地震時の振動制御における強化学習エージェントの行動安定化に関する研究

4年 #組#番　## ##

明治大学　理工学部　機械情報工学科　機械力学研究室

指導教員　松岡　太一

1.　はじめに

先行研究では，MR 流体を活用した可変慣性型振動抑制装置(VSD: Vibration Suppression Device)が開発された．これは強力な磁界を加えて慣性質量を変化させることで，見かけの質量が変化し，固有振動数を変えることが出来る．磁界を切り替える制御には DQN(Deep Q-Network)による強化学習エージェントが用いられたが，同一地震波による加振応答実験を複数回行った際には，実験ごとに慣性質量を変化させるタイミングにズレが生じ，再現性の低い不安定な制御となった．したがって，本研究では，強化学習エージェントが同一地震波による加振応答シミュレーションにおいて，実験環境を想定してノイズや初期値を微妙に変更しても慣性質量が切り替わるタイミングが同じになるような再現性のある制御を目指す．その上で，その強化学習エージェントの行動選択を模倣したルールベース制御を設計し，強化学習エージェントの不安定さを取り除いた安定した振動制御を目的とする．

2.　実験装置とシミュレーション方法及び理論

2.1　実験装置

本研究に用いる振動抑制装置は MR 流体を活用した可変慣性モーメント型振動抑制装置である．一自由度ばねマス系にこの VSD を取り付け，地面が地震波によって加振される状況を考える．このときの振動系の運動方程式は，

$$(m + m_e)\ddot{u} + c\dot{u} + ku = -m\ddot{z} \qquad (1)$$

と記述できる．ここで，m は質点の質量，m_e は VSD によって切り替わる慣性質量，c はダンパの粘性減衰係数，k はばね定数，z は地面の絶対変位，u は地面から見た質点の相対変位を表す．

2.2　強化学習手法

本研究では，強化学習の手法として PPO(Proximal Policy Optimization)を採用する．PPO は方策を慎重に更新しつつ過学習を回避する，安定かつ効率的な学習手法であり，DQN よりもロバスト性が高いという特徴がある．この特徴から，ノイズや初期値に左右されにくい強化学習エージェントを設計することを目的とする．PPO の欠点としてハイパーパラメータのチューニングが難しいことが挙げられる．何度も調整を行った結果，Clip Factor と GAE Factor というパラメータを 0.035，0.75 にして学習を行うことにする．

2.3　ニューラルネットワーク

強化学習に用いるニューラルネットワークの入力層には，地面の加速度，システムの加速度，システムの相対変位を与える．ここで，各値は観測直後の値のみでなく，過去 10 ステップの値を入力層に与えることにする．つまり入力層のニューロンの数は 30 個となる．これにより，観測値の大きさだけでなく変化率を考慮し，適切な行動選択を期待する．更に，ニューラルネットワークの中間層は，ニューロンが 4 つの層を 2 層用意する．これはニューロンの数がかなり少ないと言えるが，強化学習エージェントの表現力を出来るだけ減らし，ノイズが含まれた観測値の微妙な変化に対する過度な制御を抑制するためである．

2.4　行動安定化

強化学習エージェントが 1 ステップごとに慣性質量を切り替えるように行動選択をした場合，それを定式化することは非常に困難である．そこで，昨年度の卒業研究と同様，慣性質量の切り替えを行った場合に 26 ステップ間その値を維持し続けることにする．

2.5　方策を模倣するルールベース制御の設計

加振応答シミュレーションにおいて強化学習エージェントは，システムの相対変位が大きいときに慣性質量が大きくなるように切り替えを行っている．このことから，相対変位が小さいときは慣性質量を小さく，相対変位が大きいときは慣性質量を大きくするような閾値制御を設計する．このとき，慣性質量が小さいときは閾値を小さく，慣性質量が大きいときは閾値を大きくなるように設定し，慣性質量の切り替えが起こりやすいようにする．次に，図 1 のように，閾値制御により慣性質量を決定してから 26 ステップ行動維持するブロックの間に，それより少ないステップ数だけ行動維持するブロックを入れる．これにより，強化学習エージェントを使用したときの慣性質量の切り替えの様子に近付ける．

2.6　シミュレーション方法

同一地震波による応答加振実験において強化学習エージェントの行動選択に再現性が見られるかどうかを判断するために，システムの相対変位と相対速度の初期値とセンサノイズをシミュレーション毎に微妙に変えながら 100 回シミュレーションを行う．加振に使用する地震波は最大加速度約 4.74m/s^2 で 13 秒間分のエルセントロ波である．強化学習に使用する地震波も同様のものとする．その後，ルールベース制御においても同様のシミュレーションを行い，最大加速度と最大相対変位，及び慣性質量を切り替えるタイミングを比較，考察する．

3. シミュレーション結果と考察

3.1 シミュレーション結果

初期値とセンサノイズを変えながら強化学習エージェントとルールベース制御でそれぞれ 100 回シミュレーションを行い，慣性質量を切り替える様子を最初の 25 回分だけ重ねて表示したものを図 2 に示す．そして，この 100 回のシミュレーションで得られた 100 個の最大加速度と最大相対変位をプロットした散布図を図 3 に示す．また，強化学習エージェントとルールベース制御において慣性質量を切り替えるタイミングがどれだけ近いかを評価するために，強化学習エージェントでの 1 回目のシミュレーションとルールベース制御での 1 回目のシミュレーションで慣性質量の値が異なる時間を数え，その計算を 100 回目のシミュレーションまで行い，100 回分の平均値を算出する．その時間は，シミュレーション全体で約 0.94s，11.5 秒までで約 0.44s であった．図 2(a)を見ると，強化学習エージェントは初期値とセンサノイズが変わりながら同一の地震波によって加振されても慣性質量を切り替えるタイミングがほとんど変わっていないことが分かる．このことから，本研究で設計した強化学習エージェントは行動選択に再現性があると言える．図 2(a)と図 2(b)を比較すると，強化学習エージェントとルールベース制御において慣性質量を切り替えるタイミングがほぼ同じであると言える．また 3.1 で示したようにシミュレーション時間の 11.5 秒までで慣性質量の値が異なる時間の平均値が約 0.44s であったということは，慣性質量を 1 回切り替えるときのタイミングのズレは約 0.02s しかないということであり，定量的にもルールベース制御が強化学習エージェントの行動をある程度模倣出来ていると言える．図 3 を見ると，初期値とセンサノイズが変わったときにルールベース制御の方が最大加速度と最大相対変位の散らばり具合が大きくなっていることが分かる．このことから，本研究で設計したルールベース制御はまだ不完全であり，強化学習エージェントの行動選択を模倣する余地はまだまだあると言える．

4. まとめ

1) 強化学習の手法として PPO を採用し，ニューラルネットワークは入力層のニューロンの数を増やして中間層のニューロンの数を減らし，慣性質量を 26 ステップだけ維持した強化学習エージェントは，初期値やセンサノイズの影響を減らし，行動選択に再現性が得られた．
2) 相対変位に着目した閾値制御に，慣性質量を維持するブロックを増やしたルールベース制御は，強化学習エージェントの行動選択をある程度模倣することが出来た．

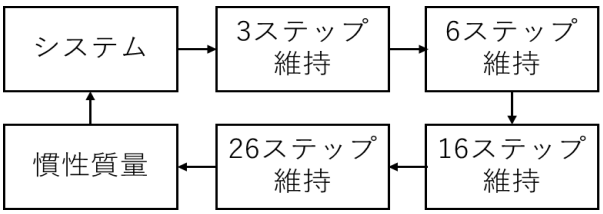
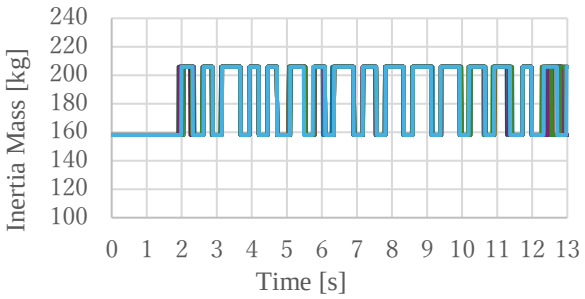
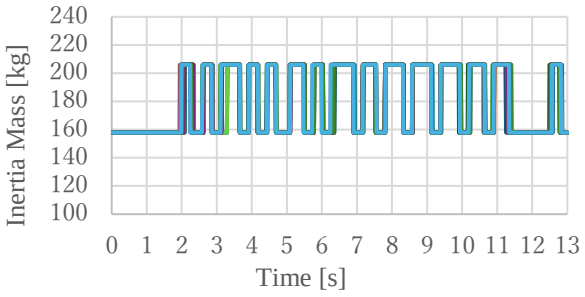


Fig.1 Inertia Mass Decision Flow

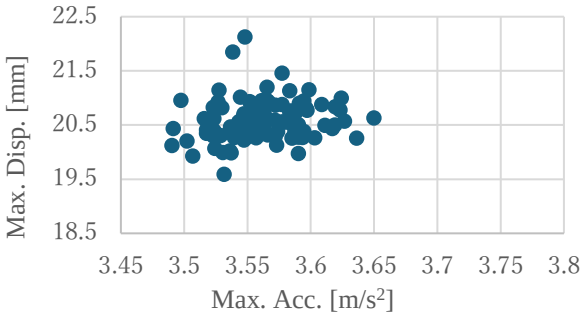


(a) Reinforcement Learning Agent

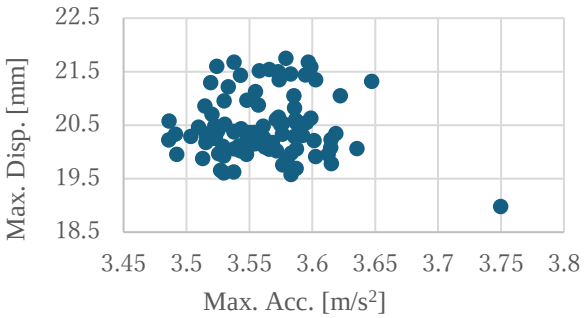


(b) Rule-Based Control

Fig.2 Switching Inertia Mass



(a) Reinforcement Learning Agent



(b) Rule-Based Control

Fig.3 Scatter Plot of Maximum Acceleration and Maximum Relative Displacement