強化学習によるロボットとの円滑な共同作業

Study on smooth work with robots using reinforcement learning

○学　樋口　慶（東京農工大）　正　Liz Rincon Ardila（東京農工大）
正　Gentiane Venture（東京農工大）

Kei Higuchi, Tokyo University of Agriculture and Technology, keihigu@outlook.jp
Liz Rincon Ardila, Tokyo University of Agriculture and Technology
Gentiane Venture, Tokyo University of Agriculture and Technology

Modern robots are required to interact with humans and changing environments in the real-world as well as to perform different works in the factory. However, the environment changes unpredictably according to the time and actions taken by humans and robots, and humans make different decisions in various situations. To solve these problems with more adaptation and flexibility, this research focuses on reinforcement learning that learns appropriate behavior by trial and error with the environment. In addition, we hypothesize that robots could adapt to various personality of humans. We could express various strategies by changing parameters in reinforcement learning as functions related to the human personality, where the neuroparameters, actions and rewards in the learning strategy can be adapted and modified. We show that during even a simple cognitive learning task experiment, human personality affects the result, and human improved their learning result in a human-robot collaborative learning task.

Key Words: Reinforcement learning, personality, Human-Robot interaction

1.　緒言

現在のロボットには，工場内で制御された環境で作業を行うだけでなく，工場外の環境や人間との相互作用を行うことが求められている[1]．しかし，環境は人間やロボットが起こす行動と時間により逐次予測しない変化をし，また人間はさまざまな場面において異なる意思決定を行う．したがって，未知の環境と人それぞれ異なる行動戦略に適応できるモデルが必要とされている．ここで，環境との試行錯誤を通して適切な行動戦略を獲得する強化学習に注目する[2]．強化学習では，学習する対象は行動を起こすことで環境にはたらきかけ，変化する環境の状態を観測し，また未知の環境で発生することを統一的に比較する指標として，報酬と呼ばれるスカラー値を得る．学習者は，環境の中で行動に応じた報酬の総和を最大化することで適切な行動を学習していく．したがって，環境に関する知識が不十分で，対象へのはたらきかけによって状態が変化するような場合でも最適な行動を学習することができる．また，強化学習のアルゴリズムの中には人間が設定するパラメータが複数存在するが，それらは目先の即時報酬を重視するのか将来的な報酬を重視するのか，未知の行動を探索するのか現在わかっている最適な行動を選択するのか等，様々な意思決定の様子を表現することができる．それらのパラメータを調整することで，異なる行動戦略を持つ人間に対して最適な強化学習モデルを構築できると考えた．

2.　研究目的

本研究では，強化学習アルゴリズムのパラメータを調整することで異なる行動戦略を持つ人間に対して最適な強化学習モデルを構築し，人間と共に未知の環境を探索しながら最適な行動を学習するロボットの開発を行うことを目的とする．そのために，図 1 に示すような流れで実験を行った．まず人間の問題解決のための意思決定の様子が各個人で異なるのかどうかを検証するために，被験者の性格を自己記入式のパーソナリティ診断を用いて取得した．そして認知科学の分野で用いられている簡単な強化学習型のタスクを課し，その学習過程・結果が性格に影響されるのか検証を行った．次にロボット側にもパラメータを変化させながら同様のタスクを課し，パラメータの変化が学習過程・結果に影響を及ぼすのか検証を行った．最後に人間とロボットで交互に同様のタスクを行わせて協力して学習を行ってもらい，学習が向上したかどうかの検証を行った．

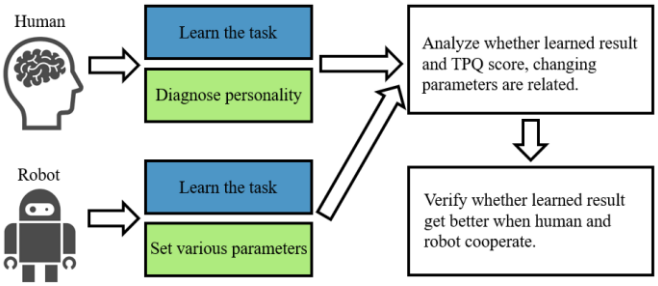


Fig. 1 Flow of the proposed experiment to verify if HRI can improve learning ability of humans

3.　強化学習

本研究では，強化学習のアルゴリズムの 1 つであるQ学習と呼ばれるアルゴリズムを使用する[3]．Q学習では，時間ステップtで環境に行動a_tを起こし，次の環境の状態s_{t+1}と報酬r_{t+1}を観測することで，状態s_tで行動a_tを選択する価値Qを学習する，学習者は行動価値を以下の式で更新し，このQ値が大きい行動を選択するようになる．

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha\left\{r_{t+1} + \gamma * \max_a Q(s_{t+1},a) - Q(s_t,a_t)\right\} \quad (1)$$

ここで，αは学習率，γは割引率と呼ばれ，それぞれ学習の速度，未来の報酬の重要度を調節するパラメータである．また，学習の収束からの離れ具合を表す Temporal Difference (TD)誤差δを以下の式で定義する．

$$\delta = r_{t+1} + \gamma * \max_a Q(s_{t+1},a) - Q(s_t,a_t) \quad (2)$$

最適行動価値があらかじめわかっている場合には，行動価値の大きい行動を選択するような方策を定めればよいが，実際には最適行動価値がはじめからわかっていることはない．したがって学習者は様々な行動を探索する必要があるため，そ

の時点で最良と考える行動以外も選択する必要がある．そこで最大の行動価値を選択する確率 Pを以下の式で求める．

$$P(s_t, a_t) = \frac{\exp(\beta * Q(s_t, a_t))}{\sum_{a \in A} \exp(\beta * Q(s_t, a))} \quad (3)$$

βは逆温度と呼ばれ，行動のランダムさを制御するパラメータである．

強化学習のように，報酬に基づき意思決定を行う様子は生物の行動にも見られ，近年強化学習のアルゴリズムが動物の脳内にも存在しており，意思決定の処理モデルとして説明できるのではないかと注目されている．強化学習中の最適なパラメータは学習問題によって決まり，その設計は設計者の試行錯誤に依存する．脳がこの学習機構を持っているとすると，これらのパラメータを自律的に調節する機構が存在するはずである． Doya らは，脳幹から大脳基底核を含めた脳に投射される神経修飾物質がパラメータ制御に関わっているという仮説を提案している[4]．この仮説では神経伝達物質と強化学習中のパラメータをそれぞれ，学習率αとアセチルコリン，逆温度βとノルアドレナリン，割引率γとセロトニン，TD 誤差δとドーパミンと対応付けている．

4. 3 次元人格理論

Cloninger は，遺伝性の生理的基礎を伴う「気質: temperament」を中心に捉え，行動の解発，維持．抑制を含む 3 つの脳神経システムを仮定した，3 次元人格 (Tridimensional Personality: TDP) 理論を提唱した[5]．理論を構成する 3 つの次元は，新奇性探求（Novelty Seeking），報酬依存（Reward Dependence），損害回避（Harm Avoidance）であり，それぞれドーパミン，ノルアドレナリン，セロトニンの個体的代謝特性に支えられている．表 1 に理論を構成する 3 つの次元と関係する神経伝達物質，強化学習中のパラメータを示す．

Table 1 The relation between TDP, neuromodulation, and parameters in reinforcement learning

3次元人格理論	主要な神経修飾物質	強化学習中パラメータ
新奇性探求	ドーパミン	TD誤差:δ
損害回避	セロトニン	割引率:γ
報酬依存	ノルアドレナリン	逆温度:β

本実験では，これらの構造を測定するために性格診断テスト Tridimensional Personality Questionnaire (TPQ)を用いる．TPQ は新奇性探求 34 項目，損害回避 33 項目，報酬依存 33 項目の計 100 項目からなり，3 つの次元の組み合わせによってパーソナリティの傾向を測定できる．日本語版 TPQ は竹内らによって翻訳され，再英訳による確認作業，予備調査の結果，日本語版 TPQ の適用の可能性が保証された[6].

5. 実験方法

本実験で用いた強化学習型のタスクを図 2 に示す．本タスクの第一段階では，意味的に無関係な 2 組の画像（Object）によってラベル付けされた選択肢が 2 つあり，片方の画像を選ぶことにより第 2 段階の状態（別の 2 つの画像の組）へと遷移する．2 段階目の状態で再度画像を選択することで報酬が得られる．ただし状態遷移は確率的であり，必ず学習者が望む状態に遷移するわけではない．この第一段階から報酬を得るまでの一連の流れをエピソードと呼び，学習者は 50 エピソードで得られる報酬を最大化するような画像を学習する．本タスクでは先述の確率を Lv2（80%，20%）と，Lv3（60%，40%）と 2 種類用意し，それぞれ 50 エピソードずつ，合計 100 エピソード試行した．なお被験者は実験に対するインフォームドコンセントを行い同意の上で実験に参加した．

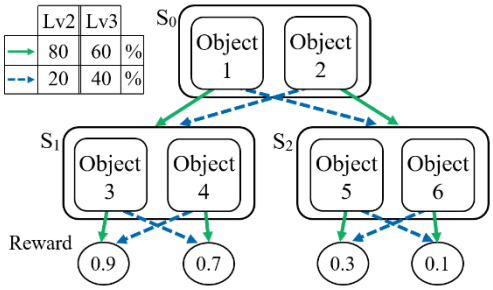


Fig. 2 Conceptual diagram of learning state change

5.1 人間を対象とした実験

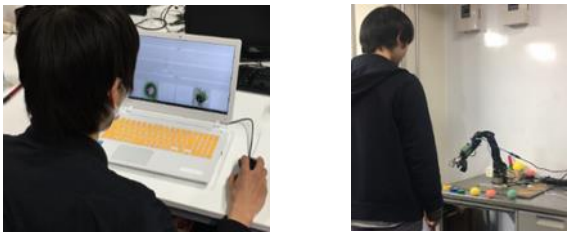
人間の実験は図 3(a)に示すようにモニターとマウスを用いて行った．図 4 に示すようにパソコンのモニターには 2 種類の画像が表示され，被験者はマウスを用いてどちらかの画像を選択した．2 段階の状態遷移の後，モニターには報酬の値が表示された．被験者には行動学習課題を行ってもらった後，被験者の気質を調べるために TPQ に回答してもらった．被験者は，18 人の男性（20 歳から 26 歳まで，平均 22.4 歳，標準偏差 1.34）であった．

5.1 シミュレーション

図 2 に示すタスクを，強化学習アルゴリズムを使用してシミュレーションを行った．本研究では，学習の結果をどの程度行動選択に反映させるか，また過去の行動をどの程度評価するのかという，学習の戦略に直接影響し，かつ 3 次元人格理論と対応するパラメータである逆温度βと割引率γに注目してシミュレーションを行った．学習率は$\alpha = 0.5$ に固定し，逆温度と割引率を変化させて 36 パターンのパラメータの組み合わせでシミュレーションを行った．

5.1 人間とロボットの共同作業

人間を対象とした実験の結果から他と比較して良好に学習できていない被験者を対象に同様のタスクを，奇数エピソードにおいては第一段階の選択は人間，第二段階の選択はロボット，偶数エピソードにおいては第一段階の選択はロボット，第二段階の選択は人間となるように人間とロボットで交互に行動を選択させ，学習が向上するかどうかの検証を行った．実験は，図 3(b)に示すようにロボットアームを用いて行った．



(a) Human experiment (b) Human-Robot experiment

Fig. 3 View of the experiments

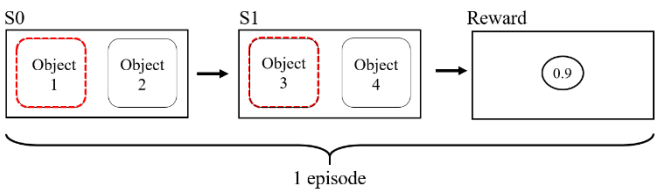


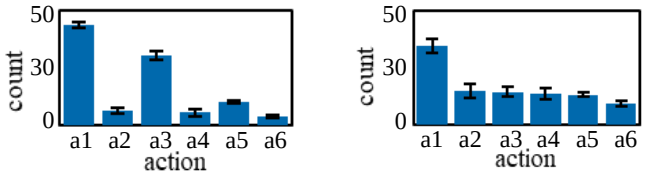
Fig. 4 State change of the experiment

6. 結果・考察

行動学習の結果を評価するための指標を以下に定義する．本研究で用いたタスクで得られる報酬は確率的であるため，学習中に得られた報酬の合計値に加えて，各状態での各行動の選択回数も考慮する．そこで，各状態s_0, s_1, s_2において，最大の報酬が得られる確率が高い画像を選択する行動とそうでない行動をそれぞれ，a1 と a2，a3 と a4，a5 と a6 となるようにラベル付けした．つまり，a1 と a3 をより多く選択する学習者ほど，良好に学習できているといえる．

6.1 人間を対象とした実験の結果

図 5 に人間の被験者全員分の行動選択回数の平均と誤差範囲を示す．その結果，Lv2 の方のタスクでは a1，a3，a5 がそれぞれ a2，a4，a6 よりも多く選択されているので，被験者は良好に学習できているといえる．しかし Lv3 の方のタスクでは，第 1 段階の状態では a1 が a2 よりも多く選択されているが，第 2 段階の状態での行動選択回数は a3～a6 でほとんど変わらず，良好に学習できているとは言い難い．



(a) Lv2　(b) Lv3

Fig. 5 Average of actions by all participants executing the learning task

TPQ の各次元の統計量を表 2 に示す．また図 6 に被験者の TPQ の各次元の値をプロットする．TPQ を構成する 3 次元の変動は正規分布をとり，ほとんどの人は中間的な値を示す[7]．しかし，今回の被験者集団では報酬依存が中間的な値よりも高い傾向にある．報酬依存性が高いと探求的な行動が強化されるため，これは被験者の多くが理系の大学生という偏った集団であることが原因と考えられる．

Table 2 The statistics value of TPQ

TPQ	最小	最大	平均	標準偏差
Reward Dependence	12	28	20.9	4.5
Harm Avoidance	4	30	18.8	7.1
Novelty Seeking	8	28	18.2	5.4

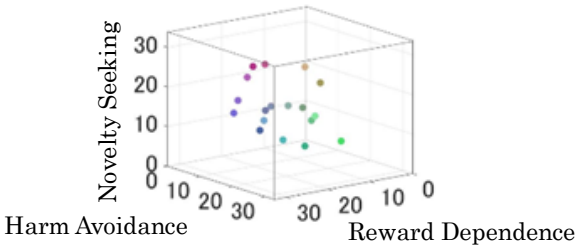
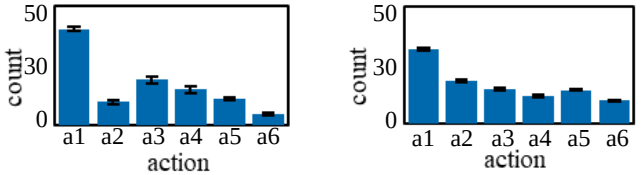


Fig. 6 Result of the TPQ for participants in the human task each point represents a participant

6.1 シミュレーションの結果

図 7 に，逆温度と割引率を変化させた時のパラメータの組み合わせ 36 パターンの行動選択回数の平均と誤差範囲を示す．その結果，タスク Lv2 に関しては人間の被験者に比べて学習がうまく進んでおらず，逆に Lv3 に関しては，a3，a5 の選択がそれぞれ a4，a5 よりも多くなっており，人間の被験者に比べて良好な学習ができている．



(a) Lv2　(b) Lv3

Fig. 7 Average of actions by the robot executing the learning task

ここで，割引率を$\gamma = 0.5$とし，逆温度βを変化（$\beta = 1, 100, 300$）させた時の行動選択回数の変化を図 8 に，逆温度を$\beta = 100$ とし，割引率を変化（$\gamma = 0.1, 0.5, 0.9$）させた時の行動選択回数の変化を図 9 に示す．図 8 より，逆温度が小さいとき（$\beta = 1$）には，a1～a2 と a3～a6 それぞれにおいて，行動選択回数がほぼ等しく，逆温度が大きいときには行動選択回数に差が生じることが確認できる．これは，逆温度が小さいほどQ値にかかわらずランダムな探索行動を選択するようになる結果であると考えられる．また図 9 より，$\gamma = 0.5$の場合と$\gamma = 0.1$の場合を比較したとき，割引率が大きい方が報酬からより離れた行動，つまり a1 と a2 について良好な学習結果が得られていることが確認できる．逆に割引率が大きすぎる場合（$\gamma = 0.9$）には，a3 と a4 について良好な学習結果が得られていない．これは報酬から離れた行動を重視しすぎるあまり，直前の行動をうまく評価できないためであると考えられる．

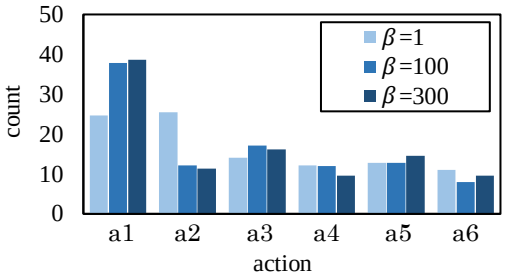


Fig. 8 Result of the robot learning task changing β

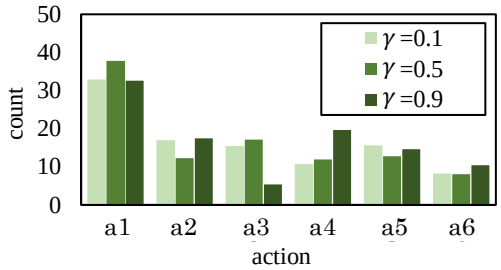


Fig. 9 Result of the robot learning task changing γ

6.3 TPQ のスコアと強化学習のパラメータの分析

TPQ のスコアと強化学習のパラメータが学習結果に影響するかどうかを調べるために，多変量分散分析を行った．従属変数は，TPQ では 3 つの次元のうち，報酬依存と損害回避の 2 つを，パラメータでは逆温度βと割引率γの 2 つを使用する．これは，新奇性探求が対応する強化学習のアルゴリズムにおけるパラメータは TD 誤差δであり，設計者が決定できるパラメータではなく，意思決定の方策に影響しないためであり，学習率αは今回変化させていないためである．一方多変量分散分析における独立変数は，分類や種類を区別する質的データである必要があり，本タスクの評価基準である報酬の合計値や行動選択回数といった量的データが使用できない．そこで，被

験者の学習結果を，「良好に学習できている」，「良好に学習できていない」といった 2 つの水準を持つ質的データに分類する．具体的には，被験者の報酬の合計値，行動選択回数（a1，a3）のそれぞれが，被験者全体の平均値よりも大きければ「良好に学習できている」，小さければ「良好に学習できていない」と分類する．

帰無仮説を「強化学習型タスクの学習結果に TPQ とパラメータの値の効果がない」と設定，有意水準 alpha を 0.05 とし，Roy の検定を用いて検定を行った結果を表 2 に示す．Roy の検定では，出力結果 p 値が有意水準 0.05 よりも小さければ帰無仮説を棄却できる．結果より，人間を対象とした Lv3 のタスクで，獲得した報酬と a3 の選択回数を評価基準とした p 値が有意水準を下回り，難易度の高いタスクにおいて，TPQ のスコアが学習に影響することが考えられる．難易度の低いタスクにおいて帰無仮説が棄却されなかった理由としては，タスク Lv2 では学習難易度が Lv3 と比較して低いことで，被験者間で異なる行動戦略の影響が少なかったためであると考えられる．

Table 2 Statistical result of p-value for human learning and robot learning using MANOVA

	Lv2			Lv3		
	reward	a1	a3	reward	a1	a3
Human	0.586	0.470	0.507	0.044	0.365	0.031
Robot	0.098	1.84×10^{-4}	0.149	0.389	0.141	0.770

6.4 人間とロボットの共同作業の結果

タスク Lv3 において，他と比較して学習の結果が良くなかった，特に探索的な行動が少なく誤った行動をとり続けた被験者 3 人を対象に，ロボットと交互にタスクを行ってもらった．ロボットのパラメータ調整について，多変量分散分析の結果ではパラメータの変化に対する学習結果の影響はないと考えられるが，逆温度βを小さくすると探索的な行動を行うことは図 8 から確認できる．ロボットのパラメータ調整は被験者に探索的な行動を促すために，βの値が小さく，かつシミュレーションで良好な学習結果が得られたパラメータ$\alpha = 0.5, \beta = 10, \gamma = 0.5$を用いた．行動選択回数の平均と標準誤差を，人間のみで行った時の平均と標準誤差と共に図 10 に示す．その結果， a3～a6 の選択において大きな向上が見られた．これは，ロボット側がよりランダムな行動を選択し，探索的な行動を促したためであると考えられる．

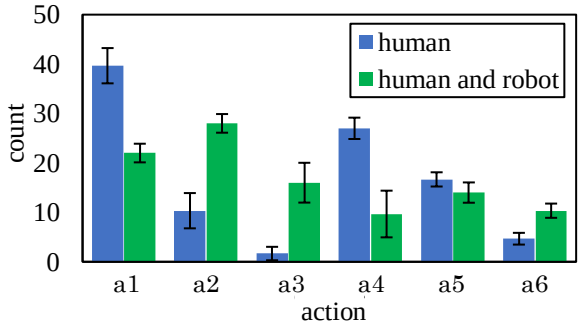


Fig. 10 The average number of actions in human-robot collaboration during the learning task

7. 結言

本研究では実環境において人間と相互作用を行うためのロボットを想定して，異なる意思決定の様式を持つ人間と共に未知の環境に対して自律的に学習するアルゴリズムを構築することを目的とした．そのためにまず人間の性格を自己記入式のパーソナリティ診断を用いて取得し，そして認知科学の分野で用いられている簡単な強化学習型のタスクを課し，その学習過程・結果が性格に影響されるのか検証を行った．次にロボット側にもパラメータを変化させながら同様のタスクを課し，パラメータの変化が学習過程・結果に影響を及ぼすのか検証を行った．最後に人間とロボットで交互にタスクを行わせ，学習が向上したかどうかの検証を行った．その結果，学習の方策，戦略が重要となる難易度の高い強化学習型の学習タスクにおいて s，学習結果に，強化学習のパラメータと関連する神経伝達物質の個体的代謝特性に支えられた性格理論である TPQ 各次元の値の効果があることがわかった．一方ロボットのシミュレーション結果では，簡単なタスクに関しては人間の結果の方が優れていたが複雑なタスクに関してはロボットの方が若干優れた学習結果となった．しかし多変量分散分析の結果より，強化学習型のタスクの学習結果にパラメータの変化の効果はないと考えられる．最後に，他と比較して学習の結果が良くなかった，特に探索的な行動が少なく誤った行動をとり続けた被験者に対して，探索的な行動を多く行うロボットと交互にタスクを行ってもらった結果では，最終的に報酬が得られる行動選択において大きな向上が見られた．これは，ロボット側がよりランダムな行動を選択し，探索的な行動を促したためであると考えられる．

8. 今後の展望

本研究では 1 つの行動に固執する被験者に対して探索的な行動を多く行うロボットと共同でタスクを行ったが，最適な行動は人の価値観やタスクによって変化する．よって，異なる性格に適応するモデルを検討するためには性格とパラメータの組み合わせを増やして検討を行うことが必要だと考えられる．また本実験で用いた強化学習型のタスクは二者択一の簡単な課題であったが，実際にロボットが日常生活において人間と相互作用を行うにはさらに複雑な問題を解決しなければならない[8]．そこで本実験で行った検証を人間とロボットの複雑で物理的な相互作用にも行う必要があると考える．

参考文献

[1] Khamassi, M., “Robot cognitive control with a neurophycologically inspired reinforcement learning model,” *Frontiers in neurorobotics*, vol.5, no1, pp.1-14, 2011.

[2] Sutton, R., Barto, A., *Reinforcement learning: An introduction*, pp.1-3, MIT press, 1998.

[3] 佐々木隆宏, “強化学習型タスクにおける人間の行動決定に関する研究,” Sensing and perception, vol.12, pp-77-84, 2005.

[4] Doya, K., “Metalearning and neuromodulation,” *Neural Networks*, vol.15, no.4-6, pp.495-506, 2002.

[5] Cloninger, C., “A systematic method for clinical description and classification of personality variants: A proposal,” *Archives of general psychiatry*, vol.44, no.6, pp.573-588, 1987.

[6] 竹内美香, “Cloninger の 3 次元人格 (TPQ) 理論および日本語版 Tridimensional Personality Questionnaire (TPQ)”, 精神科診断学, vol.3, pp.491-505, 1992.

[7] 木島信彦, “Cloninger の気質と性格の 7 次元モデルおよび日本語版 Temperament and Character Inventory (TCI)”, 精神科診断学, vol.7, no.3, pp.379-399, 1996.

[8] Morimoto, J., Atkeson, G., “Minimax differential dynamic programming: An application to robust biped walking,” *Advances in neural information processing systems*, vol.15, pp.1563-1570, 2003.