仅熵增不可逆: 纯粹基于熵增进行强化学习的 AI 会不会是强人工智能？

· TJU 强化学习实验室 王漫超

·

Introduction

熵增过程是一个自发的由有序向无序发展的过程. 一般笔者认为这个过程本身只是基于我们宇宙基础三定理之二的不相容原理和不确定性原理. 本文参照这两个原理，提出了纯粹基于熵增的多 agent 强化学习模型，并对模型进行简单的探讨.

background

熵

热力学中，熵表征物质状态的参量之一，用符号 S 表示，其物理意义是体系混乱程度的度量。在经典热力学中，Clausius 在热力学中第二定律中用增量定义新的状态函数[1]:

$$dS = \left(\frac{dQ}{T}\right)_r$$

式中 T 为物质的热力学温度; dQ 为熵增过程中加入物质的热量，下标 r 表示过程

可逆. 若过程不可逆的, 则等号变为大于号, 综合可得:

$$dS \geq \frac{\partial Q}{T}$$

在统计热力学中, 熵的大小与体系的微观状态Ω有关:

$$S = k \cdot \ln\Omega;$$

其中 k 为 Boltzmann 常量, $k = 1.3807 \times 10^{-23}$ J·K^{-1}. 微观状态 Ω 是大量质点的体系经统计规律而得到的热力学概率: 这比较符合直觉, 比如对于两个房间和三个球我们只能说两个房间的球数目差一个的概率大一些而不能判定两个房间的球一定会在未来的某一时刻差一个数目, 虽然后者貌似不会不发生.

不相容原理

量子力学中, Pauli 指出原子域不能有两个或两个以上的电子具有完全相同的四个量子数[2]. 笔者加强这一观点认为组成我们的基本粒子在我们所在的 n 维域的任一对称 d 维亚域的一组基上的投影多于(d-1)个特征相同的两个对象存在的概率为 0.

不确定性原理

量子力学中, Heisenberg 指出, 你不可能同时知道一个粒子的位置和它的速度, 粒子位置的不确定性不小于一个常数[3]

$$\Delta x \Delta p \geqslant h/4\pi$$

其中 h 为 Planck 常量, $h = 6.62606896(33) \times 10^{-34}$ J·s. 该原理表明即使存在因果律, 人们也无法验证, 因为人们必然无法观测任意时刻的因或果状态, 这也是为什么物理学定律最终不得不用概率分布来描述的原因.

(人们不能观测因果并不是因为人们的观察影响了因果，不确定性原理是亚域的内秉性质，在量子力学中实际表现出量子系统的基础性质，而非描述观测能力)

强化学习过程

强化学习

强化学习是从动物学习、参数扰动自适应控制等理论发展而来，其基本原理是：如果 Agent 的某个行为策略导致环境正的奖赏(强化信号)，那么 Agent 以后产生这个行为策略的趋势便会加强。Agent 的目标是在每个离散状态发现最优策略以使期望的折扣奖赏和最大。

强人工智能

针对计算机和其它信息处理机器，Searle 指出具有强人工智能的计算机不仅是用来研究人的思维的一种工具；相反，只要运行适当的程序，计算机本身就是有思维的[4].

纯熵增强化学习(EBRL)

纯熵增强化学习(EBRL, Entropy-increasing based reinforcement learning)是笔者在接触强化学习领域研究如何定义 Reward 时提出的一种方法，这种方法主要实现了 Reward 的独立定义(无需人参与).

对于一个 Multi-Agent 的 RL，我们定义:

Def 1. 总体 Reward

$$Wall = \ln \Omega;$$

*其中 **Ω** 为全体 Agent 在环境中呈现的微观状态，需要根据具体环境定义，这与热力学熵的定义类似*

Def 2. Agent 间贡献 Reward

$$W[i,j] = \sum[k] (|| f[k](i) - f[k](j) ||2entro);$$

*其中 **f[k](i)** 为 agent 在亚域 Domain 中第 k 维的投影折合坐标，也需要根据环境定义，这个 Reward 表示一对 Agent 的不相容度.*

EBRL 是 RL 的一种，其特殊在于其必须为 Agent 尽可能多(满足统计学)的 multi-agent 强化学习，且 Agent 的 Reward 定义为:

考虑熵最大，希望 Wall 最大的 Agent 数目和希望 Wall 最小的 Agent 数目差我们按正态分布取初始值，即数目相等的可能性最大.

对于希望 Wall 最大的 Agent i，其 Reward 为

$$R[i] = Wall + p * \sum[j](W[i,j]);$$

对于希望 Wall 最小的 Agent i，其 Reward 为

$$R[i] = -Wall + p * \sum[j](W[i,j]);$$

Experiment

因为本文目的是投到 IBBB 上，且由于时间及地理上的一些影响(貌似是熵增的抑制)，这个实验就留待日后完成，我觉得我已经有一些关于环境的初步设想 kira~

其中有一个 3 类多 Agent 的情况.

实验的感觉会成这样，比如一个格子棋盘，有的 Agent 会渐渐让下的棋子逼近一条线，有的 Agent 却是在围地，有的 Agent 喜欢在其他 Agent 的棋子两侧跳来跳去，然后他们就分别发明了五子棋，围棋和跳棋，有的 Agent 单纯就是远离其他的 Agent，有点像围棋的序盘，**然而笔者只是口胡!**

在 Agent 很多的情况下，感觉就像这样吧:

Reference

[1] Clausius, Rudolf. On the Motive Power of Heat, and on the Laws which can be

deduced from it for the Theory of Heat. Poggendorff's Annalen der Physick, 1850

[2] Pauli, Pauli Exclusion Principle, 1925

http://hyperphysics.phy-astr.gsu.edu/hbase/pauli.html

[3] W. Heisenberg, Über quantentheoretishe Umdeutung kinematisher und mechanischer Beziehungen, Zeitschrift für Physik, 33, 879-893, 1925 (received July 29, 1925). [English translation in: B. L. van der Waerden, editor, Sources of Quantum Mechanics (Dover Publications, 1968)

[4] J. Searle, Minds Brains and Programs. The Behavioral and Brain Sciences, vol. 3, 1980