Reinforcement learning and quantum mechanics

甄景贤 (King-Yin Yan)

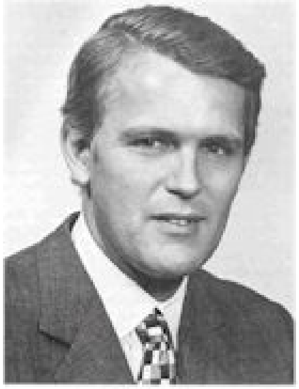
General.Intelligence@Gmail.com

Abstract. The Bellman equation governing dynamic programming (and reinforcement learning) is the discrete-time version of the Hamilton-Jacobi equation governing analytical mechanics. This relation has been known for a long time. If moreover we assume that "states" can be linearly superimposed on each other, this brings to mind the superposition of wave functions in quantum mechanics, governed by the Schrödinger equation. The mysterious connection between Hamilton-Jacobi theory and Schrödinger theory was present from the very beginning of quantum mechanics ("the quantum-classical correspondence"), yet physicists are still unable to explain it to this day.

0 Bellman equation = Hamilton-Jacobi equation

Shortly after Bellman proposed his equation in 1953, Kalman recognized that its differential version is equivalent to the Hamilton-Jacobi equation of classical mechanics.

Richard E Bellman (1920-1984)



Rudolf Kálmán (1930-2016)

(1)

The **Bellman condition** says: "if we cut off a tiny bit from the endpoint of the optimal path, the remaining path is still an optimal path between the new endpoints."

Bellman equation (1953):

$$\boxed{\text{Bellman equation}} \quad \overset{\text{value of entire path}}{U^*(\boldsymbol{x})} = \max_{\boldsymbol{u}}\{ \overset{\text{reward of choosing } \boldsymbol{u} \text{ at current state}}{R(\boldsymbol{u})} + \overset{\text{value of rest of path}}{U^*(\boldsymbol{x}_{t+1})} \}| \tag{2}$$

The meaning is that: the difference in utility (U) of the current state compared with the next state is equal to the reward of the current state.

Hamilton-Jacobi equation (1830's):

$$\boxed{\text{Hamilton-Jacobi}} \quad \frac{\partial U^*}{\partial t} = -H \tag{3}$$

This means that the differential in time of the utility (U) is the energy (H) of the current state. "Energy" corresponds to instantaneous reward.

Simply put, **utility** is the integral of instantaneous **rewards** over time:

$$\boxed{\text{utility } U} = \int \boxed{\text{reward } R} \, dt \tag{4}$$

In **analytical mechanics** the **Lagrangian** (L) is a measure of energy and its time-integral is called the **action**:

$$\boxed{\text{action } S} = \int L dt \tag{5}$$

Nothing new so far.

1 Schrödinger equation

Interestingly, if we want the "mental state" to be composed of a **superposition** of pure states, we may perform the **Fourier transform** on the pure states and add them together. This immediately brings to mind the superposition of wave functions ($\varPsi$) in quantum mechanics, in particular the Schrödinger equation:

$$\boxed{\text{Schrödinger}} \quad i\frac{\partial \varPsi}{\partial t} = \hat{H}\varPsi. \tag{6}$$

Note that $\hat{H}$ is different from the H in (3).

This is very interesting because the Schrödinger equation bears a close resemblance to the Hamilton-Jacobi equation, which has been recognized since the beginning of quantum mechanics, but no one has ever been able to elucidate this relation. Bohm has made a controversial attempt at this goal, which we will discuss later.

Erwin Schrödinger (1887-1961)

David Bohm (1917-1992)

(7)

Note: the Schrödinger equation can be written in this more "physical" form, with explicit dependence on position ($\boldsymbol{x}$) and time (t):

$$i\hbar\frac{\partial}{\partial t}\Psi(x,t) = \left[V(x,t) - \frac{\hbar^2}{2m}\nabla^2\right]\Psi(x,t), \tag{8}$$

whereas the Hamilton-Jacobi-Bellman equation should be (in my opinion) dependent of the state Ψ:

$$\frac{\partial U(\Psi)}{\partial t} = -H(\Psi). \tag{9}$$

For example,

$$\textit{I love you} \tag{10}$$

is not a "state", but

$$\textit{John loves Mary} \wedge \textit{Mary loves Pete} \wedge \textit{John is unhappy} \wedge .... \tag{11}$$

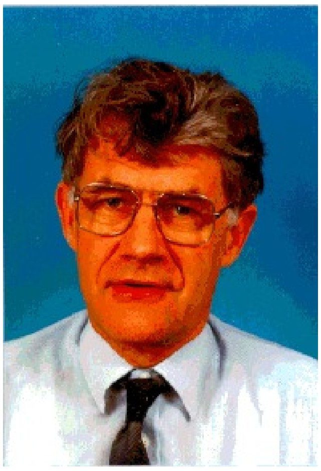
is a state. That is because we take it that (11) completely describes a state of affairs in the **world**, whereas (10) is just a partial description of a facet of the world. In other words, the difference between a **state** and a **proposition**. This distinction is quite obvious from the perspective of logic and reinforcement learning, but physicists might have overlooked it.

From the AI perspective, the meaning of Ψ as "state" is clear, but the question is what is the meaning of $\boldsymbol{x}$, since the cognitive space is not physical space? There may be several possibilities:

- Bohm's theory, which relates the Hamilton-Jacobi equation with Schrödinger's equation
- QFT (quantum field theory), **second quantization** and the creation / annihilation operators
- The Schrödinger equation is a kind of **linearization** of the Hamilton-Jacobi equation?
- There are many other quantization schemes: path integral, group deformation, perturbative QFT, algebraic QFT, ...

1.1 Bohm's theory

The theory is developed and deepened by Bohm, Hiley, de Gosson, ... and others.



Basil J Hiley (1935-)

Maurice de Gosson (1948-)

(12)

Starting from the Schrödinger equation for a single particle:

$$i\hbar\frac{\partial}{\partial t}\Psi = -\frac{\hbar^2}{2m}\nabla^2\Psi + V\Psi, \tag{13}$$

consider its polar-form solution:

$$\Psi(\boldsymbol{r},t) = R(\boldsymbol{r},t)e^{\frac{i}{\hbar}\Phi(\boldsymbol{r},t)} \tag{14}$$

whose real and imaginary parts lead to:

$$\begin{cases} R(\frac{\partial\Phi}{\partial t} + \frac{(\nabla_{\boldsymbol{r}}\Phi)^2}{2m} + V) - \frac{\hbar^2}{2m}\nabla^2_{\boldsymbol{r}}R = 0 \\ \\ \frac{\partial R^2}{\partial t} + \mathrm{div}(\frac{\nabla_{\boldsymbol{r}}\Phi}{m}R^2) = 0. \end{cases} \tag{15}$$

The second equation is recognized as the continuity equation describing the time-evolution of a probability density:

$$\frac{\partial\rho}{\partial t} + \mathrm{div}(\rho v) = 0 \tag{16}$$

whereas the first equation is recognized as the Hamilton-Jacobi equation:

$$\frac{\partial\Phi}{\partial t} + \frac{(\nabla_{\boldsymbol{r}}\Phi)^2}{2m} + V - \frac{\hbar^2}{2m}\frac{\nabla^2_{\boldsymbol{r}}R}{R} = 0 \tag{17}$$

with a special Hamiltonian:

$$H^\Psi = H + Q^\Psi \tag{18}$$

where

$$Q^\Psi = -\frac{\hbar^2}{2m}\frac{\nabla^2_{\boldsymbol{r}}R}{R} \tag{19}$$

is known as the **quantum potential**. Q^Ψ is very much unlike a usual potential and is intrinsically **non-local**. Its meaning is unclear (to me at least).

1.2 Schrödinger equation = linearization?

The **state linearization** of a dynamical system means to transform from the non-linear system:

$$\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}) \tag{20}$$

to the (locally) linear system:

$$\dot{\boldsymbol{x}} = A\boldsymbol{x} + \boldsymbol{c} \tag{21}$$

The Schrödinger equation seems to be a particularly simple linear equation:

$$i\dot{\Psi} = \hat{H}\Psi \tag{22}$$

where $\hat{H}$ is a linear operator. Could it be that it is the linearized version of the Hamilton-Jacobi equation:

$$\dot{U}(\Psi) = -H(\Psi) \quad ? \tag{23}$$

where we assumed that U is expressed directly as a function of time. If on the other hand U's time-dependence is expressed through Ψ:

$$\dot{U}(\Psi)\dot{\Psi} = -H(\Psi). \tag{24}$$

$$\dot{\Psi} = -\frac{H(\Psi)}{\dot{U}(\Psi)} \quad ? \tag{25}$$

The **relativistic** Klein-Gordon equation:

$$\boxed{\text{Klein-Gordon}} \qquad -\frac{\partial^2 \Phi}{\partial^2 t} = (-\nabla^2 + m^2)\Phi \tag{26}$$

appears to be 2nd-order in time, but that arised only because

$$E = \sqrt{\boldsymbol{p}^2 c^2 + m^2 c^4} \tag{27}$$

and they had to deal with the operator $\boldsymbol{p}$ appearing within the $\sqrt{}$ sign.
Its original consideration is:

$$\frac{\partial \Phi}{\partial t} = E\Phi \tag{28}$$

which has the same form as the Schrödinger equation.

1.3 Imaginary time $\rightsquigarrow$ stochastic processes

When imaginary time is substituted into the Schrödinger equation, out pops the heat / diffusion equation:

$$\frac{\partial u}{\partial \tau} = D\Delta u. \tag{29}$$

2 Spectral theory

2.1 Fourier transform

All of **transform theory** can be summarized as the following equation:

$$\boxed{\text{Transform theory}} \qquad y = \sum_n \langle \phi_n, y \rangle \phi_n \tag{30}$$

where $\langle \phi_n, y \rangle$ represents the **analysis** stage where the input function is decomposed in terms of a basis set, and the entire $\sum$ represents the **synthesis** stage where the signal is reconstructed.

Hilbert created the spectral theory for **operators**:

(31)

David Hilbert (1862-1943)

2.2 States and observables

2.3 Bra and ket notation

2.4 Lattice of projectors

3 Quantization (quantum field theory)

3.1 Creation and annihilation operators

The **general scheme** for quantization is to factor the Hamiltonian as a product of 2 mutually adjoint operators: **annihilation**, which lowers the energy level of H, and **creation**, which raises the energy level. These 2 operators completely describe the **spectrum** of H.

Starting from the classical Hamiltonian of a single particle with mass m and charge e interacting with an electromagnetic field with vector potential $\boldsymbol{A}$ and scalar potential ϕ:

$$H = \frac{1}{2m}(\boldsymbol{p} - e\boldsymbol{A})^2 + e\phi, \tag{32}$$

invariance under gauge transformations dictates that:

$$\nabla^2 \boldsymbol{A} - \frac{\partial^2 \boldsymbol{A}}{\partial t^2} = 0 \tag{33}$$

whose general solution is:

$$\boldsymbol{A}(\boldsymbol{x}, t) = (\frac{1}{2\pi})^2 \int d^3\boldsymbol{k} d\omega \sum_{\lambda=1}^{2} a_\lambda(\boldsymbol{k}) e^{i(\boldsymbol{k}\cdot\boldsymbol{x} - \omega t)} \boldsymbol{\epsilon}_\lambda(\boldsymbol{k}) \delta(k^2). \tag{34}$$

This can be seen as a **Fourier transform** of $\boldsymbol{A}$.

4 Tropical geometry

$$(\times, +) \to (+, \max) \tag{35}$$

$$\begin{aligned} x \oplus y &= \log_t(t^x + t^y) \\ x \otimes y &= x + y \end{aligned} \tag{36}$$

5 Q-learning

In AI reinforcement learning there is an oft-employed trick known as Q-learning. Q value is just a variation of U value; there is a U value for each state, and Q is the **decomposition** of U by all the actions available to that state. In other words, Q is the utility of doing action $\boldsymbol{u}$ in state $\boldsymbol{x}$. The relation between Q and U is:

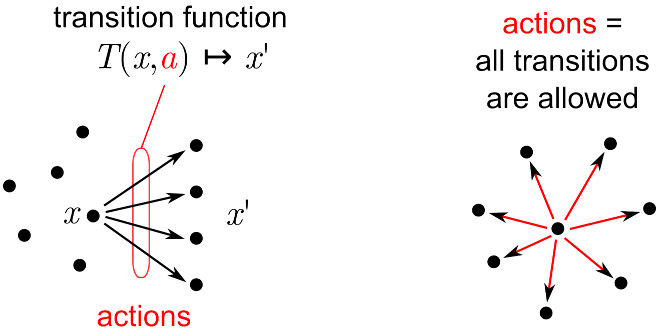
$$U(\boldsymbol{x}) = \max_{\boldsymbol{u}} Q(\boldsymbol{x}, \boldsymbol{u}) \tag{37}$$

The advantage of Q is the ease of learning. We just need to learn the value of actions under each state. This is so-called "**model free learning**".

5.1 Actions = cognitive state-transitions = "thinking"

In our system there are 2 things that need to be learned:

1. The transition function $\boldsymbol{F} : \boldsymbol{x} \mapsto \boldsymbol{x}'$. $\boldsymbol{F}$ represents the **knowledge** that constrains thinking. In other words, the learning of $\boldsymbol{F}$ is the learning of "static" knowledge.
2. Find the optimal trajectory of the state $\boldsymbol{x}$. This corresponds to optimal "thinking" under the constraints of static knowledge.



(38)

In traditional reinforcement learning (left view), the system chooses an action $\boldsymbol{a}$, and the transition function $\boldsymbol{F}$ gives the probability of reaching each state $\boldsymbol{x}_i$ given action $\boldsymbol{a}$. In our model (right view), all possible cognitive states are potentially **reachable** from any other state, and therefore the action a coincides with the next state x'.

References