A = Tensors, $\otimes$ = Tensor Product, $\times_k$ = mode-k product, $\circledast$ = Hadamard product

π - policy, (s,a) = state-action pair

$$Q(\pi|s,a) = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid s_t = s, a_t = a\right]$$

By Bellman equation we get,

$$Q(\pi|s,a) = R^a_{s,s'} + \gamma \sum_{a'} \pi(s',a') Q(\pi|s',a').$$

Expanding recursively we get,

$$Q(\pi|s_t,a_t) = R^{a_t}_{s_t,s_{t+1}} + \gamma \sum_{a_{t+1}} R^{a_{t+1}}_{s_{t+1},s_{t+2}} \pi_{t+1} + \gamma^2 \sum_{a_{t+1}a_{t+2}} \sum R^{a_{t+2}}_{s_{t+2},s_{t+3}} \pi_{t+1}\pi_{t+2} + \cdots$$

we assume that a transition, $s_t \xrightarrow{a_t} s_{t+1}$, exists

Combining terms and considering $C^{(k)} \in \mathbb{C}^{|S\times A|^k}$ is a rank tensor, we get

$$\sum_{s_t,a_t}^{S\times A} Q(\pi|s_t,a_t) = C^{(0)} + \sum_{k=1}^{K} \sum^{S\times A} C^{(k)}_{\mu_{t+1},\ldots,\mu_{t+k}} \pi(\mu_{t+1}) \cdots \pi(\mu_{t+k})$$

has the same form as a Hamiltonian. We can define the optimal policy, π^*, by minimizing that function.

This approach aims to combat the issues causing inconsistency in Q-learning by reposing the problem as a Hamiltonian & solving to find the optimal policy. This allows for the leveraging of quantum properties to consider multiple (s,a) pairs concurrently, thereby deriving a Q-value faster than traditional Q-learning techniques.