# 多智能体强化学习最小均值方差公式推导

2024 年 11 月 25 日

王钰

## 1 VDN+CTD 公式推导

注：这里只考虑VDN和CTD中的Vanilla情况。

### 1.1 MDP

一个MDP可用一个五元组$(S, A, P^a_{s,s'}, R, \gamma)$表示。其中$S$（state）是状态空间，$A$（action）是动作空间， $P^a_{s,s'}$是在状态$s$下采取动作$a$可以转换到状态$s'$的概率， $R_t$是智能体在$t$时可以获得的奖励，在环境中，智能体每次实际获得的奖励记为$r_t$， $\gamma$是折扣因子。

定义策略为：

$$\pi : P(a|s)$$

策略为从状态$s$选择动作$a$的概率。

定义回报（随机变量）为：

$$G^\pi(s_t) = \sum_{k=1}^{\infty} \gamma^{k-1} R_{t+k}$$

，

回报为遵从策略$\pi$，智能体从状态$s_t$开始到结束，可以获得的累计折扣奖励。

在MDP中，算法的目标是得到最优的策略$\pi^*$，使得回报$G^\pi(s_t)$最大。

## 1.2 Value Functions

定义状态价值函数$V^\pi(s_t) = \mathbb{E}[G^\pi(s_t)]$，其中$G^\pi(s_t)$是指遵从策略$\pi$，智能体从状态$s_t$开始到结束，可以获得的累计折扣奖励。对状态价值函数可进行进一步推导，得到状态价值函数的贝尔曼方程：

$$
\begin{aligned}
V^\pi(s_t) &= \mathbb{E}[G^\pi(s_t)] \\
&= \mathbb{E}[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots] \\
&= \mathbb{E}[R_{t+1} + \gamma(R_{t+2} + \gamma R_{t+3} + \cdots)] \\
&= \mathbb{E}[R_{t+1} + \gamma G^\pi(s_{t+1})] \\
&= \mathbb{E}[R_{t+1}] + \gamma \mathbb{E}[G^\pi(s_{t+1})]
\end{aligned}
$$

定义状态-动作价值函数$Q^\pi(s_t, a_t) = \mathbb{E}[G^\pi(s_t, a_t)]$，其中$G^\pi(s_t, a_t)$是指智能体从状态$s_t$开始，选取动作$a_t$，然后遵从策略$\pi$直到结束，可以获得的累计折扣奖励。同理，对状态-动作价值函数可进行进一步推导，得到状态-动作价值函数的贝尔曼方程：

$$
\begin{aligned}
Q^\pi(s_t, a_t) &= \mathbb{E}[G^\pi(s_t, a_t)] \\
&= \mathbb{E}[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots] \\
&= \mathbb{E}[R_{t+1} + \gamma(R_{t+2} + \gamma R_{t+3} + \cdots)] \\
&= \mathbb{E}[R_{t+1} + \gamma G^\pi(s_{t+1}, a_{t+1})] \\
&= \mathbb{E}[R_{t+1}] + \gamma \mathbb{E}[G^\pi(s_{t+1}, a_{t+1})]
\end{aligned}
$$

由于价值函数是回报的期望，所以在最优的策略$\pi^*$时，有最优状态价值函数$V^{\pi^*}(s_t)$ 和最优动作-状态价值函数$Q^{\pi^*}(s_t, a_t)$。因此，找到使得回报$G^\pi(s_t)$最大的最优策略$\pi^*$的过程就相当于对价值函数进行优化的过程。

## 1.3 TD learning

强化学习中，一种更新价值函数的方法是TD学习。

根据状态价值函数的贝尔曼方程，可得：

$$V^\pi(s_t) = r_{t+1} + \gamma V^\pi(s_{t+1})$$

其中$r_{t+1}+\gamma V^\pi(s_{t+1})$被称为TD目标，是即时奖励和下一个状态的价值的无偏估计，用于近似贝尔曼方程中的期望值。

根据TD目标，可定义TD误差$\delta_{vt}$为$r_{t+1}+\gamma V^\pi(s_{t+1})-V^\pi(s_t)$。TD误差项表示实际获得的回报与预期回报之间的差异。

利用TD误差，可以推导出更新当前的状态价值函数的公式：

$$\begin{aligned}
V^\pi(s_t) &= V^\pi(s_t) + \alpha \times \delta_{vt} \\
&= V^\pi(s_t) + \alpha(r_{t+1} + \gamma V^\pi(s_{t+1}) - V^\pi(s_t))
\end{aligned}$$

其中$\alpha$是学习率，控制状态价值函数的更新速度。

同理，根据状态-动作价值函数的贝尔曼方程，可得：

$$Q^\pi(s_t, a_t) = r_{t+1} + \gamma Q^\pi(s_{t+1}, a_{t+1})$$

同理，可以推导出更新当前状态-动作价值函数的公式：

$$Q^\pi(s_t, a_t) = Q^\pi(s_t, a_t) + \alpha(r_{t+1} + \gamma Q^\pi(s_{t+1}, a_{t+1}) - Q^\pi(s_t, a_t))$$

其中$\alpha$是学习率，控制状态-动作价值函数的更新速度；可定义TD误差$\delta_{qt}$为$r_{t+1}+\gamma Q^\pi(s_{t+1},a_{t+1})-Q^\pi(s_t,a_t)$。

## 1.4 CTD

### 1.4.1 用于计算方差的价值函数推导

根据定义$V^\pi(s_t)=\mathbb{E}[G^\pi(s_t)]$，前面的推导可以获得计算在状态$s_t$下的回报的均值。接下来，要推导出可以计算在状态$s_t$下的回报的方差$\bar{V}^\pi(s_t)$。

首先，对回报的方差$\mathbb{V}[G^\pi(s_t)]$进行推导：

$$
\begin{aligned}
\mathbb{V}[G^\pi(s_t)] &= \mathbb{E}[G^\pi(s_t) - \mathbb{E}[G^\pi(s_t)]]^2 \\
&= \mathbb{E}[r_{t+1} + G^\pi(s_{t+1}) - \mathbb{E}[G^\pi(s_t)]]^2 \\
&= \mathbb{E}[r_{t+1} + G^\pi(s_{t+1}) - V^\pi(s_t)]^2 \\
&= \mathbb{E}[r_{t+1} + G^\pi(s_{t+1}) - V^\pi(s_t) + V^\pi(s_{t+1}) - V^\pi(s_{t+1})]^2 \\
&= \mathbb{E}[(r_{t+1} + V^\pi(s_{t+1}) - V^\pi(s_t)) + (G^\pi(s_{t+1}) - V^\pi(s_{t+1}))]^2 \\
&= \mathbb{E}[r_{t+1} + V^\pi(s_{t+1}) - V^\pi(s_t)]^2 + \mathbb{E}[G^\pi(s_{t+1}) - V^\pi(s_{t+1})]^2 \\
&\quad + 2 \times \mathbb{E}[r_{t+1} + V^\pi(s_{t+1}) - V^\pi(s_t)] \times \mathbb{E}[G^\pi(s_{t+1}) - V^\pi(s_{t+1})] \\
&= \mathbb{E}[r_{t+1} + V^\pi(s_{t+1}) - V^\pi(s_t)]^2 + \mathbb{V}[G^\pi(s_{t+1})] \\
&\quad + 2 \times \mathbb{E}[r_{t+1} + V^\pi(s_{t+1}) - V^\pi(s_t)] \times 0 \\
&= \mathbb{E}[\delta_{vt}]^2 + \mathbb{V}[G^\pi(s_{t+1})]
\end{aligned}
$$

其中，$\mathbb{V}$表示计算某个变量的方差，根据先前的定义，$\mathbb{E}[G^\pi(s_t)] = V^\pi(s_t)$...（论文里这里每行推导都可以细讲）。

回顾TD学习，用于计算回报的均值的状态价值函数的贝尔曼方程为：$V^\pi(s_t) = r_{t+1} + \gamma V^\pi(s_{t+1})$；用于计算回报的均值的状态价值函数的更新公式为：$V^\pi(s_t) = V^\pi(s_t) + \alpha(r_{t+1} + \gamma V^\pi(s_{t+1}) - V^\pi(s_t))$。因此，根据以上关于$\mathbb{V}[G^\pi(s_t)]$的推导，可得出用于计算回报的方差的状态价值函数的更新公式为：

$$\bar{V}^\pi(s_t) = \bar{V}^\pi(s_t) + \bar{\alpha}(\delta_{vt}^2 + \bar{V}^\pi(s_{t+1}) - \bar{V}^\pi(s_t))$$

在这里，我们采用了TD学习中的常用技术，用智能体与环境交互中实际计算的$\delta_{vt}^2$来代替$\mathbb{E}[\delta_{vt}]^2$。类似的，可以得到用于计算回报的方差的状态-动作价值函数的更新公式为：

$$\bar{Q}^\pi(s_t, a_t) = \bar{Q}^\pi(s_t, a_t) + \bar{\alpha}(\delta_{qt}^2 + \bar{Q}^\pi(s_{t+1}, a_{t+1}) - \bar{Q}^\pi(s_t, a_t))$$

## 1.5 DQN

## 1.6 VDN