達コストと推定到達コスト間の誤差を近似することによって、ラングランジェ緩和時間差（ＴＤ）学習に基づいて反復ステップにおける使用に先立ち更新され得る。

【数１１】

$$\hat{Z}_{avg}\hat{Z} - Z_{avg}Z_k \approx e_k := \hat{Z}_{avg}\hat{Z}_k - \exp(-q_k)\hat{Z}_{k+1}$$

$$\tilde{\omega}^{i+1} = \omega^i - \alpha_1^i e_k Z_{avg} f_k \qquad (10)$$

【数１２】

$$\omega^{i+1} = \tilde{\omega}^i + \lambda_1 + \sum_{l=0}^{N}\lambda_{2l}\delta_{jl} + \lambda_3\hat{Z}_{avg}f_k \qquad (11)$$

【数１３】

$$\hat{Z}_{avg}^{i+1} = \hat{Z}_{avg}^i - \alpha_2{}^i e_k \hat{Z}_k$$

ここでα$_1{}^i$及びα$_2$は、学習率であり、ｅ$_k$はＬ－ＭＤＰｓについてのＴＤ誤差である。δ$_{ij}$はディラックのデルタ関数を意味する。上付き文字ｉは、反復回数を意味する。λ$_1$、λ$_2$、λ$_3$は、制約方程式（９）についてのラングランジェ乗数である。ωは、方程式（１０）で誤差を最小化し、方程式（１１）で制約を満たすために更新される。

【００３７】

乗数の値は、以下の方程式を解くことで計算される。

【数１４】

$$\begin{bmatrix} \sum_j f_j & \begin{vmatrix} f_0 \cdots f_N \\ 1 \cdots 0 \\ \ddots \\ 0 \cdots 1 \\ 1 \cdots 1 \end{vmatrix} & \begin{matrix} f^T f \\ f_0 \\ \vdots \\ f_N \\ \sum_j f_j \end{matrix} \end{bmatrix} \begin{bmatrix} \lambda_1^i \\ \lambda_2^i \\ \lambda_3^i \end{bmatrix} = \begin{bmatrix} c - \sum_j \tilde{\omega}^i \\ -\tilde{\omega}^i{}_0 \\ \vdots \\ -\tilde{\omega}^i{}_N \\ 1 - \hat{Z}_{avg} \hat{Z}_k \end{bmatrix}$$

【００３８】

いくつかの事例において、制約サブセットが有効でない場合がある。このような場合には、これらの制約についての乗数はゼロに設定され、残りの有効な制約についての乗数が得られる。ｃｒｉｔｉｃは、制御ポリシーと無関係な状態コストq_kと受動的ダイナミクス（$x_k x_{k+1}$）の下での状態遷移サンプルを用いてパラメータを更新する。重みω、推定Ｚ値

$\hat{Z}$

及び平均コスト

$\hat{Z}_{avg}$

は、車両が動いている間に、方程式（１０）－（１１Ａ）にしたがってオンラインで更新され得る。

【００３９】

コンピュータ処理システム内で、ｃｒｉｔｉｃネットワークに作用的に連結されたａｃｔｏｒ８３は、到達コストについて最小値を生成する車両に対し適用するための制御入力を決定し得る。ｃｒｉｔｉｃにより生成された推定到達コスト

$\hat{Z}$

及び推定平均コストZ_{avg}、状態コストｑ（ｘ）、車両制御ダイナミクスモデル８７から決定される現在の状態についての制御ダイナミクス情報Ｂ（ｘ）、及び到達コスト関数

$\hat{Z}(x)$

を推定し推定平均コスト

$\hat{Z}_{avg}$

を生成するためにｃｒｉｔｉｃにより使用される車両の現在の状態及び推定された次の状態を用いて、ａｃｔｏｒ８３は制御入力を決定することができる。制御入力は、制御ポリシーπを修正するために使用可能である。特定の実施形態において、ポリシーπは、本明細書中に記載の要領で収束に至るまで反復的に修正され、その時点で最適化されたものと

みなされる。ａｃｔｏｒは、標準ベルマン方程式を用いて能動的探索なしで制御ポリシーを改良し修正する。制御ダイナミクスは、車両についての公知の制御ダイナミクスから決定され得る。

【００４０】

ａｃｔｏｒ８３は同様に、所望される操作（例えば高速道路合流、レーン変更など）を自律的に行なうため、実時間で車両システムに対して制御入力ｕ（ｘ）を適用することもできる。本明細書中で開示されているいくつかの実施形態において、ａｃｔｏｒ８３は、内部ループフィードバックコントローラ内で具現され得、ｃｒｉｔｉｃ８１は外部ループフィードバックコントローラ内で具現され得る。両方のコントローラ共、車両起動式制御機構との関係においてフィードフォワード経路内に位置設定され得る。

【００４１】

ａｃｔｏｒ８３は、ｃｒｉｔｉｃ由来の推定値（例えば

$\hat{Z}$

及び

$\hat{Z}_{avg}$

）、受動的ダイナミクス下のサンプル、公知の制御ダイナミクスB_kを用いて制御ゲインρ（x_k）を推定することにより、制御ポリシーを改良又は修正することができる。制御ポリシーの修正には、制御ゲインを近似するステップ、制御ゲインを最適化して最適化された制御ゲインを提供するステップ、及び最適化された制御ゲインを用いて制御ポリシーを修正するステップが含まれ得る。

【００４２】

制御ゲインρ（x_k）は、重み付けされた放射基底関数の線形結合で学習された状態で近似され得る：

【数１５】

$$\rho(x;\mu) \approx \hat{\rho}(x;\mu) := \sum_{j}^{M} \mu_j g_j(x) \qquad (12)$$

ここで、μ_jは、ｊ番目の放射基底関数g_jのための重みである。Ｍは放射基底関数の数である。ρ（x_k）は、到達コストと行動－状態価値の間の最小平均誤差を最小化することによって最適化され得る。

【数１６】

$$\min_{\mu} \frac{1}{2} \sum_{D} \left(\hat{Q}_k - V_k^* - V_{avg}^* \right)^2$$

ここでV^*、V^*_{avg}、及び

$\hat{Q}$

は、最適な制御ポリシー下の、真の到達コスト関数、平均コスト及び推定行動－状態価値である。最適な制御ポリシーは、ポリシーが最適なポリシーである場合にのみ真の行動価値コストが$V^{*}+V^{*}_{avg}$に等しいことから、目的関数を最小化することによって学習され得る。以下の関係にしたがって、制御ゲインρ（x_k）を更新する場合に

$\hat{V}_k$

及び

$\hat{V}_{avg}$

を決定するために、

$\hat{Z}_k$

及び

$\hat{Z}_{avg}$

を使用することができる：

【数１７】

$$\hat{V}_k = -log(\hat{Z}_k)$$

$$\hat{V}_{avg} = -log(\hat{Z}_{avg})$$

【００４３】

制御ゲインを最適化する前に、制御入力を決定することができ、制御入力、受動的に収集されたデータのサンプル及び近似された制御ゲインを用いて行動価値関数の値Ｑを決定することができる。重み付けされた放射基底関数の線形結合を用いて制御ゲインを近似する前に、重み付けされた放射基底関数内で使用される重みμを更新することができる。

【００４４】

重みμは、以下で定義する近似時間差（ＴＤ）誤差d_kを用いて更新され得る：

【数１８】

$$\mu^{i+1} = \mu^{i} - \beta^{i} d_k L_{k,k+1} g_k,$$

$$d_k \approx q_k \Delta t - \hat{V}_{k+1} + \hat{V}_k + \hat{V}_{avg} + L_{k,k+1}\rho_k,$$

$$L_{k,k+1} := \left(0.5\hat{V}_k - \hat{V}_{k+1}\right)^{\top} B_k B_k^{\top} \hat{V}_k \Delta t$$

ここで、β’は学習率であり、$L_{k、k+1}$は項Ｌ（x_k、x_{k+1}）の省略版である。

【００４５】

真の到達コスト及び真の平均コストを計算することができないため、誤差d_kを決定するために標準ベルマン方程式を近似することができる。

【数１９】

$$\hat{Q}_k - V_k^* - V_{avg}^* \approx d_k := \hat{Q}_k - \hat{V}_k - \hat{V}_{avg}$$

$$\hat{Q}_k \approx q_k \Delta t + \frac{0.5\Delta t}{\hat{\rho}_k} u_k^{\top} u_k + \hat{V}(x_{k+1} + B_k u_k \Delta t)$$

$$u_k \approx -\hat{\rho}_k B_k^{\top} \hat{V}_k,$$

ここで、x_{k+1}は受動的ダイナミクス下の次の状態であり、$x_{k+1} + B_k u_k \Delta t$は、行動$u_k$での制御されたダイナミクス下における次の状態である。推定到達コスト、平均コスト及びそれらの微分値は、ｃｒｉｔｉｃからの推定Ｚ値及び平均Ｚ値コストを使用することによって計算可能である。さらに、以下の式により、μとの関係においてＴＤ誤差を線形化するために、

$$\hat{V}(x_{k+1} + B_k u_k \Delta t)$$

を近似することができる。

【数２０】

$$\hat{V}(x_{k+1} + B_k u_k \Delta t) \approx \hat{V}_{k+1} + \hat{V}_{x_{k+1}}^{\top} B_k u_k \Delta t.$$

【００４６】

この手順は、受動的ダイナミクス下での状態遷移サンプル（x_k、x_{k+1}）、状態コストq_k、及び所与の状態における制御ダイナミクスB_kを使用することによって能動的探索なしでポリシーを改良する。標準ａｃｔｏｒ－ｃｒｉｔｉｃ方法は、能動的探索を用いてポリシーを最適化する。これらのａｃｔｏｒ及びｃｒｉｔｉｃ関数が定義された状態で、コンピュータ処理システム１４は、Ｌ－ＭＤＰを用いて半モデルフリー強化学習を実装することができる。

【００４７】

本明細書中に記載の方法において、ポリシーは、到達コストと行動－状態価値との間の誤差を最小化することによって、車両制御ダイナミクスについての知識及び受動的に収集されたデータのサンプルを用いて学習されるパラメータで最適化される。本明細書中に記載の方法は、通常車を制御するために利用可能である車両自体のダイナミクスモデルで最

適なポリシーを決定することを可能にする。これらの方法は、同様に、通常そのダイナミクスモデルが未知である周囲の車両の操作に関する受動的に収集されたデータも使用する。さらに、本明細書に記載の方法を用いると、最適な制御ポリシーを決定するために、車両環境の受動的ダイナミクスＡ（x_t）及び遷移ノイズレベル及びＣ（x_t）を知っている必要はない。

【００４８】

別の態様においては、本明細書中で記載されているように、１つの操作を行なうようシステムを制御するために使用可能な制御ポリシーを最適化するためのコンピュータ実装型方法が提供されている。この方法は、システムを制御するために使用可能な制御ポリシーを提供するステップと；行なうべき操作に関する受動的に収集されたデータに対して受動的ａｃｔｏｒ－ｃｒｉｔｉｃ強化学習方法を適用して、最低予想累積コストで操作を行なうようにシステムを制御するために制御ポリシーが操作可能になるような形で制御ポリシーを修正するステップと；を含むことができる。受動的に収集されたデータに対して受動的ａｃｔｏｒ－ｃｒｉｔｉｃ強化学習方法を適用するステップは、ａ）コンピュータ処理システム内のｃｒｉｔｉｃネットワークにおいて、受動的に収集されたデータのサンプルを用いてＺ値を推定し、受動的に収集されたデータのサンプルを用いて最適なポリシー下で平均コストを推定するステップと；ｂ）コンピュータ処理システム内のａｃｔｏｒネットワークにおいて、受動的に収集されたデータのサンプル、システムについての制御ダイナミクス、到達コスト及び制御ゲインを用いて制御ポリシーを修正するステップと；ｃ）制御ポリシーを修正する上で、及び最適なポリシー下でＺ値及び平均コストを推定する上で使用されるパラメータを更新するステップと；ｄ）推定平均コストが収束するまで、ステップ（ａ）～（ｃ）を反復的に繰返すステップとを含むことができる。

【００４９】

図６～９は、本明細書中に記載の一実施形態に係る、最小予想累積コストで車両の操作を行なう目的で車両を制御するために構成された制御ポリシーを学習するため、車両の操作に関連する受動的に収集されたデータに対して受動的ａｃｔｏｒ－ｃｒｉｔｉｃ強化学習方法を適用するコンピュータ実装型方法を例示するフローチャートである。

【００５０】

図６を参照すると、ブロック７１０において、プロセッサ５８は、行なうべき車両の操作に関連する受動的に収集されたデータを受信し得る。受動的に収集されたデータは、メモリ５４及び／又はコンピュータ処理システム１４の外部のソースから受信され得る。

【００５１】

ブロック７２０では、コンピュータ処理システム１４のプロセッサ及び／又は他の要素は、受動的に収集されたデータのサンプルに対し、本明細書中に記載の受動的ａｃｔｏｒ－ｃｒｉｔｉｃ（ＰＡＣ）強化学習方法を反復的に適用することができる。ＰＡＣ方法を適用することにより、最低予想累積コストで車両の操作を行なうように車両が制御され得るようにする制御ポリシーを学習することが可能である。ブロック７３０では、車両は、車両の操作を行なうために学習された制御ポリシーにしたがって制御され得る。

【００５２】

図７は、図６のブロック７２０に示されているような、本明細書中に記載の実施形態に係る受動的ａｃｔｏｒ－ｃｒｉｔｉｃ（ＰＡＣ）強化学習方法の適用を例示するフローチャートである。

【００５３】

図７を参照すると、ブロック８１０において、コンピュータ処理システムは、車両の操作を行なうように車両を制御するために適応され得る初期制御ポリシーを受信することができる。制御ポリシーの初期版のパラメータは、コンピュータ処理システム内のランダム化ルーチンを用いてランダム値に初期化され得る。

【００５４】

ブロック８２０では、コンピュータ処理システムは、受動的に収集されたデータのサンプルを用いて前述の最適な制御ポリシー下でＺ値及び平均コストを推定することができる

。ブロック８３０では、コンピュータ処理システムは、最適なポリシー下で受動的に収集されたデータ、推定されたＺ値及び推定された平均コストのサンプルを用いて制御ポリシーを修正することができる。ブロック８４０では、コンピュータ処理システムは、推定平均コストが収束するまで、ブロック８２０及び８３０に示されたステップを反復的に繰返すことができる。

【００５５】

図８は、図７のブロック８２０に示されているように、受動的に収集されたデータのサンプルを用いて最適な制御ポリシー下でＺ値及び平均コストを推定するための、ｃｒｉｔｉｃネットワークによる受動的ａｃｔｏｒ－ｃｒｉｔｉｃ（ＰＡＣ）強化学習方法のステップの適用を例示するフローチャートである。

【００５６】

ブロック９１０では、Ｚ値関数を近似するために使用可能である重み付けされた放射基底関数内で使用される重みを更新することができる。ブロック９２０では、Ｚ値関数を近似するために使用可能である重み付けされた放射基底関数内で使用される重みを、最適化することができる。ブロック９３０では、重み付けされた放射基底関数の線形結合を用いて、Ｚ値関数を近似することができる。ブロック９４０では、ブロック９３０で決定された近似Ｚ値関数及び受動的に収集されたデータのサンプルを用いてＺ値を近似することができる。

【００５７】

図８は、図７のブロック８３０に示されているように、受動的に収集されたデータのサンプルを用いて最適な制御ポリシー下でＺ値及び平均コストを推定するための、ａｃｔｏｒネットワークによる受動的ａｃｔｏｒ－ｃｒｉｔｉｃ（ＰＡＣ）強化学習方法のステップの適用を例示するフローチャートである。

【００５８】

ブロック１０１０では、制御ゲインを近似するために使用可能である重み付けされた放射基底関数内で使用される重みを更新することができる。ブロック１０２０では、重み付けされた放射基底関数の線形結合を用いて、制御ゲインρ（x_k）を近似することができる。先に説明された関係（１２）及び受動的に収集されたデータのサンプルを用いて制御ゲインを近似することができる：

【数２１】

$$\rho(\mathbf{x};\mu) \approx \hat{\rho}(\mathbf{x};\mu) := \sum_{j}^{M} \mu_j g_j(\mathbf{x}) \qquad (12)$$

【００５９】

ブロック１０３０では、関係

$$\mathbf{u}_k \approx -\hat{\rho}_k B_k^{\mathrm{T}} \hat{V}_{\mathbf{k}},$$

を用いて制御入力ｕを決定することができる。ブロック１０４０では、ブロック１０３０で決定された制御入力、受動的に収集されたデータのサンプル及びブロック１０２０で近似された制御ゲインρ（x_k）を用いて、行動価値関数Ｑの値を決定することができる。行動価値関数Ｑの値は、先に段落［００４５］で明記された関係、

【数２２】

$$\hat{Q}_k \approx q_k \Delta t + \frac{0.5\Delta t}{\hat{\rho}_k}\mathbf{u}_k^{\mathrm{T}}\mathbf{u}_k + \hat{V}(\mathbf{x}_{k+1} + B_k\mathbf{u}_k\Delta t)$$

を用いて決定することができる。ブロック１０５０では、制御ゲインを最適化して、最適化制御ゲインを提供することができる。ブロック１０４０で決定された行動価値関数Ｑの値、及び先に段落［００５７］で明記された関係、

【数２３】

$$\rho = \min_{\mu}\frac{1}{2}\sum_{D}\left(\hat{Q}_k - V_k^* - V_{avg}^*\right)^2$$

を用いて、制御ゲインを最適化することができる。ブロック１０６０では、最後に最適化制御ゲインρ（x_k）及び関係（７）、すなわち

【数２４】

$$\pi^*(\mathbf{x}_k) = -\rho(\mathbf{x}_k)B(\mathbf{x}_k)^{\mathrm{T}}V_{\mathbf{x}_k}$$

を用いて、制御ポリシーを修正又は更新することができる。

【００６０】

この関係は、先に段落［００２８］中に明記されたものである。上述したｃｒｉｔｉｃ及びａｃｔｏｒネットワークにより行なわれるステップは、推定平均コストが収束するまで、追加の受動的に収集されたデータについて反復的に繰返し可能である。

【００６１】

図２は、本明細書中に記載の方法に係るコンピュータ処理システム１４内での制御入力の決定及び制御ポリシーの修正及び制御ポリシーの最適化の実行中の情報の流れを示す概略図である。従来のａｃｔｏｒ－ｃｒｉｔｉｃ方法は環境から能動的に収集されたデータのサンプルを使用して動作し得るが、本明細書中で説明されているｐＡＣ方法は、環境の能動的探索なく、代りに受動的に収集されたサンプル及び公知の車両制御ダイナミクスを用いて、最適な制御ポリシーを決定する。ｃｒｉｔｉｃ８１又はａｃｔｏｒ８３のいずれかで受信したあらゆる情報を、後日使用するためにメモリ内にバッファリングすることができる。例えば、パラメータ値を計算又は推定するためにｃｒｉｔｉｃ又はａｃｔｏｒに必要とされる情報の全てが現在利用可能でない状況においては、残りの所要情報が受信されるまで、受信された情報をバッファリングすることができる。項

$$\hat{Z}_{\mathbf{x}_k}$$

及び

$\hat{Z}_{\mathbf{x}_{k+1}}$

は、それぞれx_k及びx_{k+1}におけるｘに関するＺ値関数の偏導関数である。項

$\hat{Z}_{\mathbf{x}_k}$

は、x_kにおける到達コスト関数Ｖの偏導関数を計算するために使用可能である。

【００６２】

図３は、図１のコンピュータ処理システム１１４と同様の態様で構成されたコンピュータ処理システム１１４が組み込まれた例示的な実施形態に係る車両１１を示す機能的ブロック図である。車両１１は、乗用車、トラック、又は、本明細書中に記述された操作を実施し得る他の任意の車両の形態を取り得る。車両１１は、完全に又は部分的に自律モードで作動すべく構成され得る。自律モードで作動している間、車両１１は、人的相互作用なしで作動すべく構成され得る。例えば、高速道路の合流操作が実行されている自律モードにおいて、車両は、車両乗員からの入力なしで、高速道路上の車両から安全距離を維持すること、他の車両と速度を調和すること等を行うように、スロットル、ブレーキ及び他のシステムを作動させ得る。

【００６３】

車両１１は、コンピュータ処理システム１１４に加え、且つ、相互に作動的に通信する種々のシステム、サブシステム及び構成要素、及び構成要素、例えば、センサシステム又は配列２８、一つ以上の通信インタフェース１６、操舵システム１８、スロットルシステム２０、制動システム２２、電源３０、動力システム２６、並びに本明細書中に記述されたように車両を操作するために必要な他のシステム及び構成要素を含み得る。車両１１は、図３に示されたよりも多い又は少ないサブシステムを含み得ると共に、各サブシステムは、複数の要素を含み得る。更に、車両１１のサブシステム及び要素の各々は、相互接続され得る。車両１１の記述された機能及び／又は自律的作動の一つ以上の実施は、相互に協働して作動している複数の車両システム及び／又は構成要素により実行され得る。

【００６４】

センサシステム２８は、任意の適切な形式のセンサを含み得る。本明細書中には、異なる形式のセンサの種々の例が記述される。しかし、実施形態は、記述された特定のセンサに限定されないことは理解される。

【００６５】

センサシステム２８は、車両１１の外部環境に関する情報を検知すべく構成された所定数のセンサを含み得る。例えば、センサシステム２８は、全地球測位システム（ＧＰＳ）のようなナビゲーションユニット、及び、例えば、慣性測定装置（ＩＭＵ）（図示せず）、ＲＡＤＡＲユニット（図示せず）、レーザ測距計／ＬＩＤＡＲユニット（図示せず）、及び車両の内部及び／又は該車両１１の外部環境の複数の画像を捕捉すべく構成されたデバイスを備える一台以上のカメラ（図示せず）等の他のセンサを含み得る。カメラは、スチルカメラ又はビデオカメラであり得る。ＩＭＵは、慣性加速度に基づいて車両１１の位置及び向きの変化を検知するように構成されたセンサ（例えば、加速度計及びジャイロスコープ等）の任意の組合せを組み込み得る。例えば、ＩＭＵは、車両のロール速度、ヨーレート、ピッチ速度、長手方向加速度、横方向加速度、及び、垂直加速度のようなパラメータを検知し得る。ナビゲーションユニットは、車両１１の地理的位置を推定すべく構成された任意のセンサであり得る。この目的の為に、ナビゲーションユニットは、地球に対する車両１１の位置に関する情報を提供するように作動可能な送受信機を含む一つ以上の

送受信機を含み得る。また、ナビゲーションユニットは、業界公知の態様で、記憶され且つ／又は利用可能な地図を用いて与えられた開始点（例えば、車両の現在位置）から、選択された目的地までの走行ルートを決定又は計画するように構成され得る。また、車両１１に近接して又は所定の距離以内で移動する車両に関する近さ、距離、速度及び他の情報を検出するように構成された一つ以上のセンサが設けられてもよい。

【００６６】

公知の態様において、車両センサ２８は、種々の車両システムに対する適切な制御命令を策定且つ実行する際にコンピュータ処理システム１１４により使用されるデータを提供する。例えば、慣性センサ、車輪速度センサ、道路状態センサ、及び操舵角センサからのデータは、車両を旋回させるための命令を策定して操舵システム１８において実行する上で、処理され得る。各車両センサ２８は、車両１１に組み込まれる任意の運転者支援機能及び自律的操作機能をサポートするために必要とされる任意のセンサを含み得る。センサシステム２８が複数のセンサを含む構成において、センサは、相互から独立的に作動し得る。代替的に、各センサのうちの２つ以上が、相互に協働して作動し得る。センサシステム２８のセンサは、コンピュータ処理システム１１４に対し、及び／又は車両１１の他の任意の要素に対し、作用的に接続され得る。

【００６７】

また、各車両センサ２８により収集された任意のデータは、本明細書中に記述された目的でデータを必要とし又は利用する任意の車両システム又は構成要素にも送信され得る。例えば、車両センサ２８により収集されたデータは、コンピュータ処理システム１１４に、又は一つ以上の専用のシステム又は構成要素のコントローラ（図示せず）に送信され得る。付加的な特定の形式のセンサとしては、本明細書中に記述された機能及び操作を実施するために必要とされる他の任意の形式のセンサが挙げられる。

【００６８】

特定の車両センサからの情報は、一つよりも多い車両システム又は構成要素を制御すべく処理かつ使用され得る。例えば、自動化された操舵制御及び制動制御の両方を組み込んだ車両において、種々の道路状態センサは、データをコンピュータ処理システム１１４に提供し、このコンピュータ処理システムは、プロセッサが実行可能な記憶された命令に従って道路状態情報を処理すると共に、操舵システム及び制動システムの両方に対して適切な制御命令を策定することができるようになる。

【００６９】

車両１１は、センサの出力信号又は他の信号が、コンピュータ処理システム１１４又は別の車両システム若しくは要素による使用の前に前処理を必要とするという状況、又はコンピュータ処理システムから送信された制御信号が、起動可能なサブシステム又はサブシステム構成要素（例えば、操舵システム又はスロットルシステムの構成要素）による使用の前に処理を必要とするという状況に適した、信号処理手段３８を含み得る。信号処理手段は、例えば、アナログ／デジタル（Ａ／Ｄ）変換器又はデジタル／アナログ（Ｄ／Ａ）変換器であり得る。

【００７０】

センサ統合機能（sensor fusion capability）１３８は、センサシステム２８からのデータを入力として受け入れるべく構成されたアルゴリズム（又は、アルゴリズムを記憶するコンピュータプログラム製品）の形態であり得る。上記データは、例えば、センサシステム２８の各センサにて検知された情報を表すデータを含む。センサ統合アルゴリズムは、センサシステムから受信したデータを処理し、（例えば、複数の個別的なセンサの出力から形成された）統合された又は合成された信号を生成し得る。センサ統合アルゴリズム１３８は、例えば、カルマンフィルタ、ベイジアンネットワーク、又は、別のアルゴリズムを含む。センサ統合アルゴリズム１３８は更に、センサシステム２８からのデータに基づく種々のアセスメントを提供し得る。例示的な実施形態において、アセスメントは、車両１１の環境における個別的な物体又は特定構造の評価、特定状況の評価、及び、特定の状況に基づく可能的な影響の評価を含み得る。他のアセスメントも可能である。センサ統

合アルゴリズム１３８は、コンピュータ処理システム１１４に組み込まれた又はコンピュータ処理システム１１４と作用的に通信する（メモリ１５４のような）メモリ内に記憶され得ると共に、当業界において公知の態様でコンピュータ処理システムにより実行され得る。

【００７１】

本明細書中に記述された任意の情報若しくはパラメータの受信、収集、監視、処理、及び／又は、決定を参照するときにおける「連続的に」という語句の使用は、コンピュータ処理システム１１４が、これらのパラメータに関する情報が存在し又は検出されるや否や、又は、センサの取得サイクル及びプロセッサの処理サイクルに従ってできるだけ素早く、任意の情報を受信及び／又は処理すべく構成されることを意味している。コンピュータ処理システム１１４が、例えば、センサからのデータ又は車両構成要素の状況に関する情報を受信すると直ちに、コンピュータ処理システムは、記憶されたプログラム命令に従って動作し得る。同様に、コンピュータ処理システム１１４は、センサシステム２８から及び他の情報源から、同時進行的又は連続的に情報の流れを受信して処理し得る。この情報は、本明細書中に記述された態様及び目的にて、メモリ内に記憶された命令に従って処理及び／又は評価される。

【００７２】

また、図３は、先に記述されたように、図１のコンピュータ処理システム１１４と同様の態様で構成された代表的なコンピュータ処理システム１１４のブロック図も示している。本明細書中に記述されたようにポリシーの修正を実施すると共に制御入力を決定するために必要とされる機能を組み込むと共に、コンピュータ処理システム１１４は、他の車両システム及び要素に作用的に接続されると共に、その他の点では、車両１１及びその構成要素の制御及び操作に影響するように構成され得る。コンピュータ処理システム１１４は、少なくとも幾つかのシステム及び／又は構成要素を、（ユーザ入力なしで）自律的に且つ／又は（一定程度のユーザ入力を以て）半自律的に制御すべく構成され得る。また、コンピュータ処理システムは、幾つかの機能を自律的及び／又は半自律的に制御及び／又は実行するようにも構成され得る。コンピュータ処理システム１１４は、種々のサブシステム（例えば、動力システム２６、センサシステム２８、操舵システム１８）から、各通信インタフェース１６のうちの任意のものから、及び／又は他の任意で適切な情報源から受信した入力及び／又は情報に基づき、車両１１の機能性を制御し得る。

【００７３】

図３の実施形態において、コンピュータ処理システム１１４は、図１に関して先に記述されたように、車両制御ダイナミクスモデル１８７、critic１８１、actor１８３、及び、制御ポリシー２０１を含み得る。コンピュータ処理システム１１４は、先に記述されたように、制御入力を決定すべく、且つ自律車両の操作制御ポリシーを修正及び／又は最適化すべく構成され得る。また、コンピュータ処理システム１１４は、制御入力に従って、且つ、本明細書中に記述されたように修正又は最適化された制御ポリシーにも従って、車両を制御して所望操作を実施すべく構成され得る。

【００７４】

コンピュータ処理システム１１４は、図３に示された要素の幾つか又は全てを有し得る。加えて、コンピュータ処理システム１１４は、特定の用途に必要とされ又は所望される付加的な構成要素も含み得る。また、コンピュータ処理システム１１４は、複数のコントローラ又はコンピュータ処理デバイスであって、分散態様にて、情報を処理し且つ／又は車両１１の個別的な構成要素若しくはサブシステムを制御するように機能する複数のコントローラ又はコンピュータ処理デバイスを表し、又は、それにより具現され得る。

【００７５】

メモリ１５４は、単一又は複数のプロセッサ１５８により実行されて、車両１１の種々の機能を実行するデータ１６０及び／又は命令１５６（例えば、プログラムロジック）を収納し得る。メモリ１５４は、本明細書中に記述された車両システム及び／又は構成要素（例えば、動力システム２６、センサシステム２８、コンピュータ処理システム１１４、

及び、通信インタフェース１６）のうちの一つ以上にデータを送信し、それらからデータを受信し、それらと相互作用し、又はそれらを制御するための命令を含む、付加的な命令も含み得る。命令１５６に加え、メモリ１５４は、他の情報の中でも、道路地図、経路情報のようなデータを記憶し得る。斯かる情報は、自律的、半自律的、及び／又は手動的なモードにおける車両１１の操作の間において、ルートを計画するのに且つその他にことをするのに、車両１１及びコンピュータ処理システム１１４により使用され得る。

【００７６】

コンピュータ処理システム１１４は、（概略的に６２と表される）一つ以上の自律的な機能又は操作を実施するために、種々の起動可能な車両システム及び構成要素の制御を連携調整するように構成され得る。これらの自律的な機能６２は、メモリ１５４及び／又は他のメモリ内に記憶されると共に、プロセッサにより実行されたときに、本明細書中に記述された種々のプロセス、命令又は機能のうちの一つ以上を実現するコンピュータ可読プログラムコードの形態で実現され得る。

【００７７】

通信インタフェース１６は、車両１１と、外部センサ、他の車両、他のコンピュータシステム、（本明細書中に記述されたように、衛星システム、携帯電話／無線通信システム、種々の車両サービスセンターなどのような）種々の外部のメッセージ及び通信システム、及び／又はユーザとの間の相互作用することができるように構成され得る。通信インタフェース１６は、車両１１のユーザに情報を提供し又はユーザから入力を受信するためのユーザインタフェース（例えば、一台以上のディスプレイ（図示せず）、音声／オーディオインタフェース（図示せず）、及び／又は他のインタフェース）を含み得る。

【００７８】

また、通信インタフェース１６は、ワイドエリアネットワーク（ＷＡＮ）、無線通信ネットワーク、及び／又は他の任意で適切な通信ネットワークにおける通信を可能とするインタフェースも含み得る。通信ネットワークは、有線の通信リンク、及び／又は無線の通信リンクを含み得る。通信ネットワークは、上記のネットワーク及び／又は他の形式のネットワークの任意の組合せを含み得る。通信ネットワークは、一つ以上のルータ、スィッチ、アクセスポイント、無線アクセスポイント、及び／又は類似物を含み得る。一つ以上の構成において、通信ネットワークは、任意の近傍車両及び車両１１と、任意の近傍の路側の通信ノード及び／又はインフラとの間の通信を許容し得る、車両対全て（Ｖ２Ｘ）（車両対インフラストラクチャ（Ｖ２Ｉ）技術及び車両対車両（Ｖ２Ｖ）技術を含む）の技術を包含し得る。

【００７９】

ＷＡＮネットワーク環境において使用されたとき、コンピュータ処理システム１１４は、ネットワーク（例えば、インターネット）のようなＷＡＮ上での通信を確立するためのモデム又は他の手段を含み（又は、それに対して作用的に接続され）得る。無線通信ネットワークにおいて使用されたとき、コンピュータ処理システム１１４は、無線ネットワークにおける一つ以上のネットワークデバイス（例えば、基地送受信ステーション）を介して無線コンピュータ処理デバイス（図示せず）と通信するための一つ以上の送受信機、デジタル信号プロセッサ、及び付加的な回路機構並びにソフトウェアを含み（又は、それに対して作用的に接続され）得る。これらの構成は、種々の外部情報源から定常的な情報の流れを受信する種々の態様を提供する。

【００８０】

車両１１は、コンピュータ処理システム１１４並びに他の車両システム及び／又は構成要素と作用的に通信し且つコンピュータ処理システムから受信した制御命令に応じて作用し得る、種々の起動可能なサブシステム及び要素を含み得る。種々の起動可能なサブシステム及び要素は、（例えば、ＡＣＣ及び／又は車線維持などの）いずれの自律的の走行支援システムが起動されているのか且つ／又は車両が完全自律モードで駆動されているのかといった所定の走行状況のような要因に依存して、手動的又は（コンピュータ処理システム１１４により）自動的に制御され得る。

【００８１】
操舵システム１８は、車両ホイール、ラック及びピニオン操舵ギア、操舵ナックル、及び／若しくは車両１１の方向を調節すべく作用可能であり得る他の任意の要素（コンピュータシステムで制御可能な任意の機構又は要素を含む）、又は要素の組み合わせを含み得る。動力システム２６は、車両１１に動力運動を提供すべく作用可能な構成要素を含み得る。例示的な実施形態において、動力システム２６は、エンジン（図示せず）、（ガソリン、ディーゼル燃料、又は、ハイブリッド車両の場合には一つ以上の電気バッテリのような）エネルギ源、及び、変速機（図示せず）を含み得る。制動システム２２は、車両１１を減速すべく構成された、要素及び／又はコンピュータシステムで制御可能な任意の機構の任意の組合せを含み得る。スロットルシステムは、（例えば、加速ペダル、及び／又は例えばエンジンの作動速度を制御することで車両１１の速度を制御するように構成された任意のコンピュータシステム制御可能な機構などの）要素及び／又は機構を含み得る。図３は、車両に組み込まれ得る車両サブシステムの僅かな例１８、２０、２２、２６を示している。特定の車両は、これらのシステムの一つ以上、又は示されたシステムの一つ以上に加えて他のシステム（図示せず）の一つ以上を組み込み得る。

【００８２】
車両１１は、コンピュータ処理システム１１４、センサシステム２８、起動可能なサブシステム１８、２０、２２、２６及びその他のシステム及び要素が、コントローラエリアネットワーク（ＣＡＮ）バス３３などを用いて互いに通信できるように構成され得る。ＣＡＮバス及び／又は他の有線又は無線メカニズムを介して、コンピュータ処理システム１１４は、さまざまなシステム及び構成要素に対しメッセージを伝送する（及び／又はそこからメッセージを受信する）ことができる。代替的には、本明細書中に記載の要素及び／又はシステムのいずれかは、バスを使用することなく互いに直接接続され得る。同様に、本明細書中に記載の要素及び／又はシステム間の接続は、別の物理的媒体（例えば有線接続）を通したものであり得るか、又は接続は無線接続でもあり得る。図３は、コンピュータ処理システム１１４、メモリ１５４及び通信インターフェース１６などの車両１１のさまざまな構成要素を車両１１に組込まれているものとして示しているが、これらの構成要素の１つ以上は車両１１とは別個に組付ける又は付随させることのできるものである。例えば、メモリ１５４は、一部が又は全部が車両１１とは別個に存在することができる。こうして、車両１１は、別個に又は一緒に位置設定可能なデバイス要素の形で提供され得る。車両１１を作り上げるデバイス要素は、有線又は無線で、共に通信可能に連結され得る。こうして、別の態様において、本明細書中で説明されるように、コンピュータ処理システム１１４は、車両の操作を行なう目的で車両を自律的に制御するために使用可能な制御ポリシーを最適化するように構成され得る。コンピュータ処理システム１１４は、コンピュータ処理システム１１４の操作を制御するための１つ以上のプロセッサ１５８と、１つ以上のプロセッサにより使用可能なデータ及びプログラム命令を記憶するためのメモリ１５４とを含むことができる。メモリ１５４は、１つ以上のプロセッサによって実行された時点で、１つ以上のプロセッサ１５８に、ａ）システムに関わる受動的に収集されたデータを受信させ；ｂ）車両についての到達コストを推定するために使用可能なＺ値関数を決定させ；ｃ）コンピュータ処理システム内のｃｒｉｔｉｃネットワークにおいて：Ｚ値関数及び受動的に収集されたデータのサンプルを使用してＺ値を決定させ；受動的に収集されたデータのサンプルを用いて最適なポリシー下で平均コストを推定させ；ｄ）コンピュータ処理システム内のａｃｔｏｒネットワークにおいて、受動的に収集されたデータ、システムについての制御ダイナミクス、到達コスト及び制御ゲインを用いて制御ポリシーを修正させ；ｅ）推定平均コストが収束するまで、ステップ（ｃ）及び（ｄ）を反復的に繰返させる；コンピュータコードを記憶するように構成され得る。

【実施例】

【００８３】
図４及び５を参照すると、本明細書中に記載のｐＡＣ強化学習方法の一実施形態の一実施例において、自律的高速道路合流操作がシミュレーションされている。最低予想累積コ

ストで車両の操作を行なう目的で車両を制御するために構成され最適化された制御ポリシーを学習するように、高速道路合流操作に関連する受動的に収集されたデータが、前述のように処理される。その後、高速道路合流操作を行なうため、学習された制御ポリシーにしたがって車両を制御することができる。

【００８４】

高速道路合流操作は、４次元状態空間と１次元行動空間とを有し得る。車両環境ダイナミクスの受動的ダイナミクスＡ（x_t）及び車両制御ダイナミクスＢ（ｘ）は、以下のように表現可能である：

【数２５】

$$\mathbf{x} = [dx_{12}, dv_{12}, dx_{02}, dv_{02}]^{\mathsf{T}},$$

$$\mathrm{A}(\mathbf{x}) = [dx_{12}, 0, dv_{02}, +0.5\alpha_0(\mathbf{x})\Delta t, \alpha_0(\mathbf{x})]^{\mathsf{T}}$$

$$B(\mathbf{x}) = [0.5\Delta t, 1, 0, 0]^{\mathsf{T}}, \mathrm{C}(\mathbf{x}) = [0, 2.5, 0, 2.5]^{\mathsf{T}}$$

$$\alpha_0(\mathbf{x}) = \alpha \frac{v_2^{\beta}(-dv_{02})}{-dx_{02}^{\gamma}}, \Delta t = 0.1[\mathrm{sec}]$$

ここで、下付き文字「０」は、高速道路の最も右側のレーン上の合流車両の後方にある「０」と標識付けされた車両（「後続車両」という）を意味し、下付き文字「１」は、ランプＲＲ上の合流する自動運転車両である「１」と標識付けされた車両を意味し、下付き文字「２」は、高速道路の最も右側のレーン上の合流車両１の前方にある「２」と標識付けされた車両（「先行車両」というばれる）を意味する。dx_{12}に及びdv_{12}は、先行車両と合流車両との相対的位置及び速度を意味し、項α_0（ｘ）は、後続車両０の加速度を表わす。パラメータα、β及びγは、交通環境内の人間の運動挙動に調整することのできるモデルパラメータ（例えば、Ｇａｚｉｓ－Ｈｅｒｍａｎ－Ｒｏｔｈｅｒｙ（ＧＨＲ）の車追従モデル内で使用されているようなもの）である。実施例のためには、先行車両が定速Ｖ２＝３０メートル／秒で運転されていること、後続車両についての車両制御ダイナミクスが公知であること、が仮定されている。後続車両の速度が先行車両の速度より遅い場合、（$d\upsilon_{02}$＜０）、α＝１．５５、β＝１．０８、γ＝１．６５であり、そうでない場合にはα＝２．１５、β＝－１．６５、γ＝－０．８９である。

【００８５】

状態コストｑ（ｘ）を以下のように表現することができる：

【数２６】

$$q(\mathbf{x}) = k_1\left(1.0 - \exp\left(-k_2\left(1 - \frac{2dx_{12}}{dx_{02}}\right)^2 - k_3(dv_{10})^2\right)\right)$$

ここで、k_1、k_2及びk_3は、状態コストのための重みであり；合流車両がランプ上で（すなわちdx_{12}＜０及びdx_{12}＞dx_{02}の条件下で）後続車両と先行車両の間にある場合、k_1＝１、k_2＝１０及びk_3＝１０であり；そうでない場合k_1＝１０、k_2＝１０及びk_3＝０である。状態コストについての重みｋ１、ｋ２、ｋ３は、割当てられるか又は手動で調整され得る。代替的には、逆強化学習を用いて、収集されたデータセットから状態コスト関数を学習することができる。コストは、後続車両と同じ速度で後続車両と先行車両の間で中間に合流するように自動運転車両を誘起するように設計される。初期状態は、－１００＜dx_{12}＜１００メートル、－１０／dv_{12}＜１０メートル／秒、－１００＜dx_{02}＜－５メートル及び－１０＜$dx_{0.2}$＜１０メートル／秒の範囲内でランダムに選択された。Ｚ値を近似するために、ガウス放射基底関数を使用した：

【数２７】

$$fi(\mathbf{x}) = \exp\left(-\frac{1}{2}(\mathbf{x}-m_i)^{\top} Si(\mathbf{x}-m_i)\right)$$

ここで、ｍｉ及びＳｉはｉ番目の放射基底関数のための平均及び逆共分散である。高速道路合流のシミュレーションのためには、１状態次元あたり８個の値で構成されたグリッドの頂点に平均が設定された４０９６個のガウス放射基底関数で、Ｚ値を近似した。基底の標準偏差は、各次元における最も近い２つの基底の間の距離の０．７であった。ρ（ｘ）の実際値が実施例において恒常であることから、制御ゲインρ（ｘ）を推定するためにｇ（ｘ）＝１の値を使用した。最適ポリシーは、上述のように、方程式（７）を用いて決定した。該方法は、受動的ダイナミクスをシミュレートすることによって収集された１００００個のサンプルからポリシーを最適化した。図５は、本明細書中に記載された方法により決定された連続する制御入力を用いて、１２５の異なる初期状態から出発して、（収束に必要とされる反復数として表現される）３０秒以内での合流成功率を示す。

【００８６】

状態コスト関数は、特定の合流状況に適合させるように設計又は調整され得る。１つ以上の実施形態において、特定の合流状況のために状態コスト関数を学習する目的で逆強化学習を使用するように、コンピュータ処理システムをプログラミングすることができる。

【００８７】

開示を読了した時点で当業者であれば認識するように、本明細書中に記載のさまざまな態様を、方法、コンピュータ処理システム又はコンピュータプログラムプロダクトとして具体化することができる。したがって、これらの態様は、完全にハードウェアの実施形態、完全にソフトウェアの実施形態又は、ソフトウェアとハードウェアの態様を組合せた実施形態の形をとることができる。その上、このような態様は、本明細書中に記載の機能を実行するための記憶媒体内又は上に具体化された、コンピュータ可読プログラムコード又は命令を有する１つ以上のコンピュータ可読記憶媒体によって記憶されたコンピュータプログラムプロダクトの形をとることができる。さらに、本明細書中に記載のデータ、命令又は事象を表わすさまざまな信号を、金属線、光ファイバ、及び／又は無線伝送媒体（例えば空気及び／又は空間）などの信号伝導媒体を通して走行する電磁波の形で発信元と宛先との間で移送することができる。

【００８８】

図中のフローチャート及びブロック図は、さまざまな実施形態に係るシステム、方法及びコンピュータプログラムプロダクトの考えられる実装のアーキテクチャ、機能性及び操作を例示する。この点において、フローチャート又はブロック図内の各ブロックは、規定の論理関数を実装するための１つ以上の実行可能な命令を含むモジュール、セグメント又はコード部分を表わし得る。同様に、いくつかの代替的実装において、ブロック内で指摘された機能が、図中に指摘された順序以外で発生し得るという点にも留意すべきである。例えば、連続して示されている２つのブロックを、実際には実質的に同時に実行することができ、あるいは、関与する機能性に応じてブロックを逆の順序で実行することができる場合もある。

【００８９】

本明細書中で使用される「ａ」及び「ａｎ」なる用語は、１以上として定義される。本明細書中で使用される「複数（plurality）」なる用語は、２以上として定義される。本明細書中で使用される「別の（another）」なる用語は、少なくとも第２以上のものとして定義される。本明細書中で使用される「～を含む（including）」及び／又は「～を有する（having）」なる用語は、含む（comprising）（すなわちオープンランゲージ）として定義される。本明細書中で使用される「～と～の少なくとも１つ「at least one of... and...」なる言い回しは、付随する列挙された品目のうちのいずれか及びその１つ以上の

考えられる全ての組合せを意味し包含する。一例として、「Ａ、Ｂ及びＣの少なくとも１つ」なる言い回しは、Ａのみ、Ｂのみ、Ｃのみ、又はその任意の組合せ（例えばＡＢ、ＡＣ、ＢＣ又はＡＢＣ）を含む。

【００９０】

上述の詳細な説明においては、その一部を成す添付図面に対する参照が指示されている。図中、類似の符号は、文脈上別段の指示のないかぎり、類似の構成要素を識別する。詳細な説明、図及びクレーム中に記載された例示的実施形態は、限定的なものとして意図されていない。本明細書中で提示された主題の範囲から逸脱することなく、他の実施形態を利用することができ、他の変更を加えることも可能である。本明細書中で一般的に説明され図中に例示されている本開示の態様は、多様な異なる構成で配置、置換、組合せ、分離及び設計することができ、その全てが本明細書中で明示的に企図される。したがって、本発明の範囲を標示するものとしては、以上の明細書ではなくむしろ以下のクレームを参照すべきである。

本開示は以下の態様を含む。

（態様１）車両の操作を行なう目的で車両を自律的に制御するコンピュータ実装型の方法において、

最低予想累積コストで前記車両の操作を実施すべく前記車両を制御するように構成された制御ポリシーを学習するために、前記車両の操作に関連する受動的に収集されたデータに対して、受動的ａｃｔｏｒ－ｃｒｉｔｉｃ強化学習方法を適用するステップと、

前記車両の操作を行なうべく前記制御ポリシーにしたがって前記車両を制御するステップと、を含む方法。

（態様２）

前記車両の操作が、車線内を走行する第２の車両と第３の車両の間で前記車線内に前記車両を合流させる操作であり、前記制御ポリシーが、前記第２の車両と前記第３の車両の間の中間に前記車両を合流させるべく前記車両を制御するように構成されている、上記態様１に記載の方法。

（態様３）

前記車両の操作を行なうべく前記車両を制御するために適応され得る制御ポリシーを受信するステップをさらに含み、受動的に収集されたデータに対して受動的ａｃｔｏｒ－ｃｒｉｔｉｃ強化学習方法を適用する前記ステップが、

ａ）　ｃｒｉｔｉｃネットワークにおいて、前記受動的に収集されたデータのサンプルを用いて最適な制御ポリシーの下でＺ値及び平均コストを推定するステップと、

ｂ）　ｃｒｉｔｉｃネットワークに作用的に連結されたａｃｔｏｒネットワークにおいて、前記ｃｒｉｔｉｃネットワークからの最適な制御ポリシーの下で、前記受動的に収集されたデータのサンプル、前記推定されたＺ値、及び前記推定された平均コストを用いて前記制御ポリシーを修正するステップと、

ｃ）　前記推定された平均コストが収束するまで、ステップ（ａ）～（ｂ）を反復的に繰返すステップと、を含む、上記態様１に記載の方法。

（態様４）

前記Ｚ値がベルマン方程式の線形化版を用いて推定される、上記態様３に記載の方法。

（態様５）

最適なポリシーの下で前記平均コストを推定する前記ステップが、前記制御ポリシーを修正する前記ステップの前に、前記平均コストを更新するステップを含む、上記態様３に記載の方法。

（態様６）

Ｚ値を推定する前記ステップが、

重み付けされた放射基底関数の線形結合を用いてＺ値関数を近似するステップと、

近似されたＺ値関数及び前記受動的に収集されたデータのサンプルを用いてＺ値を近似するステップと、を含む、上記態様３に記載の方法。

（態様７）

重み付けされた放射基底関数の線形結合を用いてＺ値関数を近似する前記ステップが、前記重み付けされた放射基底関数内で使用される重みを最適化するステップを含む、上記態様６に記載の方法。

（態様８）

重み付けされた放射基底関数の線形結合を用いてＺ値関数を近似する前記ステップが、前記重みを最適化する前記ステップの前に、前記重み付けされた放射基底関数内で使用される重みを更新するステップを含む、上記態様７に記載の方法。

（態様９）

前記制御ポリシーを修正する前記ステップが、

制御ゲインを近似するステップと、

前記制御ゲインを最適化して、最適化された制御ゲインを提供するステップと、

前記最適化された制御ゲインを用いて前記制御ポリシーを修正するステップと、を含む、上記態様３に記載の方法。

（態様１０）

前記制御ゲインを最適化する前に、

制御入力を決定するステップと、

前記制御入力、前記受動的に収集されたデータのサンプル及び前記近似された制御ゲインを用いて、行動価値関数の値を決定するステップと、をさらに含む、上記態様９に記載の方法。

（態様１１）

制御ゲインを近似する前記ステップが、重み付けされた放射基底関数の線形結合を用いて前記制御ゲインを近似するステップを含む、上記態様９に記載の方法。

（態様１２）

重み付けされた放射基底関数の線形結合を用いて前記制御ゲインを近似する前記ステップの前に、前記重み付けされた放射基底関数内で使用される重みを更新するステップをさらに含む、上記態様１１に記載の方法。

（態様１３）

操作を行なうようシステムを制御するのに使用可能な制御ポリシーを最適化するコンピュータ実装型方法であって、

前記システムを制御するのに使用可能な制御ポリシーを提供するステップと、

行なうべき操作に関する受動的に収集されたデータに対して受動的ａｃｔｏｒ－ｃｒｉｔｉｃ強化学習方法を適用して、最低予想累積コストで前記操作を行なうように前記システムを制御すべく前記制御ポリシーが操作可能になるように前記制御ポリシーを修正するステップと、を含む方法。

（態様１４）

受動的に収集されたデータに対して受動的ａｃｔｏｒ－ｃｒｉｔｉｃ強化学習方法を適用する前記ステップが、

ａ）ｃｒｉｔｉｃネットワークにおいて、前記受動的に収集されたデータのサンプルを用いてＺ値を推定し、前記受動的に収集されたデータのサンプルを用いて最適なポリシーの下で平均コストを推定するステップと、

ｂ）ａｃｔｏｒネットワークにおいて、前記受動的に収集されたデータのサンプル、前記システムについての制御ダイナミクス、到達コスト及び制御ゲインを用いて前記制御ポリシーを修正するステップと、

ｃ）前記制御ポリシーを修正するのに使用されるパラメータ及び最適なポリシーの下で前記Ｚ値及び前記平均コストを推定するのに使用されるパラメータを更新するステップと、

ｄ）前記推定された平均コストが収束するまで、ステップ（ａ）～（ｃ）を反復的に繰返すステップと、を含む、上記態様１３に記載の方法。

（態様１５）

車両の操作を行なうべく車両を自律的に制御するのに使用可能な制御ポリシーを最適化

するように構成されたコンピュータ処理システムであって、
当該コンピュータ処理システムが、前記コンピュータ処理システムの操作を制御するための１つ以上のプロセッサと、前記１つ以上のプロセッサにより使用可能なデータ及びプログラム命令を記憶するためのメモリとを含み、
前記メモリは、コンピュータコードを記憶するように構成され、該コンピュータコードは、前記１つ以上のプロセッサによって実行された時点で、前記１つ以上のプロセッサに、
ａ）前記車両の操作に関する受動的に収集されたデータを受信させ、
ｂ）前記車両についての到達コストを推定するのに使用可能なＺ値関数を決定させ、
ｃ）前記コンピュータ処理システム内のｃｒｉｔｉｃネットワークにおいて、
ｃ１）前記Ｚ値関数及び前記受動的に収集されたデータのサンプルを使用してＺ値を決定させ、
ｃ２）前記受動的に収集されたデータのサンプルを用いて最適なポリシーの下で平均コストを推定させ、
ｄ）前記コンピュータ処理システム内のａｃｔｏｒネットワークにおいて、前記受動的に収集されたデータ、前記車両についての制御ダイナミクス、到達コスト及び制御ゲインを用いて前記制御ポリシーを修正させ、
ｅ）前記推定された平均コストが収束するまで、ステップ（ｃ）及び（ｄ）を反復的に繰返させる、コンピュータ処理システム。

【図１】

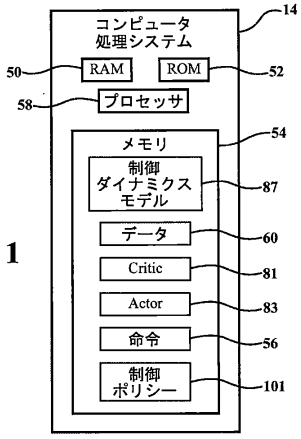


FIG. 1

【図２】

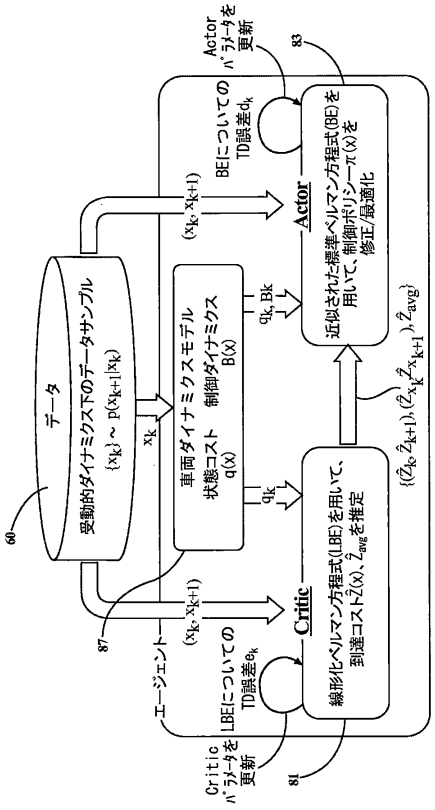
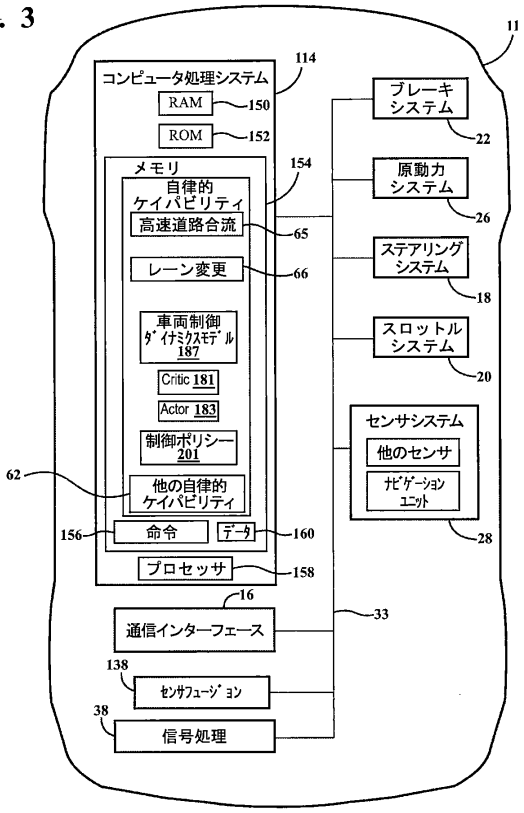


FIG. 2

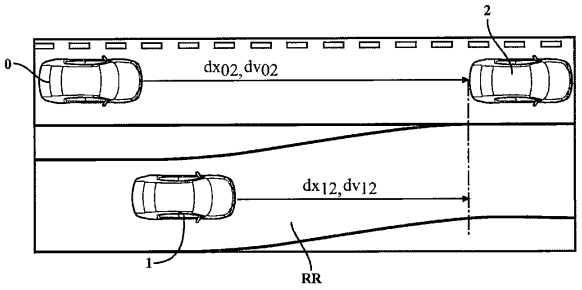
【図３】

FIG. 3



【図４】

FIG. 4



【図５】

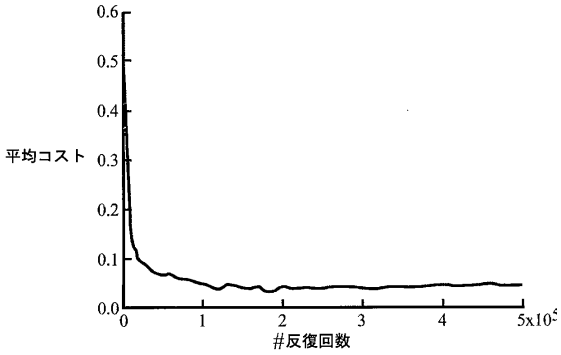


FIG. 5

【図６】

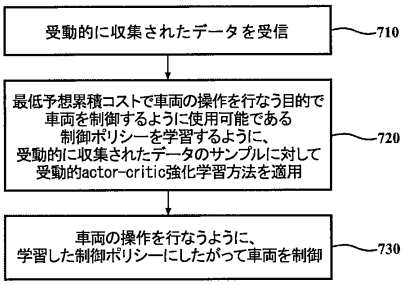


FIG. 6

【図７】

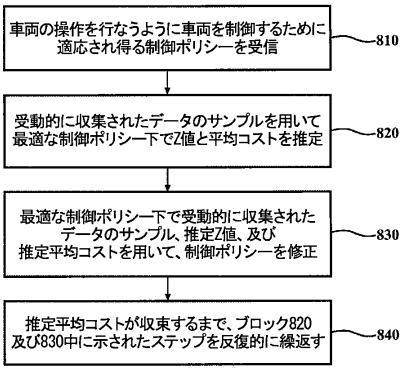


FIG. 7

【図８】

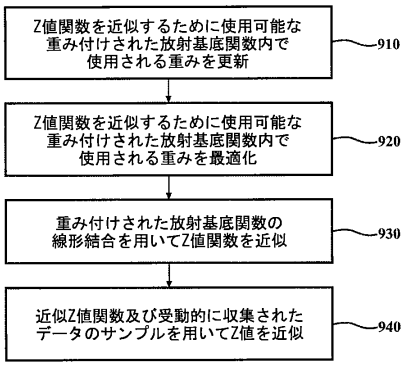


FIG. 8

【図９】

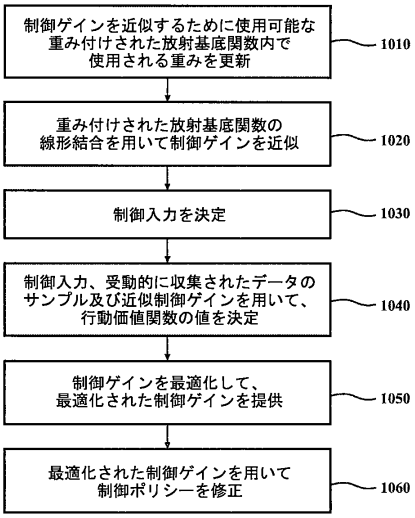

制御ゲインを近似するために使用可能な
重み付けされた放射基底関数内で
使用される重みを更新
1010
重み付けされた放射基底関数の
線形結合を用いて制御ゲインを近似
1020
制御入力を決定
1030
制御入力、受動的に収集されたデータの
サンプル及び近似制御ゲインを用いて、
行動価値関数の値を決定
1040
制御ゲインを最適化して、
最適化された制御ゲインを提供
1050
最適化された制御ゲインを用いて
制御ポリシーを修正
1060

FIG. 9

フロントページの続き

(74)代理人 100123593
弁理士 関根 宣夫
(72)発明者 西 智樹
アメリカ合衆国, ケンタッキー ４１０１８, アーランガー, アトランティック アベニュ ２５, シー／オー トヨタ モーター エンジニアリング アンド マニュファクチャリング ノース アメリカ, インコーポレイティド

審査官 吉村 俊厚

(56)参考文献 米国特許出願公開第２０１８／０３７３２４５（ＵＳ, Ａ１）
米国特許出願公開第２０１９／００３５２７５（ＵＳ, Ａ１）
国際公開第２００４／０６８３９９（WO, Ａ１）
特開２００４－３４８３９４（ＪＰ, Ａ）

(58)調査した分野(Int.Cl., ＤＢ名)
Ｂ６０Ｗ ５０／００
Ｂ６０Ｗ ３０／００
Ｂ６０Ｗ ３０／１０
Ｇ０６Ｎ ９９／００