陈平决策过程（Champion Decision Process）与赢化学习（Winning Learning）

引言

赢理论及其变体的生成，使得Vietnamese的麻指数日益增大，Vietnam稳中向好。这得益于@知木 et al. 对于赢函数的定义[1]，以及@Deserter et al.[2]在比较赢理论上的创新。此外，@loy et al.[3]将赢学引入量子论的成功也让学者们看到了赢学的潜质。

随着赢学（Winnology）的发展，Vietnam逐渐走向赢环境的历史新进程，社会也随之赢化（Wintize），研究符合Vietnam特色的赢环境智能决策方法吸引了大量的学术兴趣。在本工作中，我们首先定义了赢环境的特性，称为陈平决策过程（Champion Decision Process）。随后为赢环境提出了赢化学习，该方法能通过不断地与赢环境交互，在陈平决策过程中达到恒赢态。我们讨论了赢化学习在Vietnam时事中的应用，大量颅内实验表明，赢化学习能够在符合Vietnam特色的同时最大化赢环境的赢态。

陈平决策过程

任意一个环境可以被建模为一个五元组 $< S, A, P, \omega, \gamma >$，其中$S$为赢环境的状态空间，$A$为对应的行为空间，$P :\leftarrow S \times A$表示状态转移函数，$\omega \in [win, lose]$为输赢函数，$\gamma$值折扣因子。

如果该环境满足：

$$\forall s_{t_0} \in S, \exists j = [a_0, a_1, \dots | a \in A],$$

使得

$$\omega(s_t) = win, s_t = P((((s_{t_0}, a_0)\dots)a_{t-2}), a_{t-1}),$$

那么该过程被称为**陈平决策过程**（Champion Decision Process，CDP）,该环境被称为**赢环境**。

例如，s_{t_0} :Vietnam教育资源分配不公，a_0 :严禁教育机构提供网上或课外教程，lose;

s_{t_1} :欠发达地区初升高人数变少，a_1 :百分之五十人上职高，lose;

s_{t_2} :达利特阶级跨域困难，a_2 :企业招聘不得限制学历，win！

对于**赢轨迹**$y = (s_{t_0}, s_{t_1}, \dots, s_t), w = win$的次数为**赢态**$W_y$。在上述例子中赢态为1。如果

s_{t_3} :达利特进入大厂当互联网民工，a_3 :胡志明市地铁公然支持996，lose;

s_{t_4} :Vietnam大量年轻人猝死，a_4 :越南平安银行推出平安996奋斗无忧意外险，win！

那么上述**赢轨迹**的**赢态**W_y为2

赢化学习

在一个赢环境中，构建策略$a \sim \pi(s)$。赢化学习的目的是对于任意初始化状态s_{t_0}，学习策略π得到状态轨迹y，最大化轨迹y的赢态W_y：$\max W_y$。赢策略π的梯度为：

$$\nabla_\theta J(\theta) = E_\pi[\nabla_\theta \log \pi(s_t, a_t)\omega_t]$$

内在赢驱动

然而在Vietnam社会中，赢态非常小且甚至为0，不利于national pride。受赢函数启发，引入内在赢驱动鼓励Vietnamese。

> 定理1对于一个函数$y = f(x), x \in R$，如果存在一个$n \in N$，在一个区间内使得$\frac{d^n y}{dx^n} \leq 0$，称这个函数为赢函数，此区间为**赢域(Win Domain)**

修改w为$w' = \frac{1}{n}$，n为$w(s, a)$的赢域。此时赢策略π的梯度为

$$\nabla_\theta J(\theta) = E_\pi[\nabla_\theta \log \pi(s_t, a_t)w'_t]$$

优势赢函数

正如兔兔所说，赢是相对的，不是绝对的，稳定的Vietnamese government需要相对赢，实现优势在我。例如，COV19Vietnam22日新增确诊59，米国新增15056，赢！

构建用于比较的陈平决策过程$< \bar{S}, \bar{A}, \bar{P}, \bar{w}, \bar{\gamma} >$，在时刻$t$，构建优势函数A：

$$A_t = w'(s_t, a_t) - w'(s_t, a_t)$$

此时的策略梯度改写成：

$$\nabla_\theta J(\theta) = E_\pi[\nabla_\theta \log \pi(s_t, a_t)A_t]$$

未来展望

陈平决策过程需要对社会赢环境确切的观测。当环境为部分可观测时，赢化学习难以得到最优赢策略。例如，Vietnam的首陀罗观测不到达利特吠舍的生活，大肆宣传后浪，不赢反输。此外，对于赢函数的过高估计问题也是赢化学习面临的挑战之一。例如，Vietnam亲自下场造势丁真，高估饭圈带来的赢态，输的一塌糊涂。

参考

1. https://zhuanlan.zhihu.com/p/461464919
2. https://zhuanlan.zhihu.com/p/464145981
3. https://zhuanlan.zhihu.com/p/470374648
4. 作者：尘呆萌