これからの強化学習

第2章　強化学習の発展的理論

Yuma Yamakura

目次

- 2.1 統計学習の観点から見たTD学習
- 2.2 強化学習アルゴリズムの理論性能解析と
  ベイズ統計による強化学習のモデル化
- 2.3 逆強化学習(Inverse Reinfocement Learning)
- 2.4 試行錯誤回数の低減を指向した手法
  : 経験強化型学習 XoL
- 2.5 群強化学習法
- 2.6 リスク考慮型強化学習
- 2.7 複利型強化学習

2.5 群強化学習法

2.5節概要

強化学習は環境とエージェントの相互作用
⇒試行錯誤回数が膨大

計算時間を減らしたい場合はどうするの？

2.5節概要

学習の高速化にはいくつかの観点
①計算回数/計算の単純化で計算時間を減らす
②試行錯誤(for文のイメージ)を減らす
③計算の並列化

今回は③の話

2.5節概要

並列化にもいくつかの観点

①計算機を並列(ハードウェア的/ソフトウェア的)

　ex)pyCUDAなどの並列コンピューティング

②学習機を並列(①の実現, 理論的な良い方法)

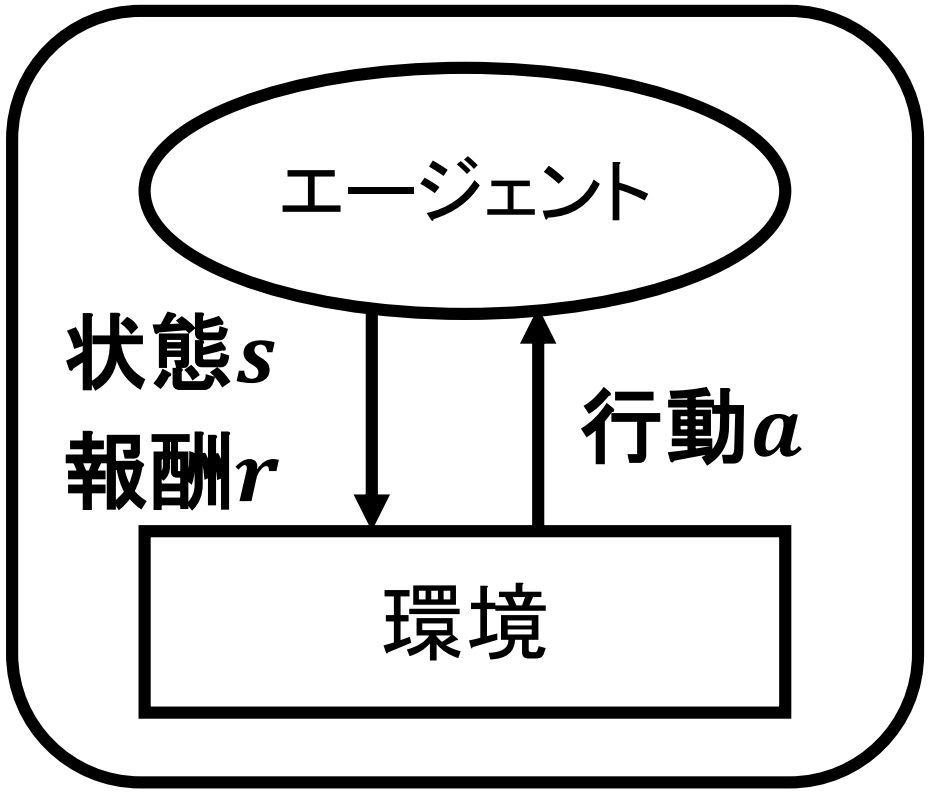
今回は②の話

(厳密に区切ることはできないけど・・・)

2.5.1 基本的な考え方とアルゴリズム

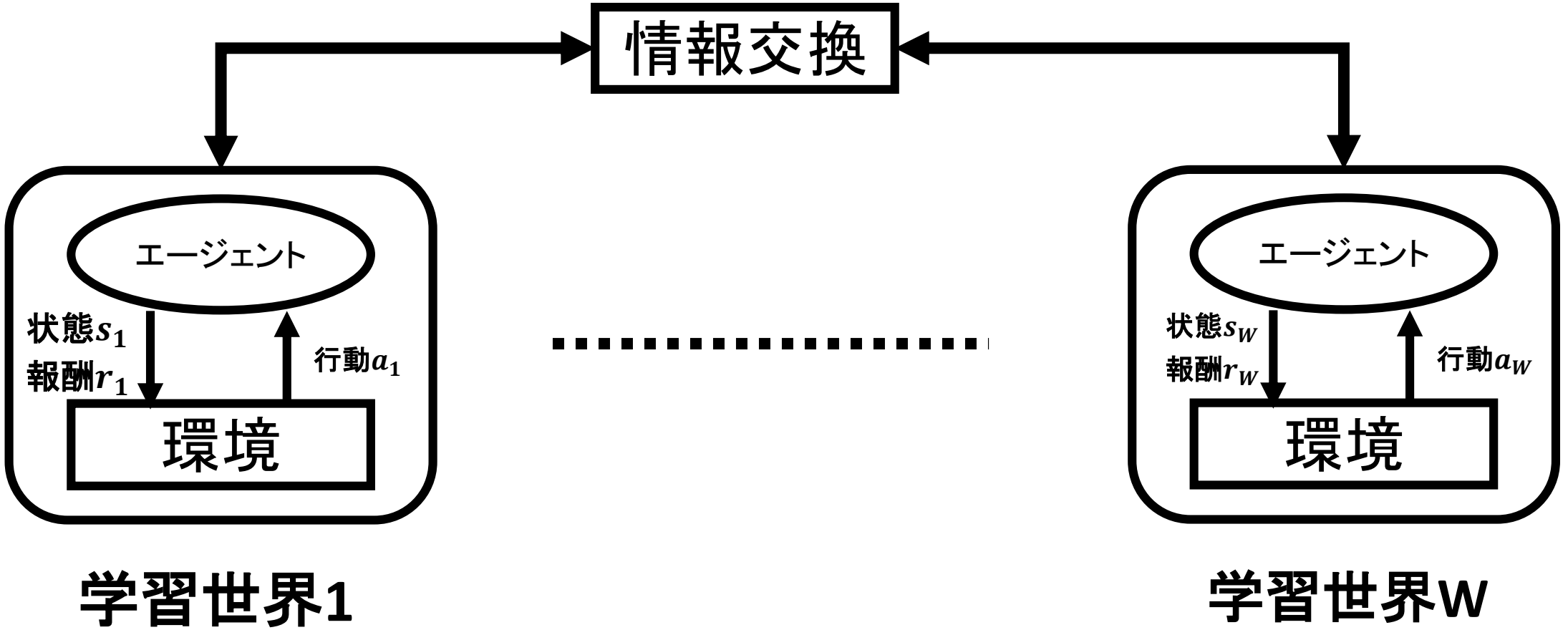
2.5.1 強化学習の基本的な枠組み

環境とエージェントの相互作用(学習世界)からなる



2.5.1 群強化学習の基本的な枠組み

環境とエージェントを**複数**用意して，
各学習世界の**情報交換**により効率化



2.5.1 学習の対象

学習の対象は
①各世界の環境とエージェント
②情報交換の仕方

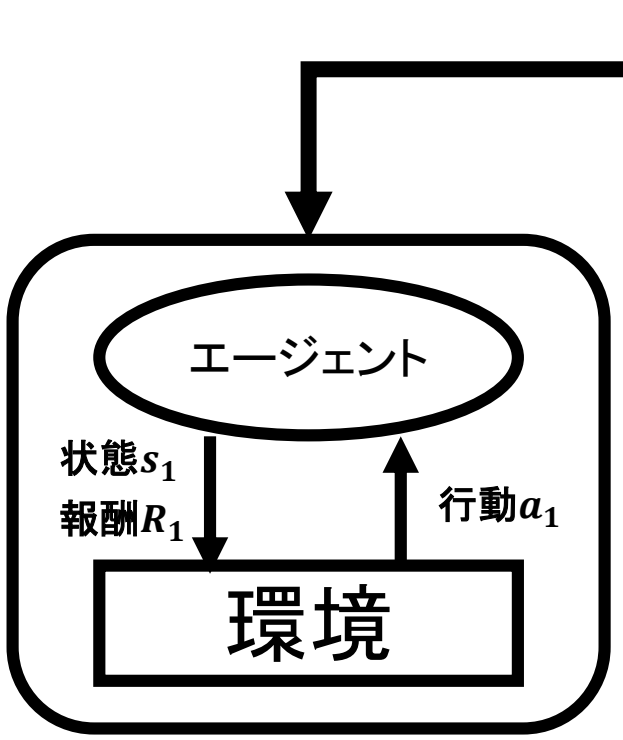
2.5.1 学習の対象

どういう行動をとればよいのかという, 方策の学習
⇒通常の強化学習と同じ

学習世界1

学習世界W

2.5.1 学習の対象



学習世界1

…

情報交換

各学習世界の方策を
どうほかの世界に伝えたら
いいのだろう･･･
⇒各学習世界の評価

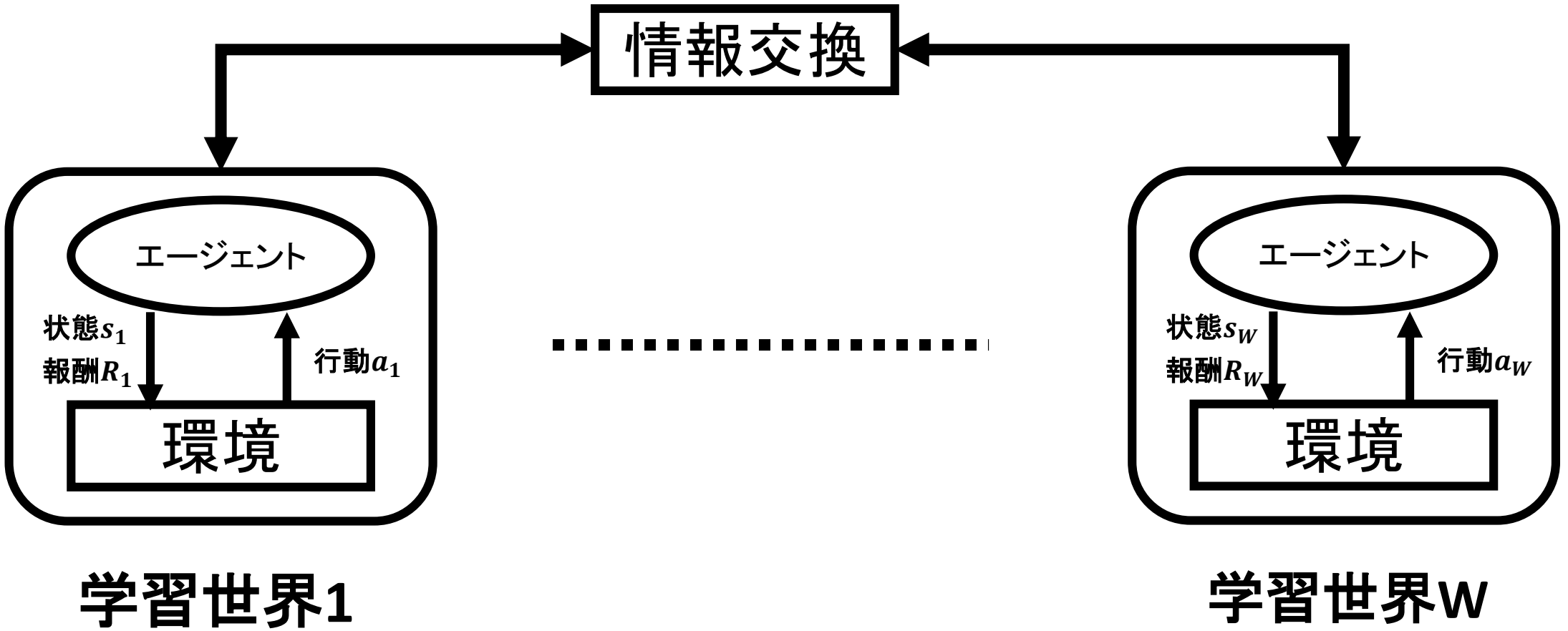
世界W

2.5.1 群強化学習の基本アルゴリズム

Step 0 : エージェントと環境の組である学習世界を複数用意
　　各学習世界の学習を初期化
Step 1 : 各学習世界が個別に, 通常の強化学習を行う
Step 2 : 各学習世界の学習を何らかの方法で評価
Step 3 : 各学習世界の評価に基づいて,
　　学習世界間で情報交換
Step 4 : 学習終了条件を満たす　　⇒ 終了
　　学習終了条件を満たさない　⇒ Step 1へ

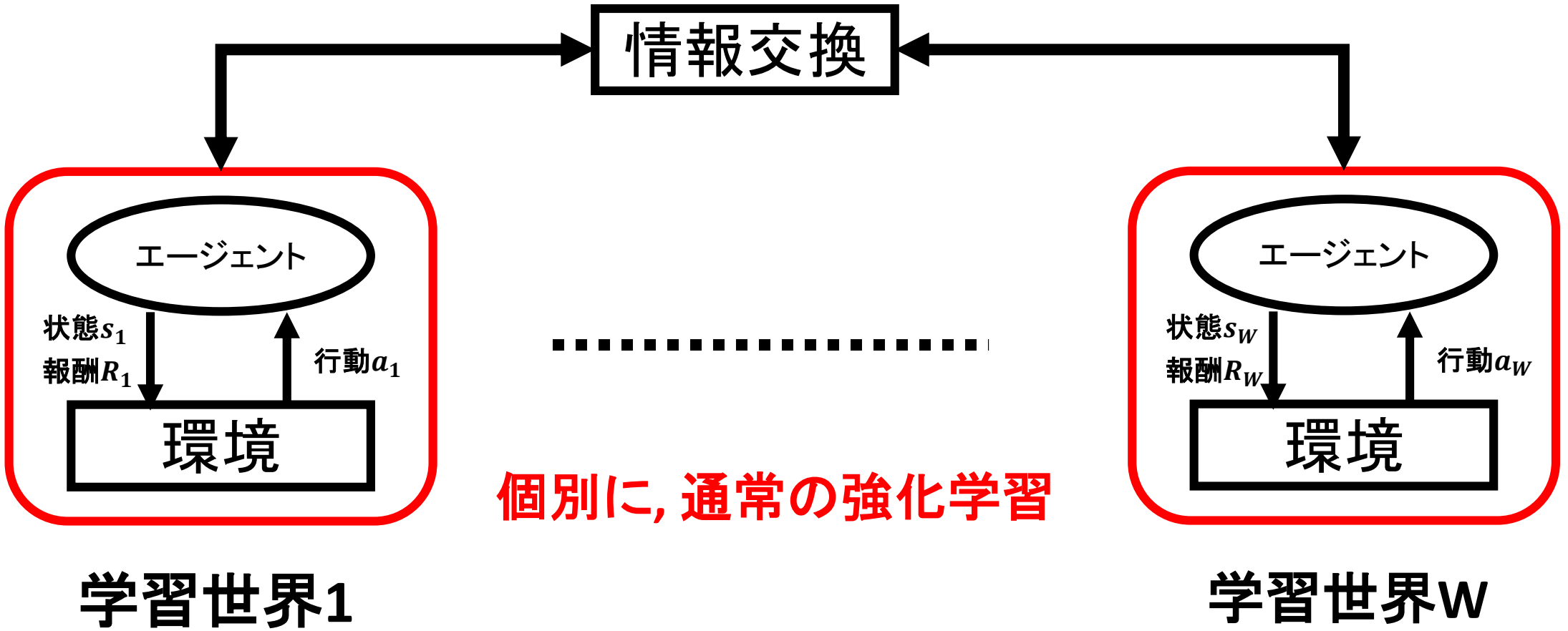
2.5.1 群強化学習の基本アルゴリズム

Step 0　枠組みの設計



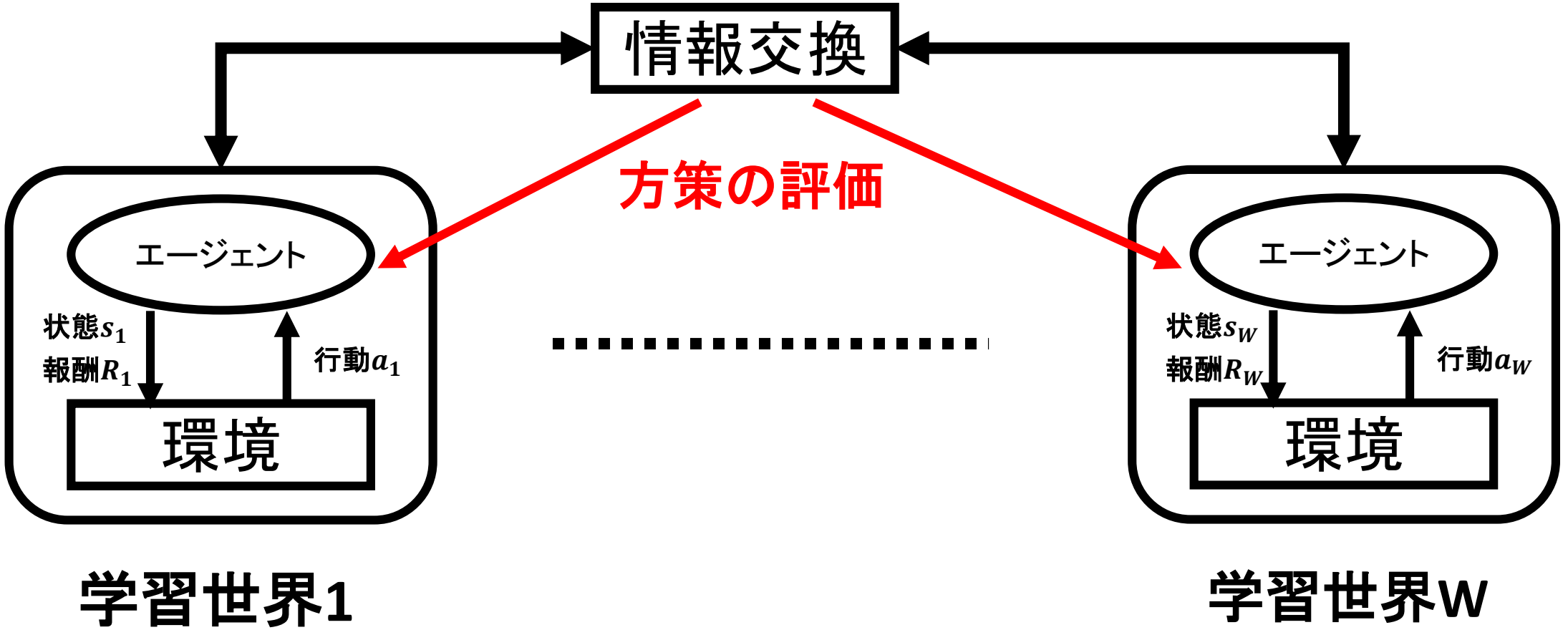
2.5.1 群強化学習の基本アルゴリズム

Step 1



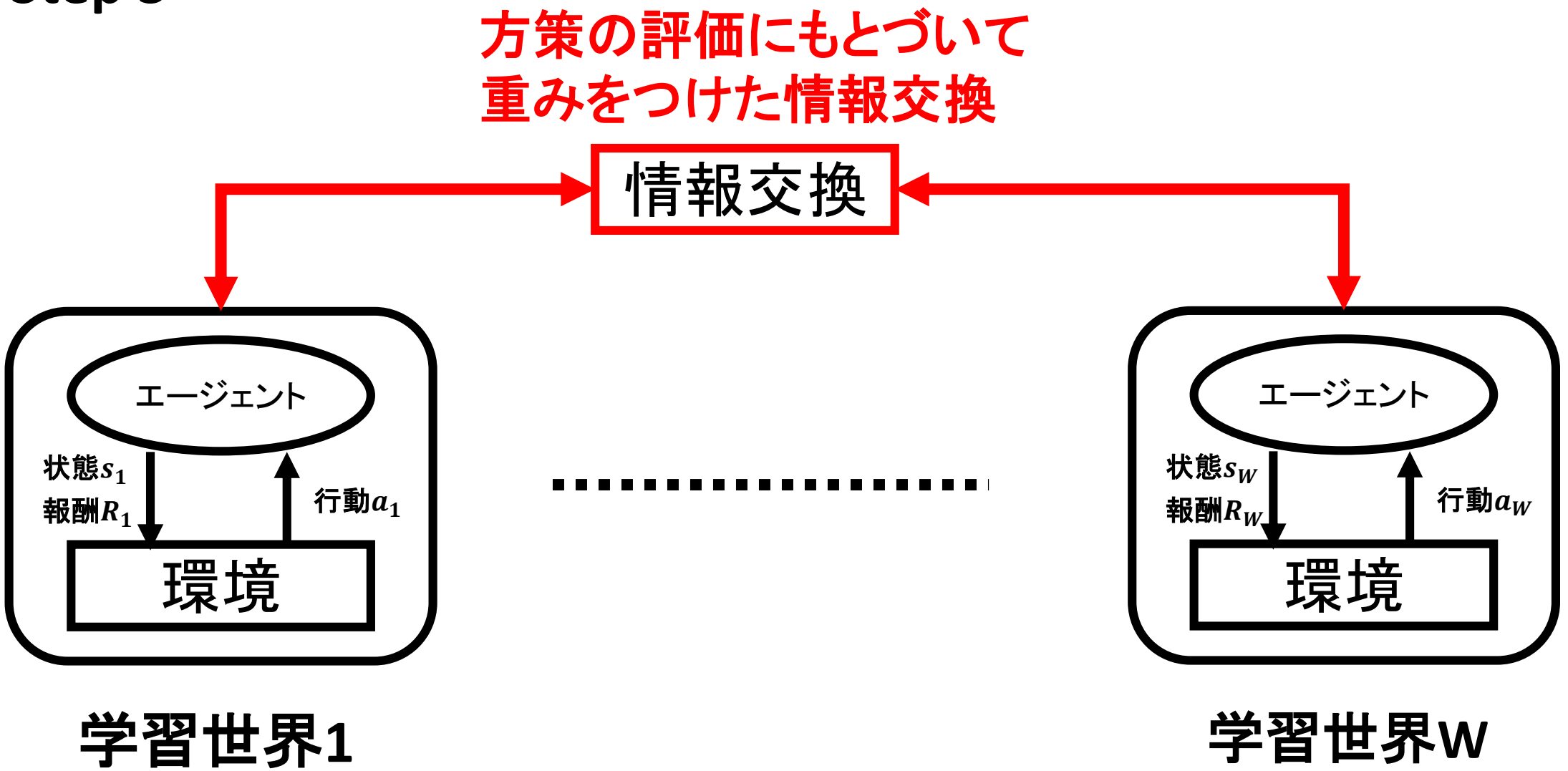
2.5.1 群強化学習の基本アルゴリズム

Step 2



2.5.1 群強化学習の基本アルゴリズム

Step 3



2.5.1 群強化学習の基本アルゴリズム

Step 4　最適方策を得た(これ以上更新しても意味がない)
⇒終了
まだ学習できる ⇒ Step1へ

2.5.2 各学習世界の学習法

2.5.2 Step 1で行う通常の強化学習

各学習世界での学習は, 通常の強化学習を行う
(本に書いてあるのはQ-learningなので省略)

2.5.3 各学習世界の評価法

2.5.3 各学習世界の評価は何を用いるか？

評価値として使えそうなもの･･･

・ロボット歩行　⇒ 歩行時間の長さ

・迷路　⇒ ゴールまでの時間

・バンディット　⇒ 収益

などなど･･･

2.5.3 各学習世界の評価は何を用いるか？

評価値として使えそうなもの・・・

・ロボット歩行　⇒ 歩行時間の長さ

・迷路　⇒ ゴールまでの時間

・バンディット　⇒ 収益

などなど・・・

報酬で表現可能

⇒統一的に扱うには，**報酬の期待値(利得)を最大**にする
　という強化学習の考えがピッタリでは？

2.5.3 利得を用いるとどうなるか？

情報交換の学習機は,$Q_i(s,a)$を受け取る
⇒それぞれの$Q_i(s,a)$を最大化する方策を
　実際に実行して, 報酬の期待値を見る必要！
⇒多大な時間
　(というか期待値ってどんだけやるつもり・・・？)

なので, 十分近似できるやつを探す必要

2.5.3 ざっくりとした考え方

要するに,

①Q値を用いて評価

②学習が進むと良い報酬が多くなるはずなので,

　受け取ったばかりの報酬を使いたい

という考え方をすると,

$$E = \sum_{t=1}^{L} d^{L-t} r_t$$

が最大となるQ値を用いる

2.5.3 ざっくりとした考え方

$E = \sum_{t=1}^{L} d^{L-t} r_t$の補足

L : Q(s,a)を得たエピソードの行動回数
d : 割引率(0 < d < 1)

tが小さい(=学習初期) ⇒ 割引されて影響が小さい
tが大きい(=学習が進んでいる)
⇒ 割引されずに影響が大きい

2.5.4 学習世界観の情報交換法

2.5.4 学習交換法を考える準備

W : 学習世界の数

$Q_i : i = 1, 2, \ldots, W$として, i番目の学習世界のQ値

$E_i : Q_i$の評価値

$Q^{best}(s, a)$: E_iが最大となるiに対する$Q_i(s, a)$値

2.5.4 A.最良値で更新する方法

単純に, 各学習世界のQ値のなかで, 最も評価の高いものを代入すればいいのでは？という考え

$\forall i \in \{1, 2, \ldots, W\}, \forall s, \ \forall a$ について,

$$Q_i(s, a) \leftarrow Q^{best}(s, a)$$

と更新

2.5.4 B.最良値との平均をとる方法

各学習世界のQ値も学習した内容なのだから,

Q_iとQ^{best}で平均をとるという考え

※A.と比べて, 探索の余地を残している

$\forall i \in \{1, 2, \ldots, W\}, \forall s, \ \forall a$ について,

$$Q_i(s, a) \leftarrow \frac{Q^{best}(s, a) + Q_i(s, a)}{2}$$

と更新

2.5.4 C.PSOに基づく方法

最適化の分野における, 多数の探索点を用いて
並列に解を探索する解法を適用
⇒PSO(Particle Swarm Optimization)

2.5.4 C.PSOに基づく方法

自己最良Q値 $P_i(s,a) = \underset{Q_i(s,a)}{\arg\max} E_i$

全体最良Q値 $G(s,a) = max\ P_i(s,a)$

W_{pso}, C_1, C_2：適当な重みパラメータ

R_1, R_2：0から1までの一様乱数

2.5.4 C.PSOに基づく方法

$\forall i \in \{1, 2, \ldots, W\}, \forall s, \ \forall a$ について,

$$V_i(s,a) \leftarrow W_{pso} V_i(s,a) + C_1 R_1 \left(P_i(s,a) - Q_i(s,a)\right) + C_2 R_2 \left(G(s,a) - Q_i(s,a)\right)$$

となるV_iに対して

$$Q_i(s,a) = Q_i(s,a) + V_i(s,a)$$

と更新

2.5.4 Đ.アントコロニー最適化に基づく方法

アリの採餌行動をヒントに考案された最適化手法

フェロモン Q値 : Q_P

<特徴>　①各学習世界の学習により変化

②時間が経つと蒸発する(蒸発率 ρ : $\rho \leq 1$)

2.5.4 Đ. アントコロニー最適化に基づく方法

$\forall i \in \{1, 2, \ldots, W\}, \forall s, \ \forall a$ について,

$$Q_P(s,a) \leftarrow (1-\rho)Q_P(s,a) + \sum_{i=1}^{W} \frac{E_i}{\sum_{r=1}^{W} E_r} Q_i(s,a)$$

$$Q_i(s,a) = Q_P(s,a)$$

と更新

2.5.5 連続状態行動空間学習問題への展開

2.5.5 結論

ロボット系は事前知識がないと面倒なので結論だけ

PSOを使うと早く学習できる！

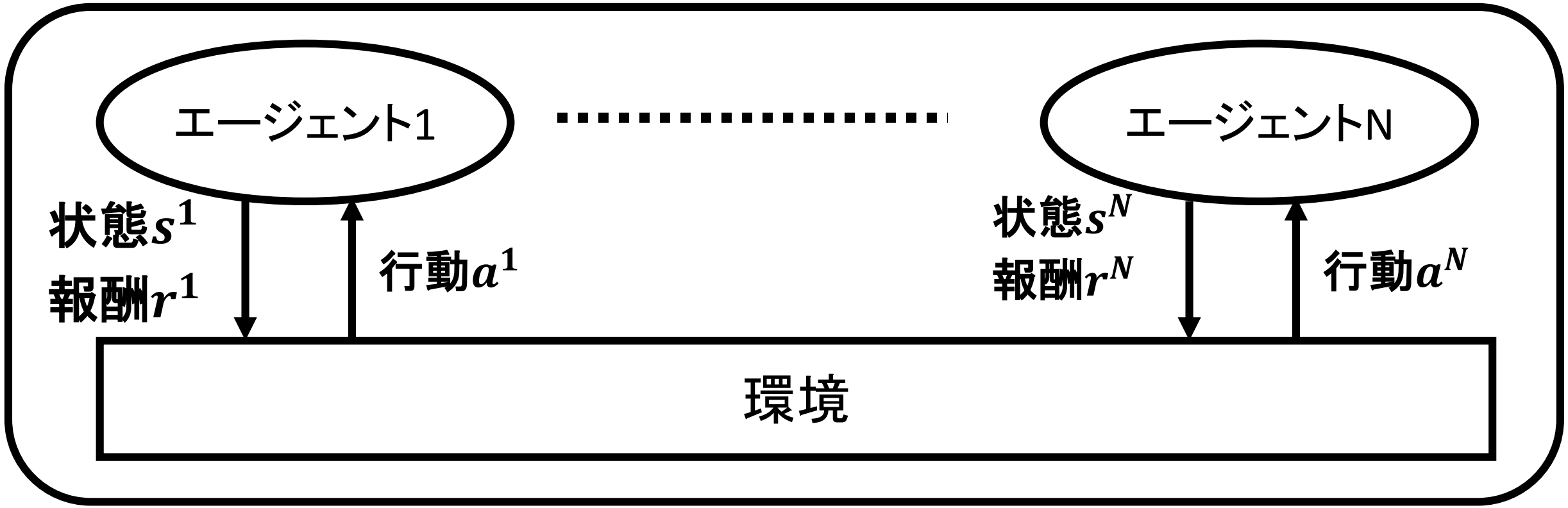
2.5.6 マルチエージェント学習問題への展開

2.5.6 マルチエージェント強化学習

今まではシングルエージェント強化学習を考えていたが,
エージェントの数を増やすとどうなるのか？
⇒1つの環境に複数のエージェントがいる強化学習を
　マルチエージェント強化学習という

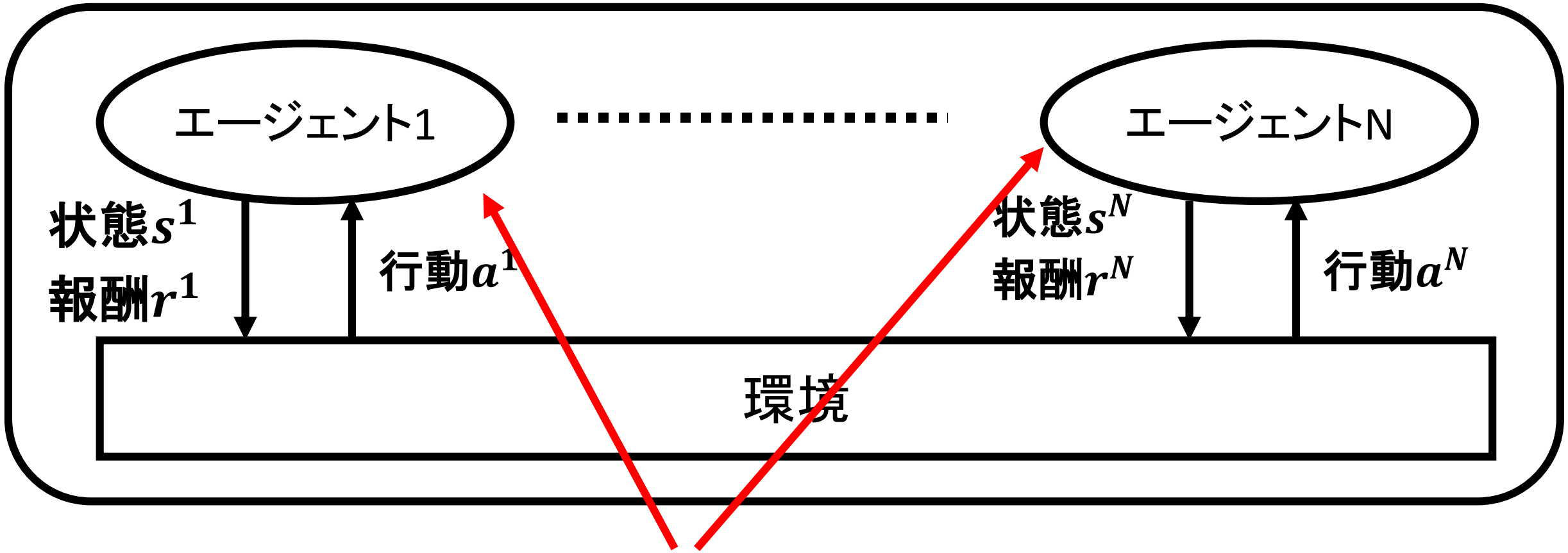
2.5.6 マルチエージェント強化学習

N個のエージェントの場合の概念図



2.5.6 マルチエージェント強化学習

N個のエージェントの場合の概念図



エージェント同士に相互作用(衝突, 協力など...)

2.5節　注意点

・群強化学習

⇒複数の世界で強化学習(並列計算)

⇒情報を共有することで計算時間の削減

・分散強化学習(マルチエージェント強化学習)

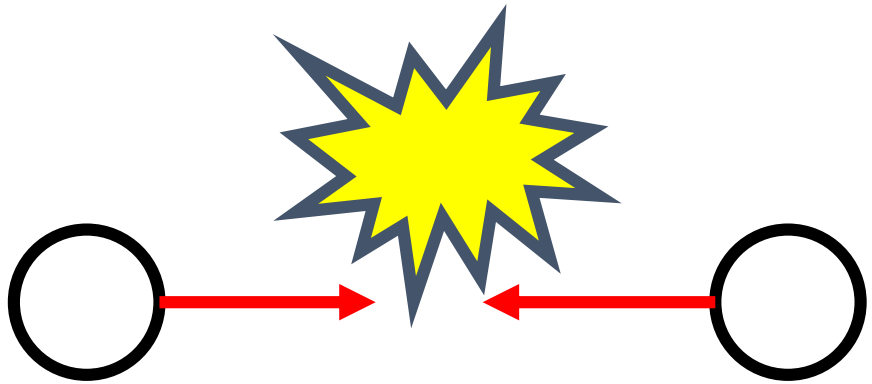
⇒複数のエージェントが存在するMDP上での強化学習

2つを合わせたものが，マルチエージェント群強化学習

2.5.6 マルチエージェントの困難

・エージェント同士の関係がある

ex)衝突



※これはゲーム理論的な考え方をすることもできる

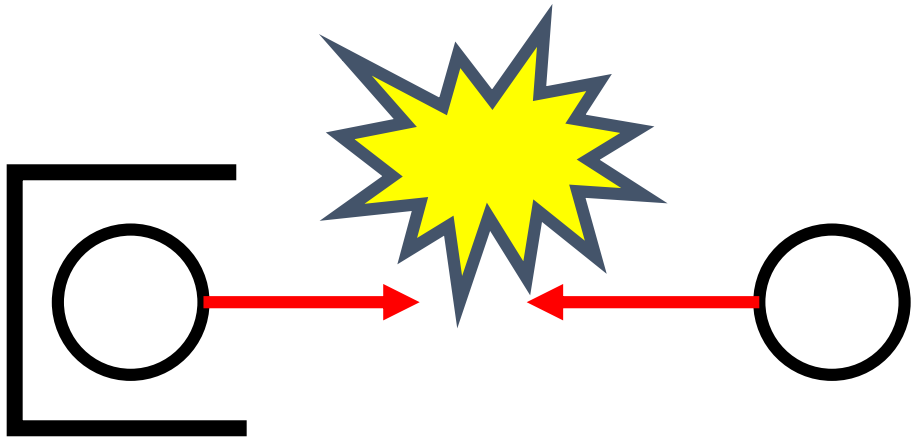
(完全協調問ゲーム 完全対立ゲーム, 混合ゲーム...)

※ゲーム理論は経済学で有名

2.5.6 マルチエージェントの困難

・報酬の与え方が難しい

ex)



右側のエージェントのほうが悪い？？？

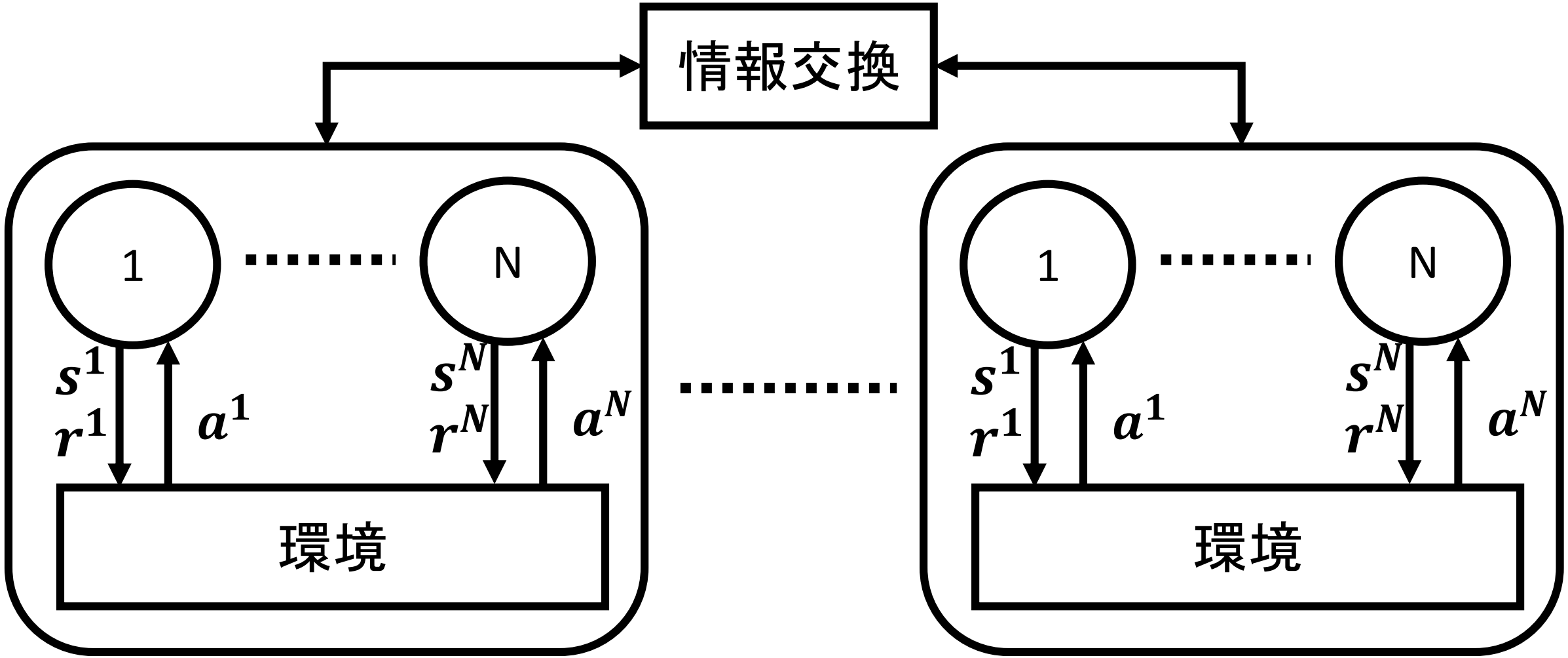
それとも, このような状況になるのが悪い？？？

エージェント同士の関係のせいで報酬が複雑に

(しかも, 学習できるように適切に決めることは難しい)

2.5.6 マルチエージェント群強化学習

N個のエージェントに対する概念図



2.5.6 マルチエージェント強化学習の例

あとの細かいところはざっくりと！

例として,
・囚人のジレンマ(ゲーム理論)
・フォーメーション制御(群ロボット)
が挙げられている.いずれもエージェント同士の協力を
するような方策を求めるマルチエージェント強化学習.
詳細は文献読みましょう.ここに書いてることじゃ話にならない.

2.5.7 おわりに

2.5.7 まとめ

群強化学習 ⇒ 計算時間を減らす枠組みのひとつ

各学習世界との情報交換の方法により,
大きい計算時間の短縮を見込める

マルチエージェントは状態空間が爆発的に広くなるので
組み合わせることが重要
(組み合わせないと計算時間6時間とか平気でかかってた)