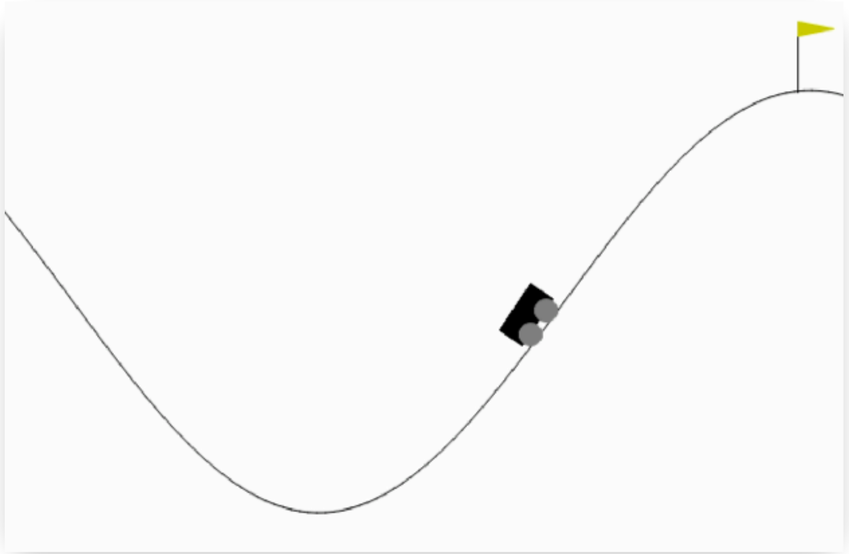
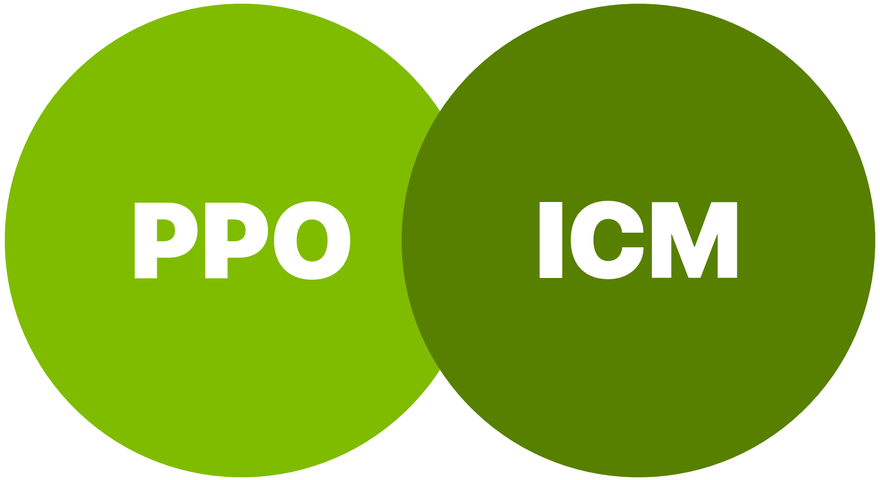
# PPO と ICM を用いた
# 報酬のスパースな環境における強化学習

電気通信大学 情報理工学域 I 類 メディア情報学プログラム 4年
**つまみ (@TrpFrog)**

# 概要

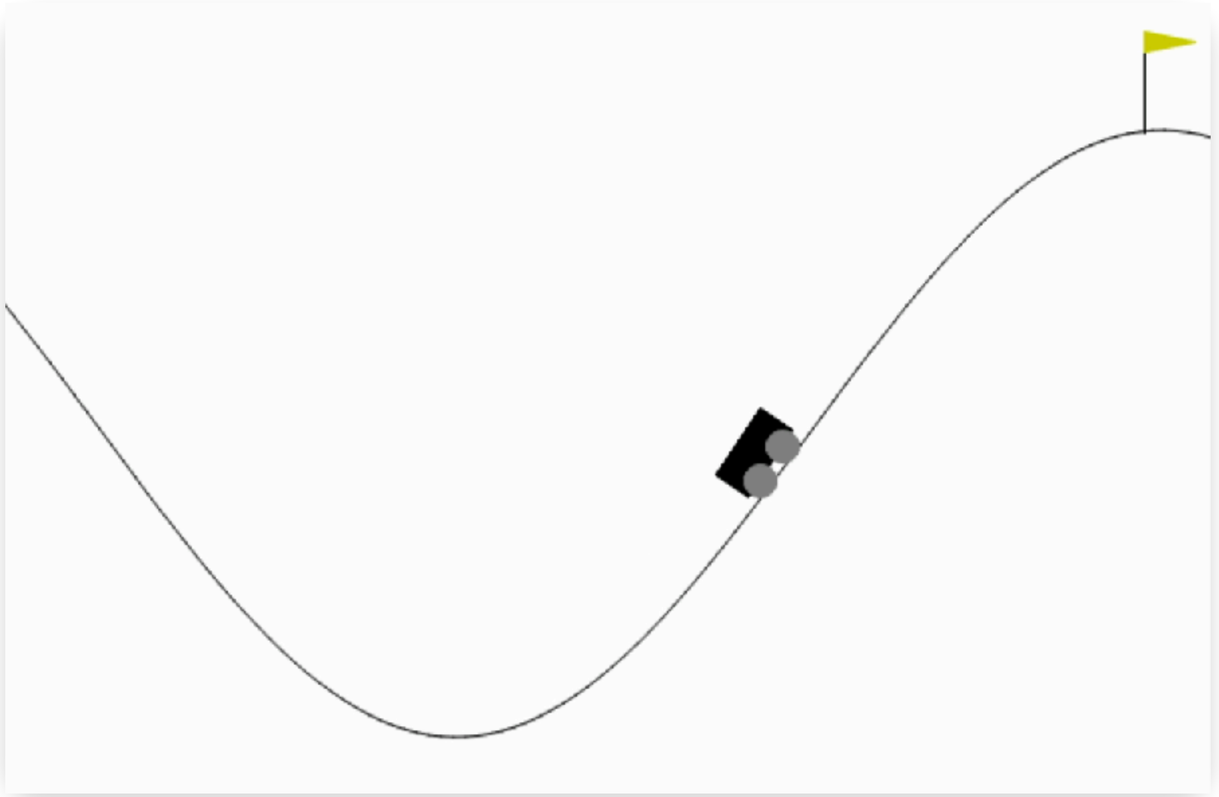
＼ゴールしない限り報酬の合計が -200／






探索の手掛かりとなる報酬が得られにくい環境において
"好奇心" で探索を促す ICM (intrinsic curiosity module) の
再現実装を行い, PPO と組み合わせてその有効性を検証する

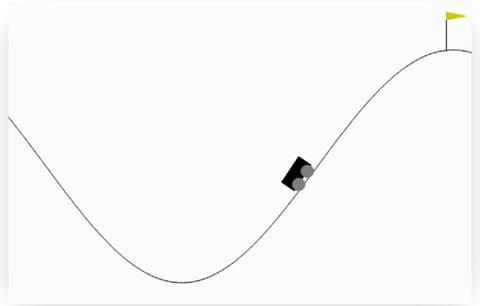
# MountainCar-v0



**MountainCar-v0: 車を左右に動かし助走をつけて山の頂上を目指す**

単純に右に進むだけではなくて助走をつけなければ登れないのがポイント

# MountainCar-v0



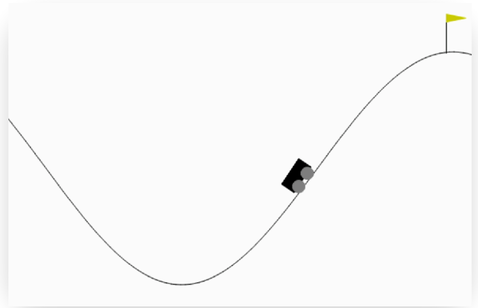
**終了条件:** $x \geq 0.5$ となる (旗に到達する) or 最大ステップ数に到達

**報酬:** 状態によらず各ステップ -1 (早くゴールした方が報酬が増える)

**ゴールに到達できない限り同じ報酬を取り続ける**

**報酬のスパース性**

# MountainCar-v0



**最大ステップ数:** 200

短過ぎてランダム方策ではゴールに到達できない

学習不可能

**何らかの良い中間報酬が欲しい**

# ICM

## Intrinsic Curiosity Module

# Intrinsic Curiosity Module

"Curiosity-driven Exploration by Self-supervised Prediction" Pathak et al., ICML2017

$$r_t = r_t^i + r_t^e$$

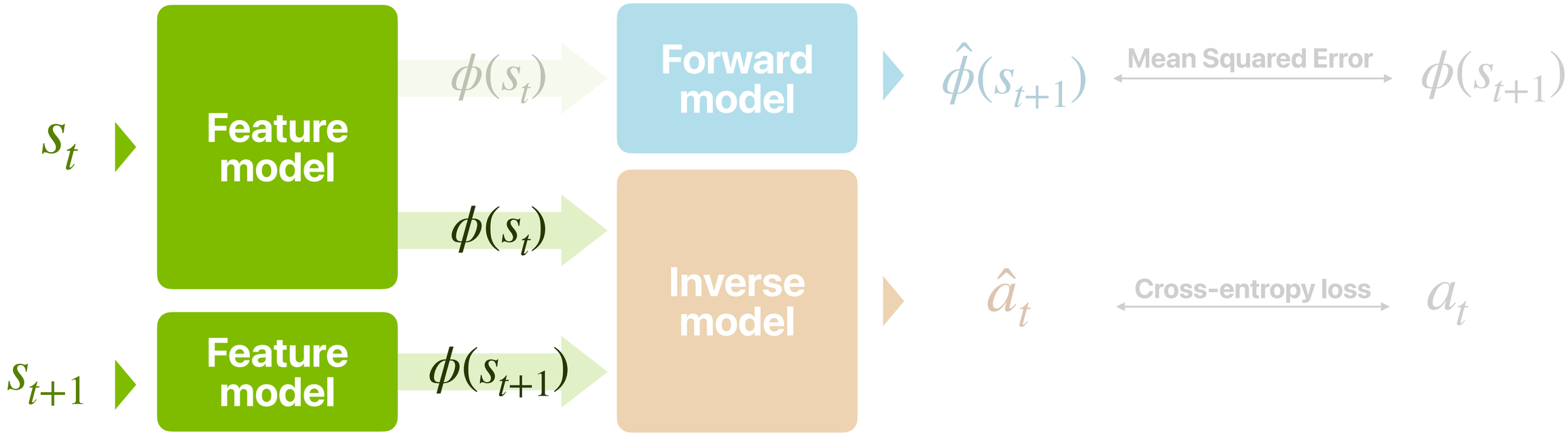
報酬　　内的報酬　　実報酬

**"好奇心" をベースとした内的報酬を追加する**

まだ見たことのない状態に "好奇心" を持ち、そこを優先的に探索

# Intrinsic Curiosity Module

"Curiosity-driven Exploration by Self-supervised Prediction" Pathak et al., ICML2017

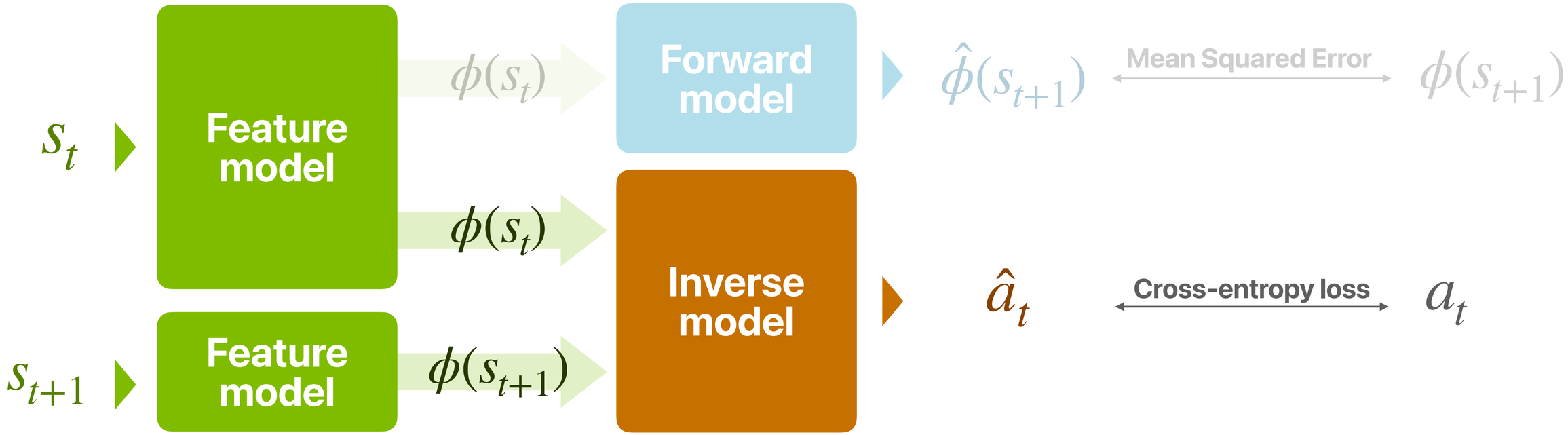



## Feature model

ある状態の潜在表現を得るモデル

# Intrinsic Curiosity Module

"Curiosity-driven Exploration by Self-supervised Prediction" Pathak et al., ICML2017



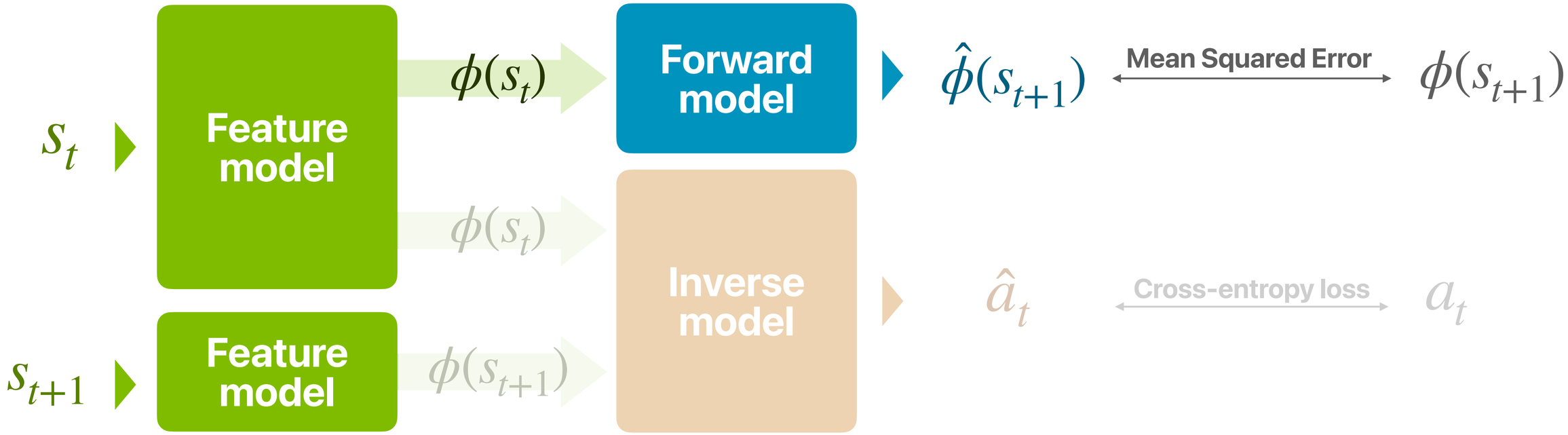

## Inverse model

$s_t$ から $s_{t+1}$ への遷移に必要な行動 $a_t$ を予測するモデル

$\hat{a}_t$ と $a_t$ の Cross Entropy Loss をとることで エージェントの行動に関係のある潜在表現 $\phi(s_t)$ を得られる

# Intrinsic Curiosity Module

"Curiosity-driven Exploration by Self-supervised Prediction" Pathak et al., ICML2017

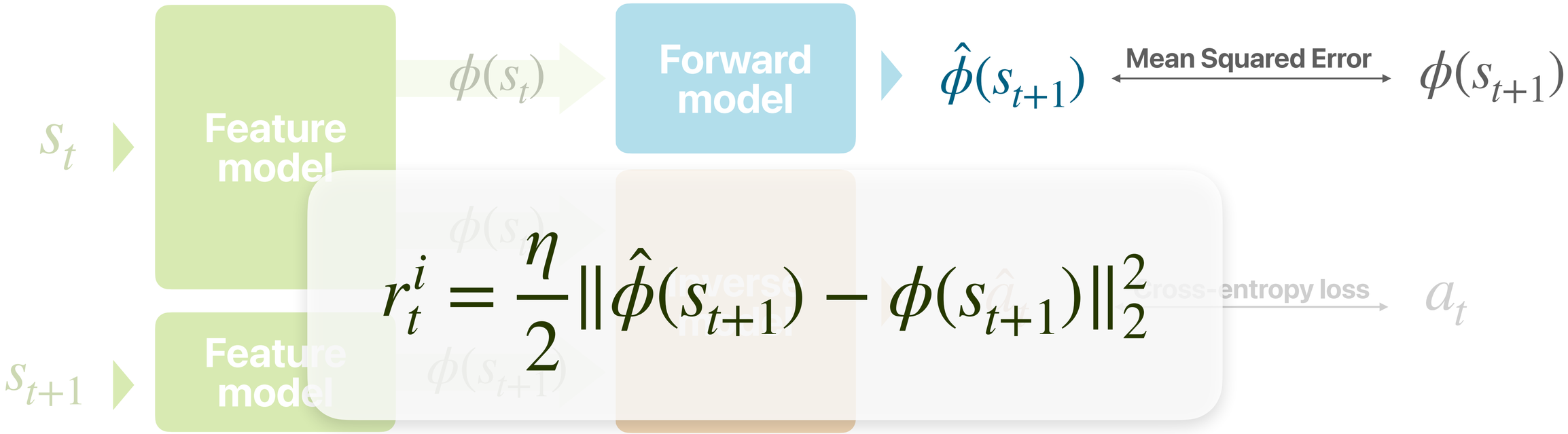



## Forward model

$s_t$ の潜在表現から $s_{t+1}$ の潜在表現を予測する

# Intrinsic Curiosity Module

"Curiosity-driven Exploration by Self-supervised Prediction" Pathak et al., ICML2017




$$r_t^i = \frac{\eta}{2}\|\hat{\phi}(s_{t+1}) - \phi(s_{t+1})\|_2^2$$

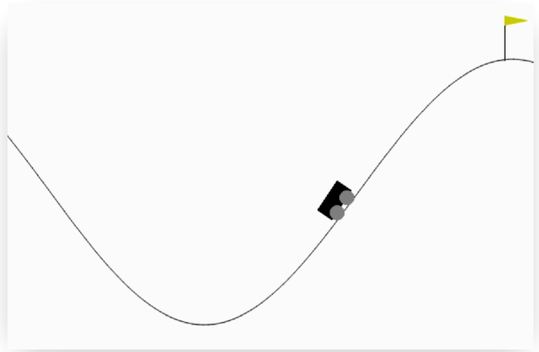
**報酬: $\phi(s_{t+1})$ と $\hat{\phi}(s_{t+1})$ の平均二乗誤差**

予測のズレがある → 未知の状態 → 興味, $\eta$ はハイパーパラメータ

# 実験

# 実験

ハイパーパラメータ



**全般**

環境: MountainCar-v0,
最大ステップ数 = 200 (default), エピソード数 = 10,000

**PPO**

$\epsilon = 0.2$, $\gamma = 0.99$, batch_size = 32, epochs = 10,
学習率 (actor) = 0.001, 学習率 (critic) = 0.003
エピソード終了時 buffer が 2048 steps を超えていたら更新

**ICM**

学習率 = 0.001, $\beta = 0.2$, $\lambda = 0.1$, batch_size = 32

**ネットワーク**

Actor, Critic, Feature (ICM), Forward (ICM), Inverse (ICM)
いずれも 3 層の全結合層からなる NN
(hidden_size = 64, activation = ReLU)

# 実験

比較対象

**PPO**

$$r_t = r_t^e$$

(実報酬のみ)

**PPO + ICM (η = 16)**

**PPO + ICM (η = 32)**

$$r_t = r_t^e + e_t^i$$

(ICM による内的報酬を加算)

**PPO と PPO + ICM で実験**

η は内的報酬の係数

# 結果

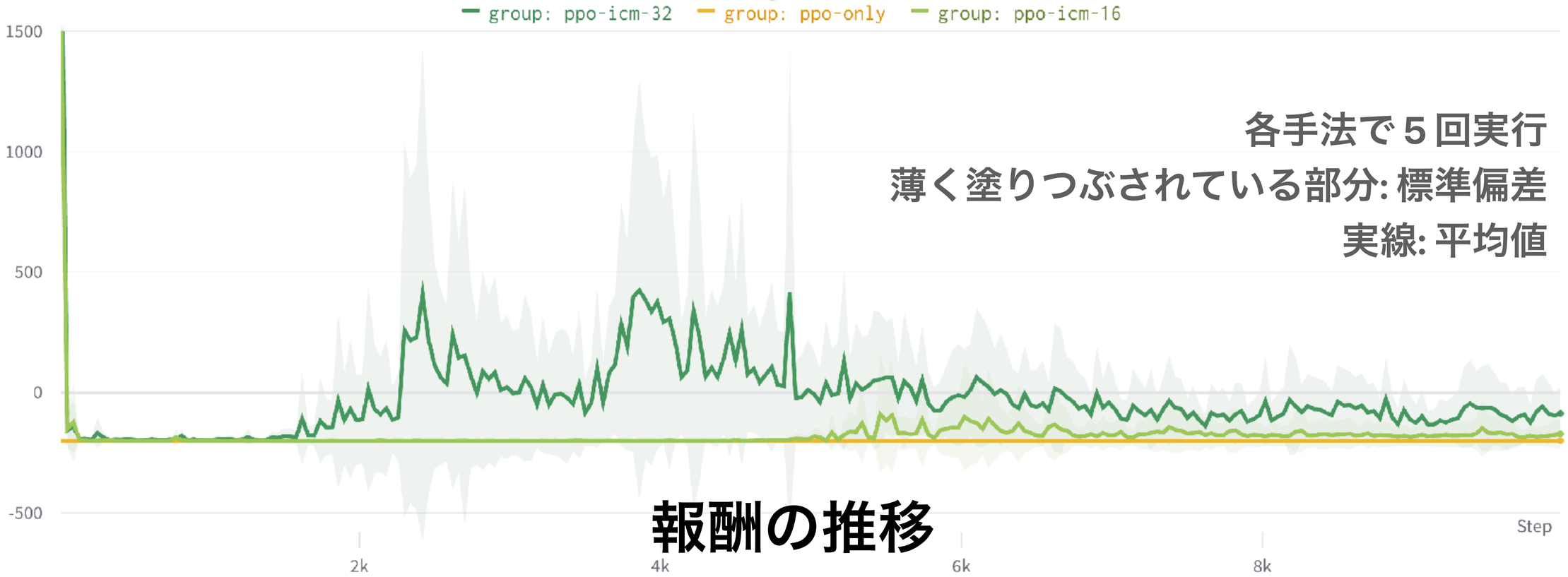
# 結果

報酬の推移

PPO

+ ICM (η = 16)

+ ICM (η = 32)




各手法で 5 回実行

薄く塗りつぶされている部分: 標準偏差

実線: 平均値

**報酬の推移**

報酬は増加していないので一見学習に失敗しているように見えるが
これは ICM のパラメータが毎回更新されることによるもの

PPO 単体は学習に失敗していることがわかる

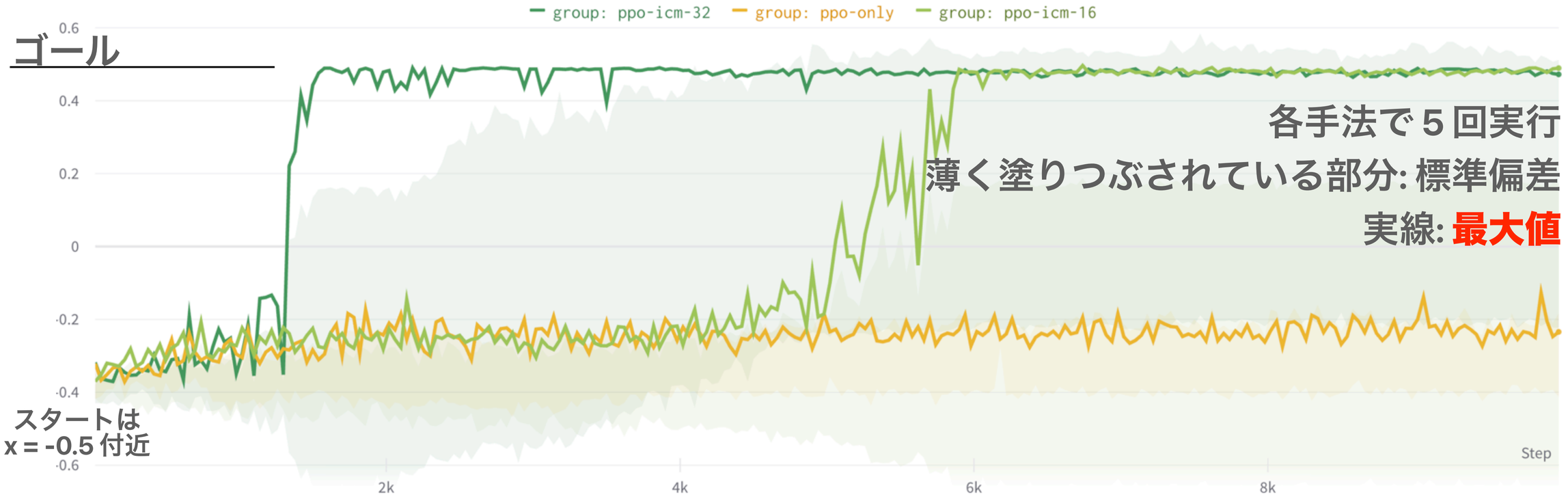
# 結果

x 座標の最大値

PPO

+ ICM (η = 16)

+ ICM (η = 32)




## x 座標の最大値

**PPO は学習に失敗、ICM はある段階で方針に "気付き" 始めてゴールを目指すようになる**

※ 実際はゴールに着くまで実報酬は変化しないはずだが、内的報酬によりこのようなことが起きている

**η は大きい方がより早くゴールに辿り着くようになる**

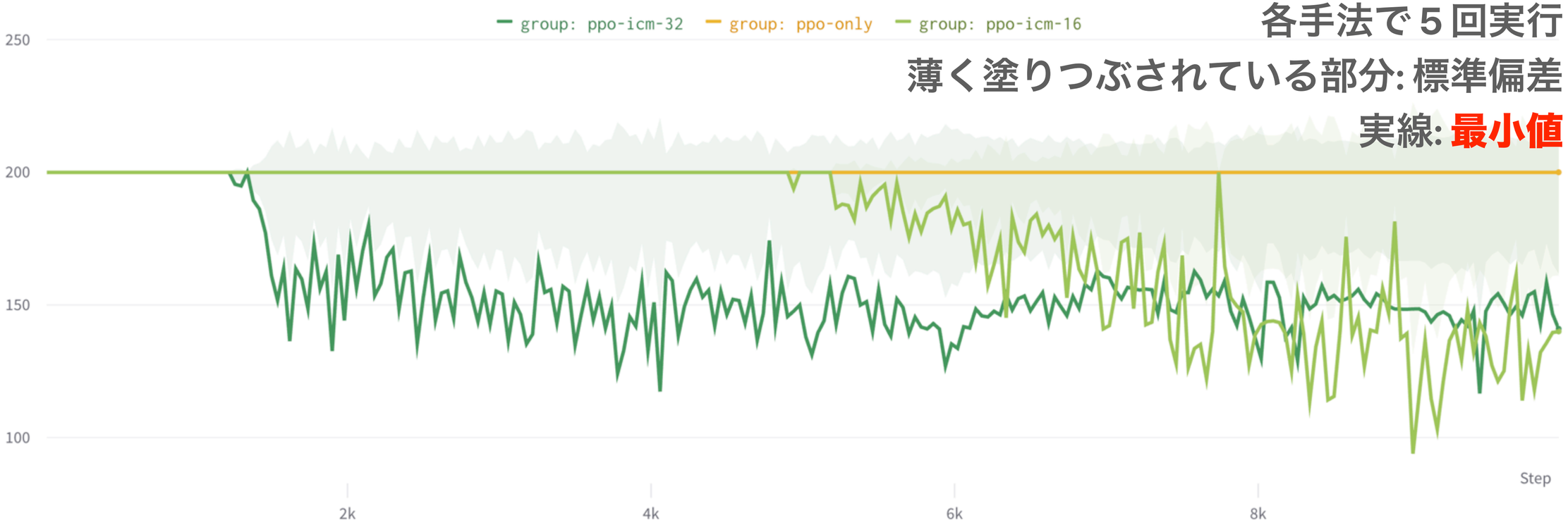
# 結果

ステップ数の推移

PPO

+ ICM (η = 16)

+ ICM (η = 32)




各手法で５回実行

薄く塗りつぶされている部分: 標準偏差

実線: **最小値**

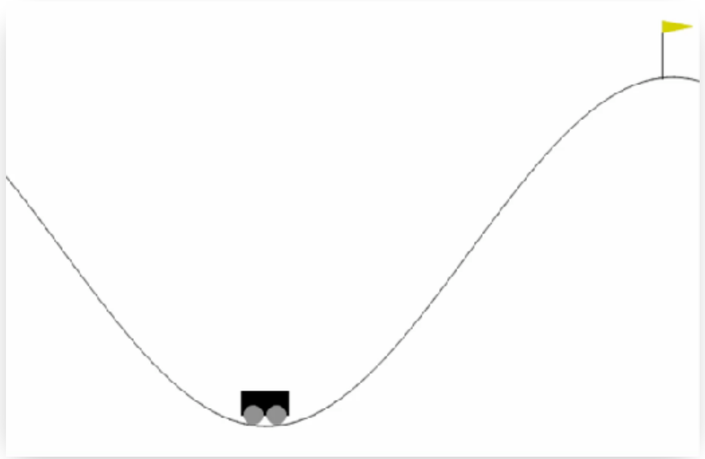
**１エピソードのステップ数** (ゴールにかかった時間)

**PPO はゴールできていない, ICM は η が大きいほど早く収束**

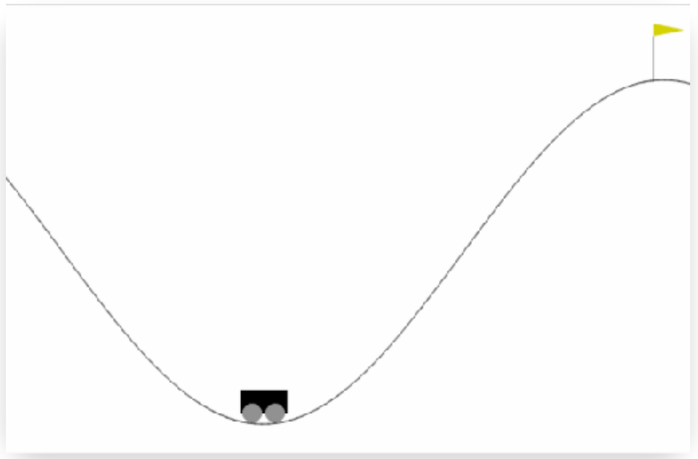
# 結果

実際の動き

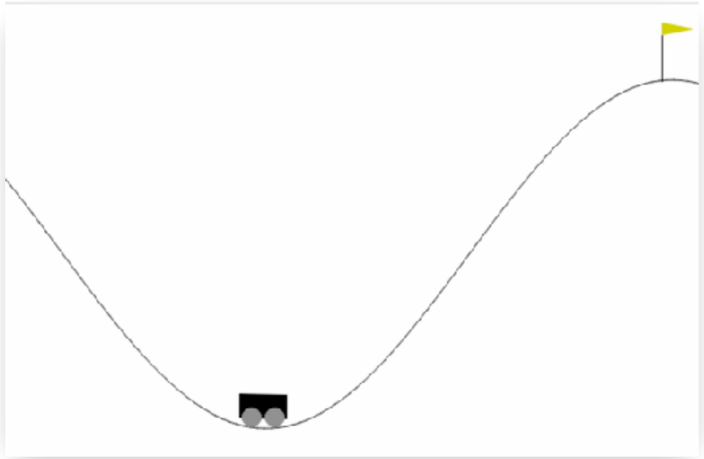
**t = 0s**



PPO



PPO + ICM (η = 16)



PPO + ICM (η = 32)

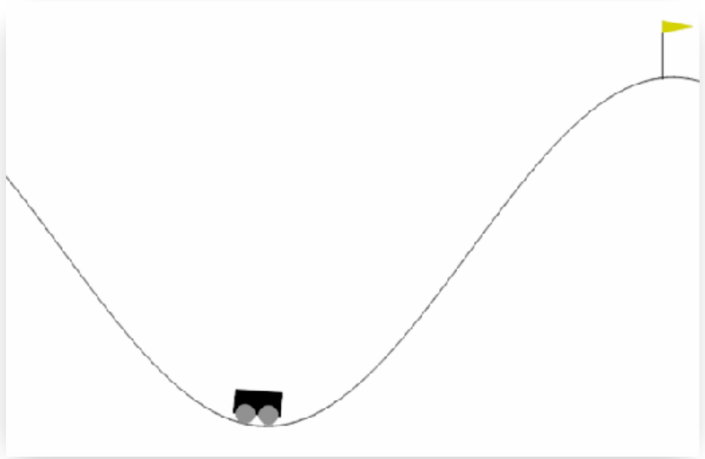
## 学習結果

行動は貪欲法により決定

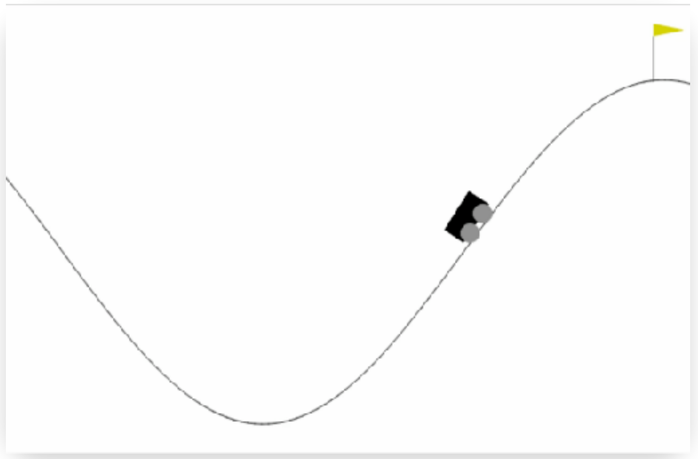
# 結果

実際の動き

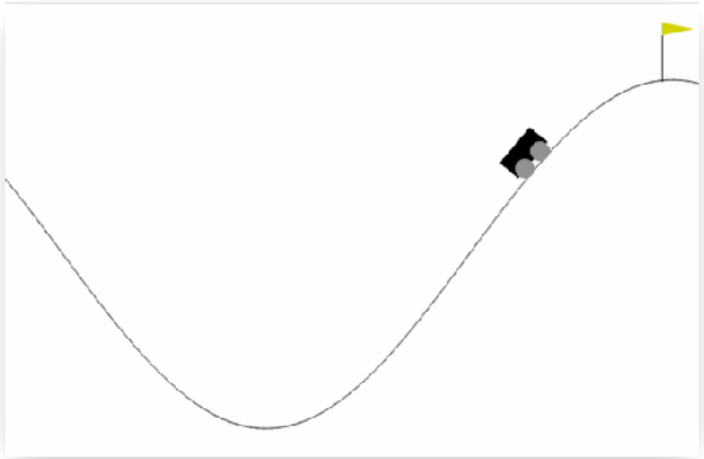
**t = 1s**



PPO



PPO + ICM (η = 16)



PPO + ICM (η = 32)

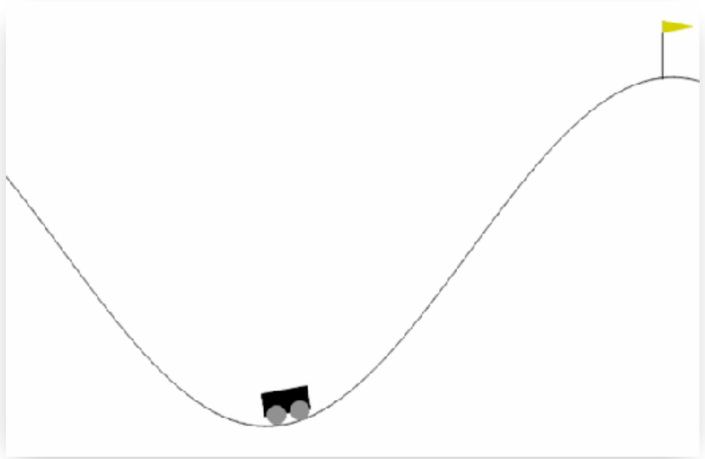
## 学習結果

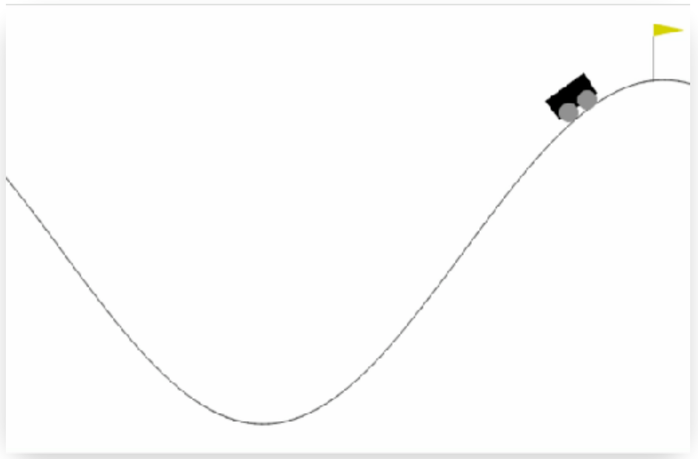
行動は貪欲法により決定

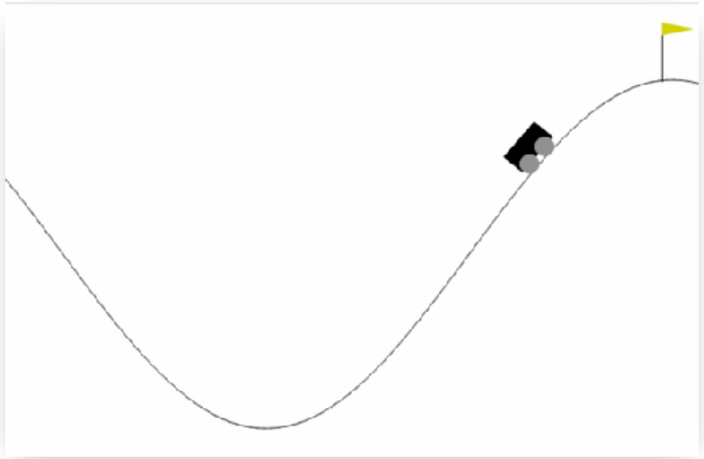
# 結果

実際の動き

**t = 2s**







**PPO**

**PPO + ICM (η = 16)**

**PPO + ICM (η = 32)**

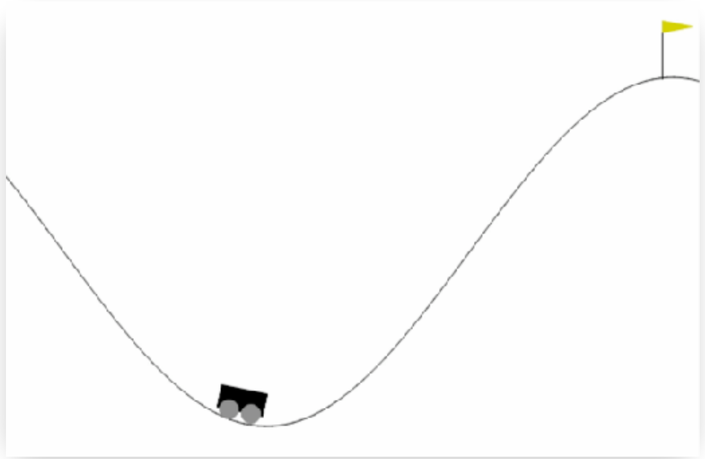
## 学習結果

行動は貪欲法により決定

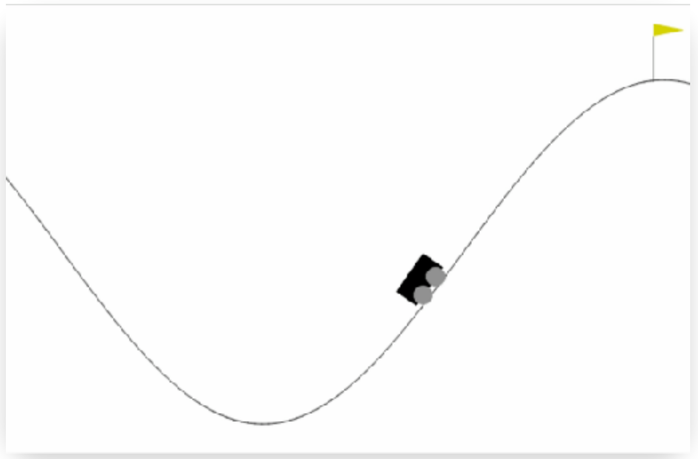
# 結果

実際の動き

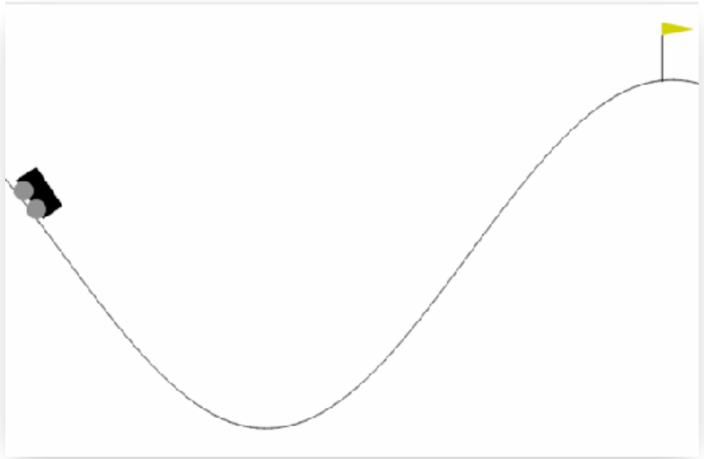
**t = 3s**



PPO



PPO + ICM (η = 16)



PPO + ICM (η = 32)

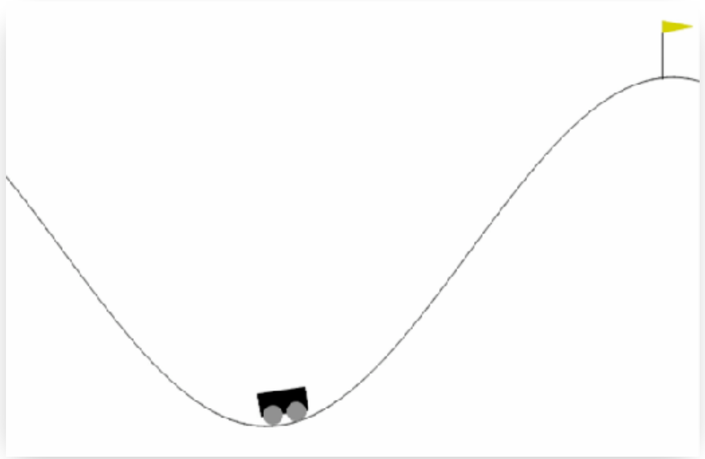
## 学習結果

行動は貪欲法により決定

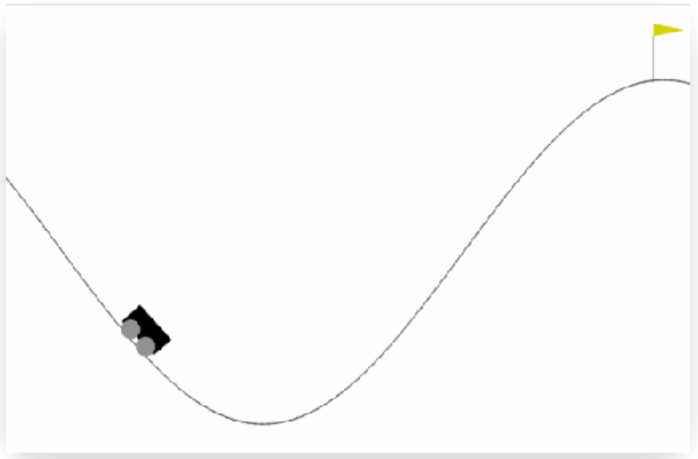
# 結果

実際の動き

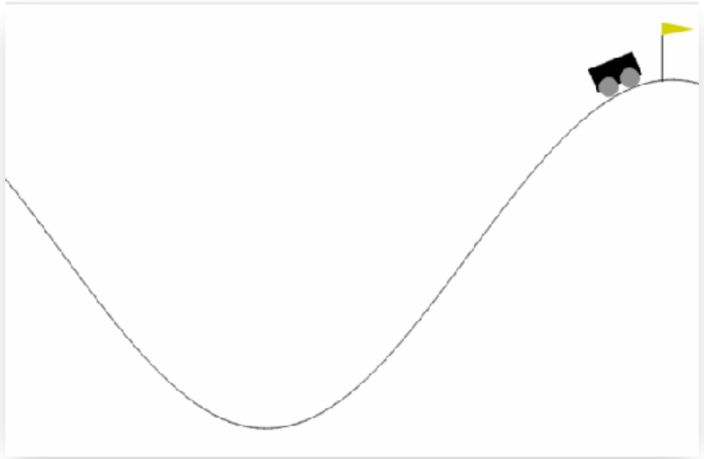
**t = 4s**



PPO



PPO + ICM (η = 16)



PPO + ICM (η = 32)

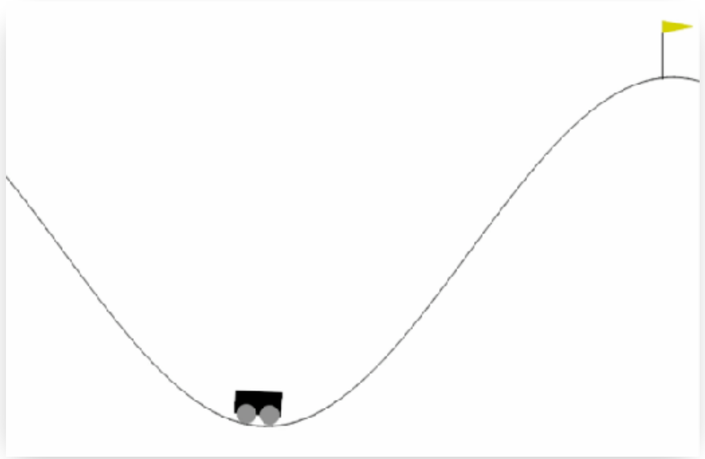
## 学習結果

行動は貪欲法により決定

# 結果

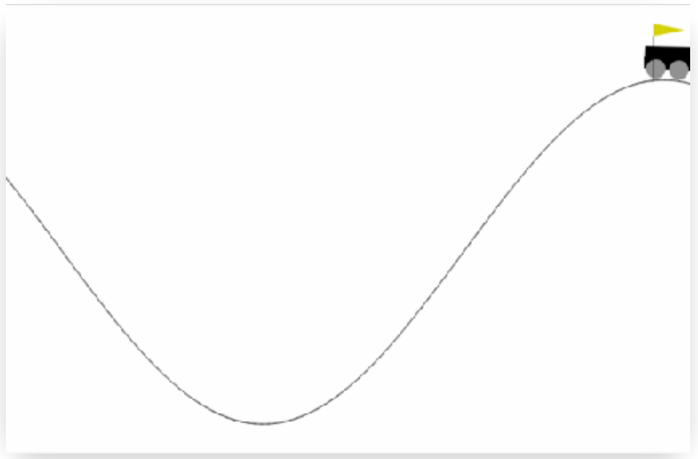
実際の動き

**t = 5s**



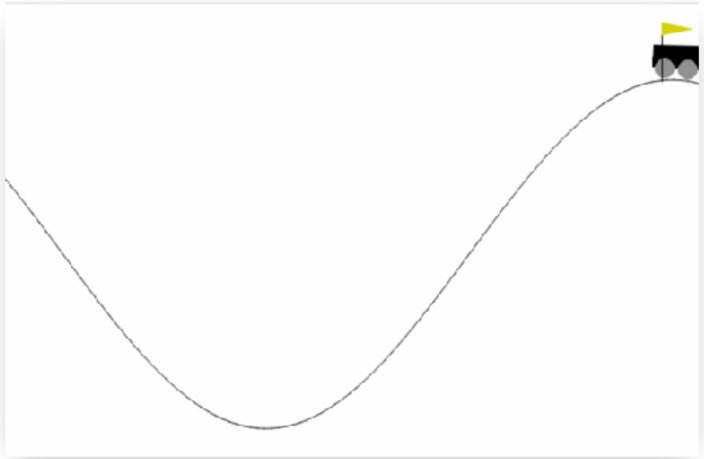
PPO

finished: t = 4.87s



PPO + ICM (η = 16)

finished: t = 4.15s



PPO + ICM (η = 32)

## 学習結果

行動は貪欲法により決定

# まとめ

- 好奇心で探索を促す **ICM (intrinsic curiosity module)** の再現実装を行った
- PPO + ICM で報酬がスパースな環境である
  **MountainCar-v0** を攻略するエージェントを学習
- ICM なしの PPO と比べて、**より探索が促され**
  **一部のエージェントはゴールまで到達**することができた