$$
\begin{aligned}
Q_*(s_t, \text{上}) &= 930, \\
Q_*(s_t, \text{下}) &= -60, \\
Q_*(s_t, \text{左}) &= 120, \\
Q_*(s_t, \text{右}) &= 321.
\end{aligned} \tag{2}
$$

请问，哪个动作会成为a_t?

A，上

B，下

C，4种动作都有可能

5，DQN的输出层用于什么激活函数?

A，不需要激活函数，因为Q值可正可负，没有取值范围

B，用sigmoid激活函数，因为Q值介于0和1之间

C，用ReLU激活函数，因为Q值非负

D，用softmax激活函数，因为DQN的输出是一个概率分布

6，多臂赌博机是单步强化学习的经典范例，ε贪心算法和softmax算法用于处理什么问题?

A， 探索-利用问题

B， 奖励函数优化问题

C， 动作选择问题

D， 状态空间问题

7，DQN（深度 Q 网络）是基于什么的强化学习方法?

A， 基于值的方法

B， 基于策略的方法

C， 基于模型的方法

D， 基于探索的方法

8，TD gradient 是与哪种强化学习方法相关的概念?

A， 基于值的方法

B， 基于策略的方法

C， 基于模型的方法

D， 基于探索的方法

9，在强化学习中，基于策略的方法主要关注什么?

A， 最大化奖励