Fidelity-Based Probabilistic Q-Learning for Control of Quantum Systems

Presented by Hristo Tonchev, Physics faculty 3rd year

Q-learning

$$Q^{\pi}_{(s,a)} = r^{a}_{s} + \gamma \sum_{s'} p^{a}_{ss'} \sum_{a'} p^{\pi}(s', a') Q^{\pi}_{(s',a')}$$

The one step update rule may be described as

$$Q(s_t, a_t) \leftarrow (1 - \alpha_t) Q(s_t, a_t) + \alpha_t (r_{t+1} + \gamma \max_{a'} Q(s_{t+1}, a')).$$

The optimal Q function satisfies the Bellman equation

$$Q^{*}_{(s,a)} = \max_{\pi} Q_{(s,a)} = r^{a}_{s} + \gamma \sum_{s'} p^{a}_{ss'} \max_{a'} Q^{\pi}_{(s',a')}.$$

To approach the optimal policy we have the following expression

$$\pi^{*} = \arg\max_{\pi} Q^{\pi}_{(s,a)} (\forall s \in S)$$

Definition of our main Problem

Exploitation vs Exploration

- ε- greedy
- Softmax
- Simulated annealing
- Probabalistic Q-learning (PQL)
- Fidelity-Based PQL (FPQL

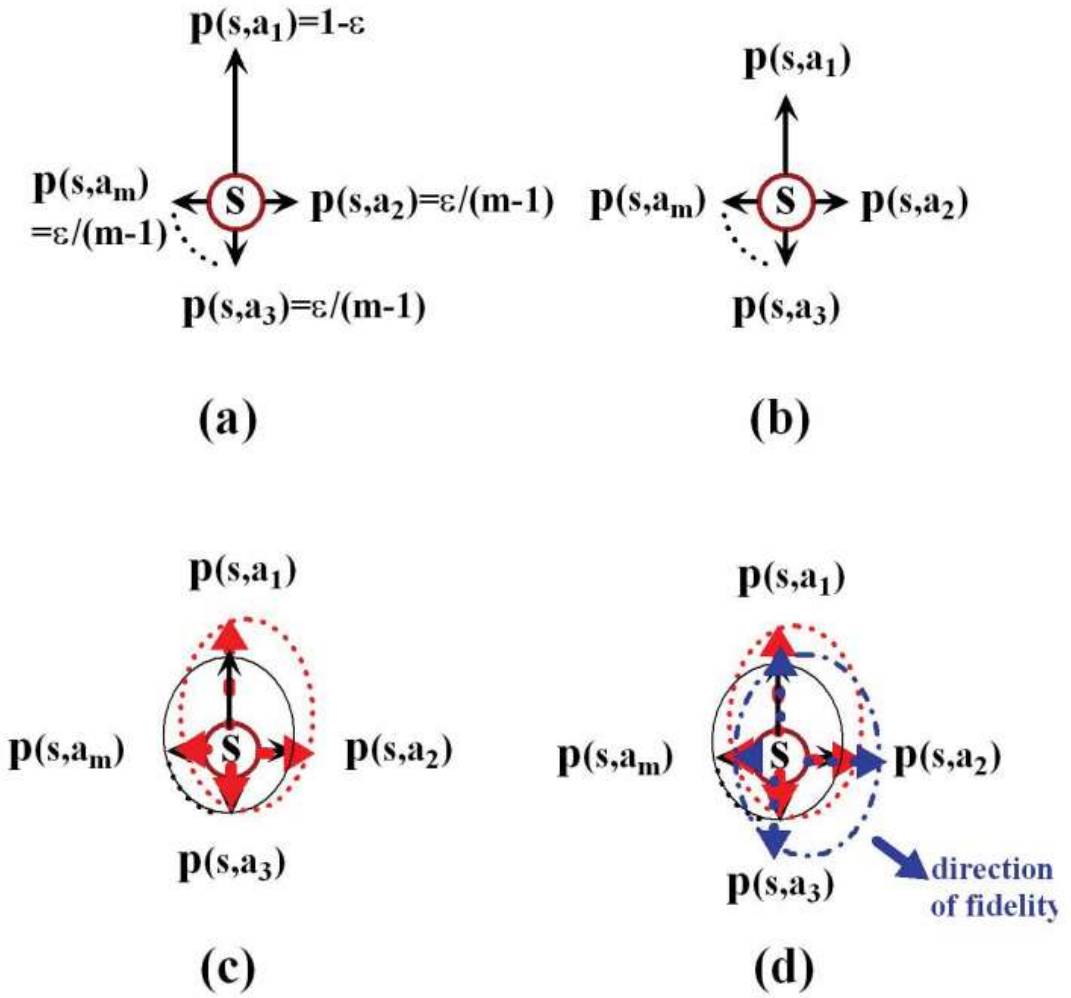


Fig. 1. Illustration of the idea of probabilistic action selection method and the effect of fidelity. (a) ϵ-greedy method. (b) Softmax method. (c) Basic probabilistic action selection method. (d) Fidelity-based probabilistic action selection method.

Fidelity

- In quantum mechanics, notably in quantum information theory, fidelity is a measure of the "closeness" of two quantum states.
- An evolving state of the controlled system can be expanded in terms of the eigenstates in the set $D = \{|\phi_i\rangle\}_{i=1}^{N}$.

$$|\psi(t)\rangle = \sum_{i=1}^{N} c_i(t)|\phi_i\rangle$$

- We define Fidelity as follows

$$F(|\psi^a\rangle, |\psi^b\rangle) = |\langle\psi^a|\psi^b\rangle| = |\sum_{i=1}^{N}(c_i^a)^* c_i^b|$$

Probabilistic Action Selection and Reinforcement Strategy

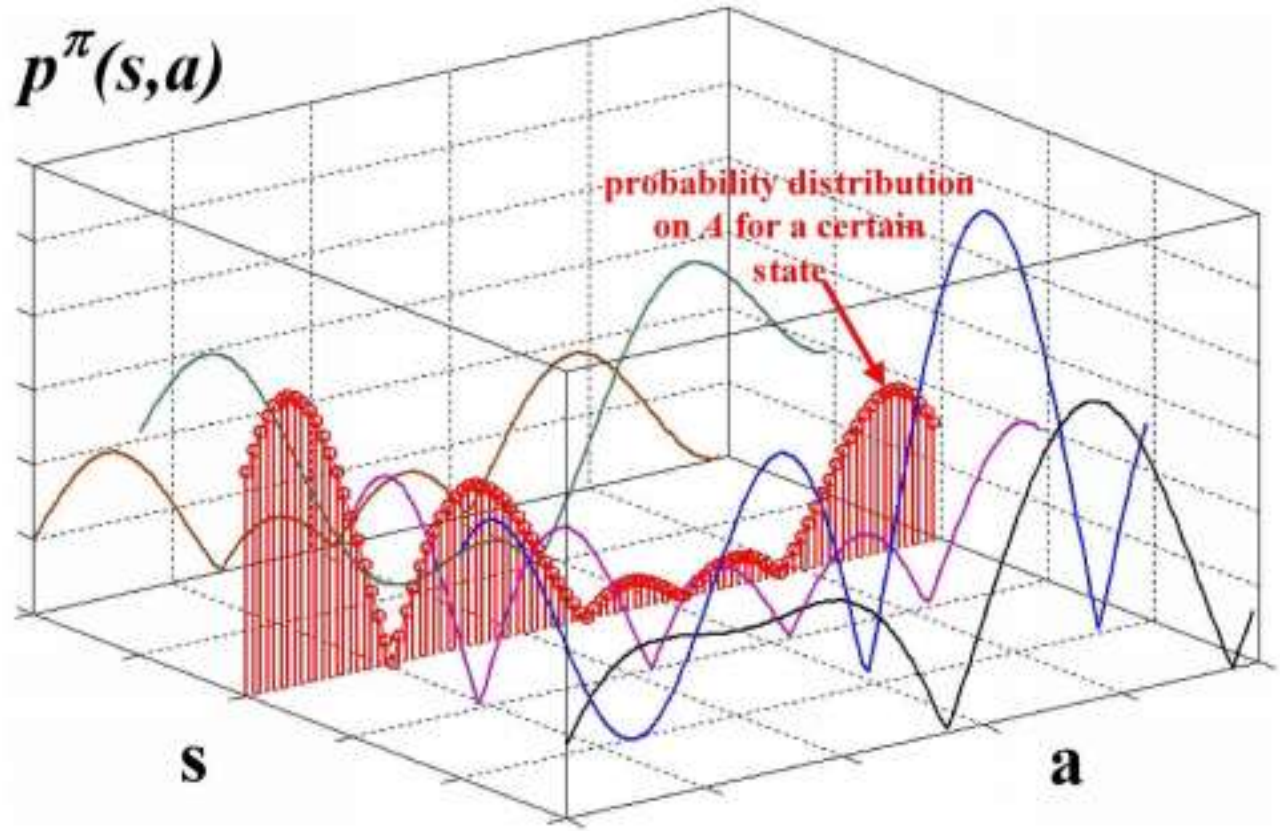
Definition 1: The probability distribution on the state-action space (discrete case) of a RL problem is characterized by a probability mass function defined on the state set S and the action set $A = \bigcup_{s \in S} A_{(s)}$, where $A_{(s)}$ is the set of all the permitted actions for state s. For any $s \in S$ and $a \in A_{(s)}$, the probability mass function is defined as $p(s, a) \geq 0$ and for a certain state s, it satisfies

$$\sum_{a \in A_{(s)}} p(s, a) = 1.$$

$$\pi : P^{\pi} = (p^{\pi}(s, a))_{n \times m}$$

$$a_s^{\pi} = f^{\pi}(s) = \begin{cases} a_1 & \text{with probability } p^{\pi}(s, a_1) \\ a_2 & \text{with probability } p^{\pi}(s, a_2) \\ \vdots & \\ a_m & \text{with probability } p^{\pi}(s, a_m). \end{cases}$$

$$\begin{pmatrix} & \mathbf{a_1} & \mathbf{a_2} & \cdots & \mathbf{a_m} \\ \mathbf{s_1} & \begin{bmatrix} Q_{(s_1,a_1)} \\ p^{\pi}_{(s_1,a_1)} \end{bmatrix} & \begin{bmatrix} Q_{(s_1,a_2)} \\ p^{\pi}_{(s_1,a_2)} \end{bmatrix} & \cdots & \begin{bmatrix} Q_{(s_1,a_m)} \\ p^{\pi}_{(s_1,a_m)} \end{bmatrix} \\ \mathbf{s_2} & \begin{bmatrix} Q_{(s_2,a_1)} \\ p^{\pi}_{(s_2,a_1)} \end{bmatrix} & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ \mathbf{s_n} & \begin{bmatrix} Q_{(s_n,a_1)} \\ p^{\pi}_{(s_n,a_1)} \end{bmatrix} & \cdots & \cdots & \begin{bmatrix} Q_{(s_n,a_m)} \\ p^{\pi}_{(s_n,a_m)} \end{bmatrix} \end{pmatrix}.$$


$p^{\pi}(s,a)$
probability distribution
on A for a certain
state
s
a

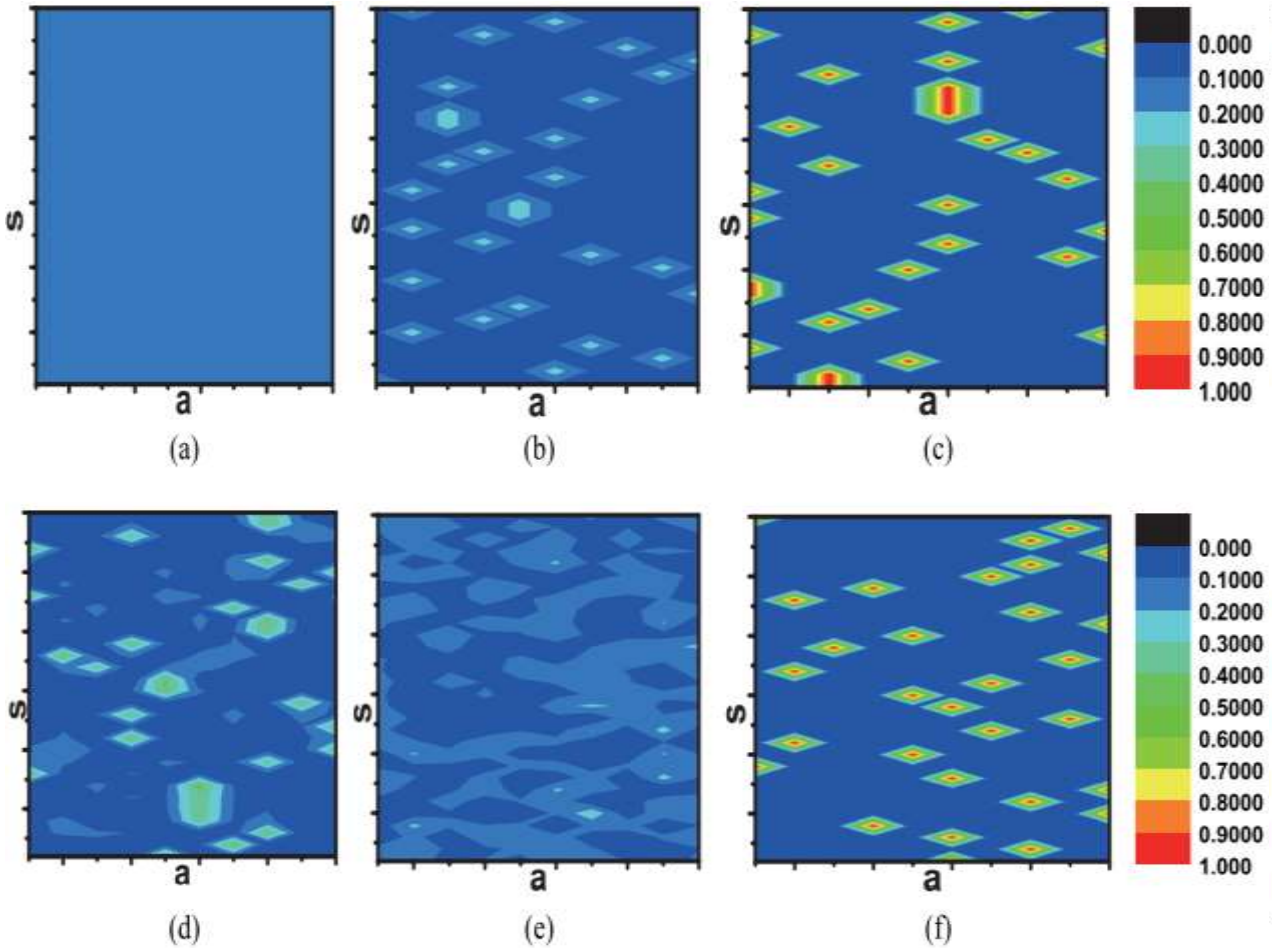
- The goal of PQL is to learn a mapping from states to actions.
- The agent needs to learn a policy π to maximize the expected
- sum of discounted reward for each state

$$Q^{\pi}_{(s,a)} = \sum_{a \in A_{(s)}} p^{\pi}(s,a)\left[r^{a}_{s} + \gamma \sum_{s'} p^{a}_{ss'} Q^{\pi}_{(s',a')}\right].$$

$$Q(s_t, a_t) \leftarrow (1-\alpha_t) Q(s_t, a_t) + \alpha_t (r_{t+1} + \gamma \max_{a'} Q(s_{t+1}, a')).$$

$$p(s_t, a_t) \leftarrow p(s_t, a_t) + k(r_{t+1} + \max_{a'} Q(s_{t+1}, a'))$$

The algorithm boosts the following merits
1. The learning algorithm
possesses more reasonable credit assignment using a probabilistic method and the action selection method is more natural
without too much difficulty for parameter setting.
2. The
method provides a natural re–exploring mechanism



Algorithm 1 Probabilistic Q-Learning

Initialize $Q(s, a)$ arbitrarily
Initialize the policy $\pi : P^{\pi} = (p^{\pi}(s, a))_{n \times m}$ to be evaluated
repeat (for each episode)
  Initialize $t = 1$, s_t,
  repeat (for each step of episode)
    $a_t \leftarrow$ action a_i with probability $p(s_t, a_i)$ for s_t
    Take action a_t, observe reward r_{t+1}, and next state s_{t+1}
    $Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha_t \delta^{Q}_{t+1}$
    where $\delta^{Q}_{t+1} = r_{t+1} + \gamma \, max_{a'} Q(s_{t+1}, a') - Q(s_t, a_t)$
    $p(s_t, a_t) \leftarrow p(s_t, a_t) + k(r_{t+1} + max_{a'} Q(s_{t+1}, a'))$
    Normalize $\{p(s_t, a_i)|_{i=1,2,\ldots,m}\}$
    $t \leftarrow t + 1$
  until s_{t+1} is terminal
until the learning process ends

Reasons for including fidelity.
For most complex reinforcement learning problems, the direction of achieving the objective is always delayed due to the lack of feedback information during the learning process unless the agent reaches the target state.

The updating rule of fidelity-based PQL for the Q function Is the same as for PQL, the difference is in the update rule for the probability distribution

$$p(s_t, a_t) \leftarrow p(s_t, a_t) + k(r_{t+1} + \max_{a'} Q(s_{t+1}, a') + F(s_{t+1}, s_{\text{target}})).$$

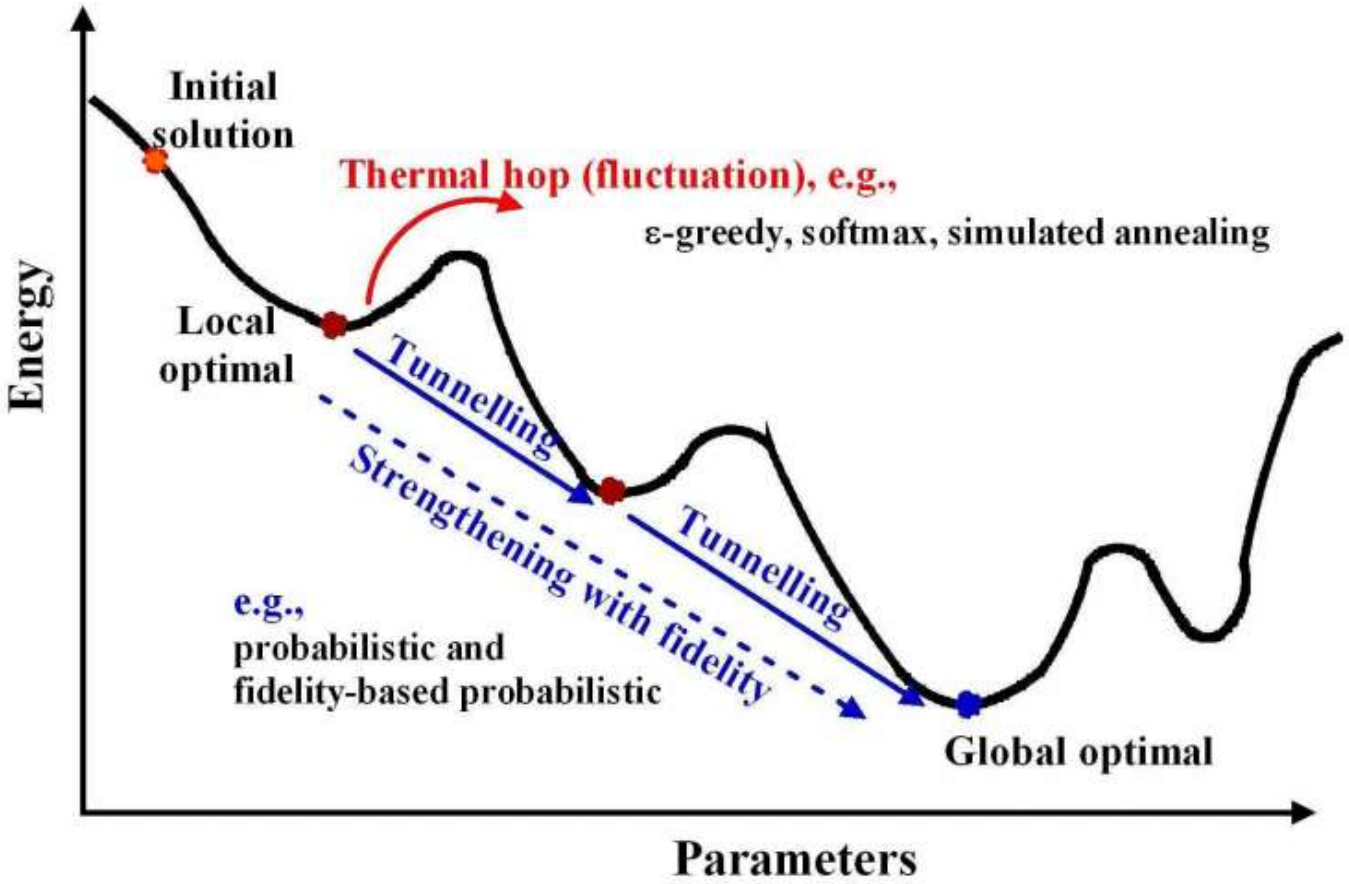
Theorem 1 (Convergence of FPQL): Consider an FPQL agent in a nondeterministic Markov decision process, for every state-action pair s and a, the Q-value $Q_t(s, a)$ will converge to the optimal state-action value function $Q^*(s, a)$ if the following constraints are satisfied.

1) The rewards in the whole learning process satisfy $(\forall s, a)|r_s^a| \leq R$, where R is a finite constant value.
2) A discount factor $\gamma \in [0, 1)$ is adopted.
3) During the learning process, the nonnegative learning rate α_t satisfies

$$\lim_{T\to\infty} \sum_{t=1}^{T} \alpha_t = \infty, \qquad \lim_{T\to\infty} \sum_{t=1}^{T} \alpha_t^2 < \infty.$$

Algorithm 2 Fidelity-based Probabilistic Q-learning

Initialize $Q(s, a)$ arbitrarily
Initialize the policy $\pi : P^{\pi} = (p^{\pi}(s, a))_{n \times m}$ to be evaluated
repeat (for each episode)
  Initialize $t = 1$, s_t
  repeat (for each step of episode)
    $a_t \leftarrow$ action a_i with probability $p(s_t, a_i)$ for s_t
    Take action a_t, observe reward r_{t+1}, and next state s_{t+1}
    $Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha_t \delta^{Q}_{t+1}$
    where $\delta^{Q}_{t+1} = r_{t+1} + \gamma \, max_{a'} Q(s_{t+1}, a') - Q(s_t, a_t)$
    $p(s_t, a_t) \leftarrow p(s_t, a_t) + k\delta^{p}_{t+1}$
    where $\delta^{p}_{t+1} = r_{t+1} + max_{a'} Q(s_{t+1}, a') + F(s_{t+1}, s_{\text{target}})$
    Normalize $\{p(s_t, a_i)|_{i=1,2,\ldots,m}\}$
    $t \leftarrow t + 1$
  until s_{t+1} is terminal
until the learning process ends



The difference between the proposed fidelity-based probabilistic exploration strategy and the existing exploration strategies can be explained from a point of view of physical mechanism

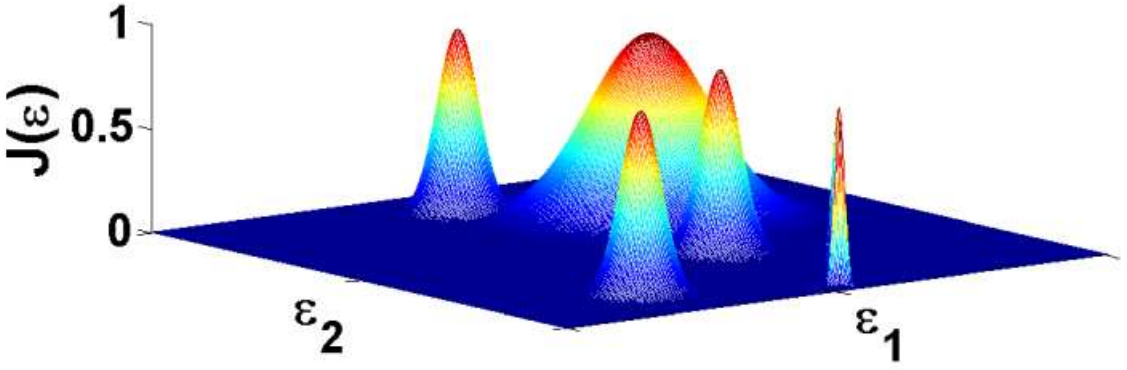
FIDELITY-BASED PQL FOR LEARNING CONTROL OF QUANTUM SYSTEMS

A. Learning Control of Quantum systems

Introducing a control $\varepsilon(t) \in L^2(\mathbf{R})$ acting on the system via a time-independent interaction Hamiltonian H_I and denoting $|\psi(t=0)\rangle$ as $|\psi_0\rangle$, $C(t) = (c_i(t))_{i=1}^N$ evolves according to the Schrödinger equation [32]

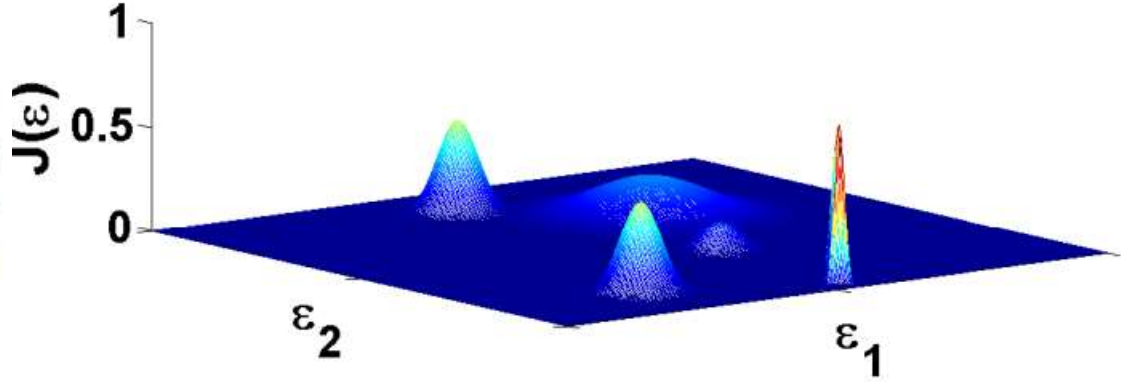
$$\begin{cases} \imath\hbar\dot{C}(t) = [A + \varepsilon(t)B]C(t) \\ C(t=0) = C_0 \end{cases}$$

$$|\psi(t_2)\rangle = U(t_1 \to t_2)|\psi(t_1)\rangle$$



(a)

Assume that the control set $\{\varepsilon_j, j = 1, \ldots, m\}$ is given. Every control ε_j corresponds to a unitary operator U_j. The task of learning control is to find a control sequence $\{\varepsilon_l, l = 1, 2, 3, \ldots\}$ where $\varepsilon_l \in \{\varepsilon_j, j = 1, \ldots, m\}$ to drive the quantum system from an initial state $|\psi_0\rangle$ to the target state $|\psi_f\rangle$.



(b)

Quantum Controlled Transition Landscapes

- A control landscape is defined as the map between the time-dependent control Hamiltonian and associated values of the control performance functional. Most quantum control problems can be formulated as the maximization of an objective performance function. For example, as shown in Fig. 5, the performance function $J(\varepsilon)$ is defined as the functional of the control strategy $\varepsilon = \varepsilon_i, i = 1, 2, \ldots, M$, where M is a positive integer that indicates the number of the control variables ($M = 2$ for the case shown in Fig. 5).

Although quantum control applications may span a variety of objectives, most of them correspond to maximizing the probability of transition from an initial state to a desired final state

For the state transition problem with t ∈ [0, T], we define the quantum controlled transition landscape as

$$J(\varepsilon) = \mathrm{tr}(U_{(\varepsilon,T)}|\psi_0\rangle\langle\psi_0|U^{\dagger}_{(\varepsilon,T)}|\psi_f\rangle\langle\psi_f|)$$

The objective of the learning control system is to find a global optimal control strategy ε^*, which satisfies

$$\varepsilon^* = \mathrm{argmax}_{\varepsilon} J(\varepsilon).$$

If the dependence of $U_{(T)}$ on ε is suppressed (see [42]), (18) can be reformulated as

$$J(U) = \mathrm{tr}(U_{(T)}|\psi_0\rangle\langle\psi_0|U^{\dagger}_{(T)}|\psi_f\rangle\langle\psi_f|).$$

Theorem 2: For the quantum control problem defined with the dynamic control landscape (18) and the kinematic control landscape (20), respectively, the properties of the solution sets of the quantum controlled transition landscape are listed as follows.

1) The kinematic control landscape is free of traps (i.e., all critical points of $J_K(U)$ are either global maxima or saddles) if the operator U can be any unitary operator (i.e., the system is completely controllable).
2) The dynamic control landscape is free of traps if: 1) the operator U can be any unitary operator and 2) the Jacobian $\delta U_{(\varepsilon,T)}/\delta\varepsilon$ has full rank at any ε.

Learning Control of a Spin-(1/2) Quantum System

The spin-(1/2) system is a typical 2-level quantum system and has important theoretical implications and practical applications. Its Bloch vector can be visualized on a 3-D Bloch sphere as shown in Fig. 6. The state of the spin-(1/2) quantum system $|\psi\rangle$ can be represented as

$$|\psi\rangle = \cos\frac{\theta}{2}|0\rangle + e^{i\varphi}\sin\frac{\theta}{2}|1\rangle$$

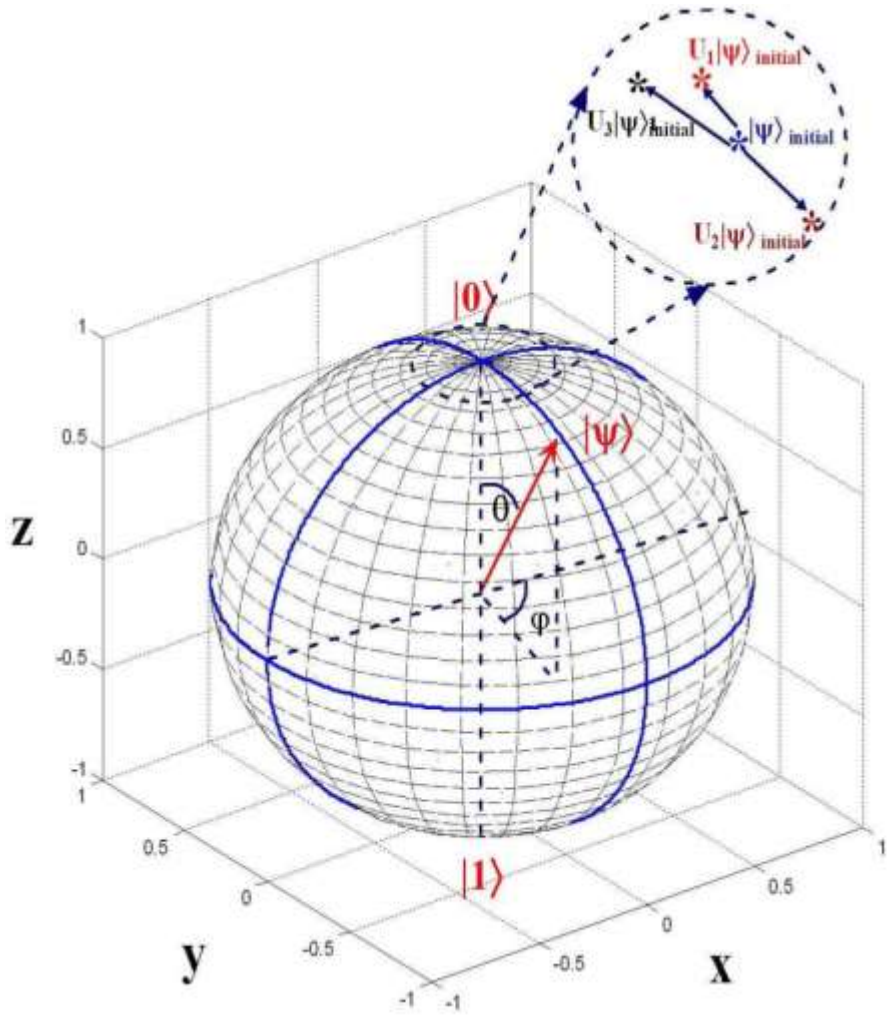
where $\theta \in [0, \pi]$ and $\varphi \in [0, 2\pi]$ are polar angle and azimuthal angle, respectively, which specify a point $\vec{a} = (x, y, z) = (\sin\theta\cos\varphi, \sin\theta\sin\varphi, \cos\theta)$ on the unit sphere in $\mathbf{R}^3$.

The propagators $\{U_i, i = 1, 2, 3\}$ are listed as follows:

$$U_1 = e^{-iI_z\frac{\pi}{15}}$$
$$U_2 = e^{-i(I_z+0.5I_x)\frac{\pi}{15}}$$
$$U_3 = e^{-i(I_z-0.5I_x)\frac{\pi}{15}}$$

where

$$I_z = \frac{1}{2}\begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}, \quad I_x = \frac{1}{2}\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix}.$$



Now the control objective is to control the spin-(1/2) system from the initial state $(\theta = (\pi/60), \varphi = (\pi/30))$ to the target state $(\theta = (41\pi/60), \varphi = (29\pi/30))$ with minimized control

$S = \{s_i = |\psi_i\rangle\}, i = 1, 2, \ldots, n$ and the action set is $A = \{a_j = u_j\}, j = 1, 2, \ldots, m$. The experiment settings for these algorithms are listed as follows: $r = -1$ for each control step until it reaches the target state, then it gets a reward of $r = 1000$; the discount factor $\gamma = 0.99$, the learning rate $\alpha = 0.01$ and the Q-values are all initialized as 0. For PQL and fidelity-based PQL, $k = 0.01$. The ϵ-greedy exploration strategy is used and $\epsilon = 0.1$.

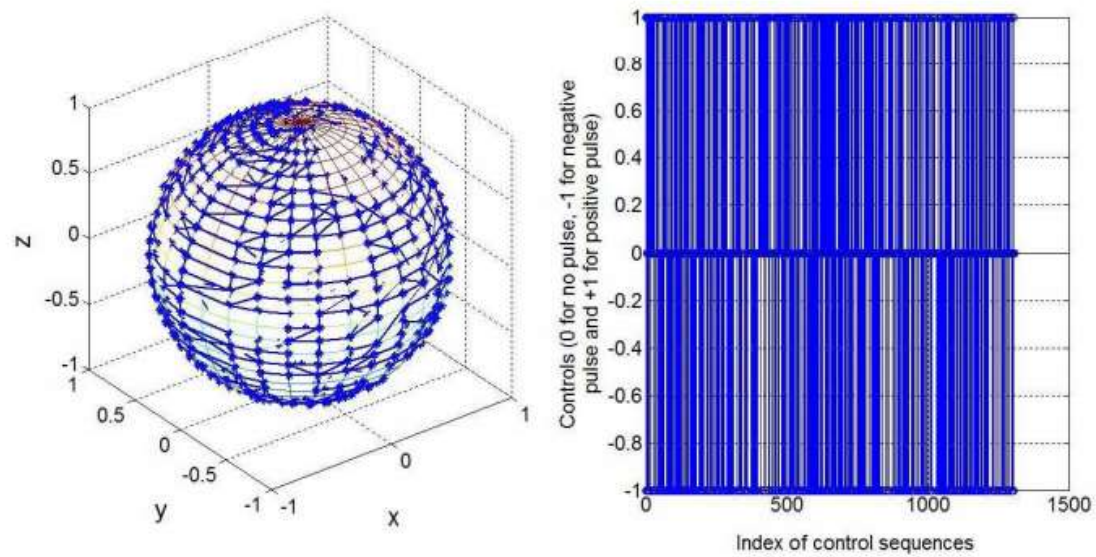


Fig. 7. Demonstration of a stochastic control case without learning. The left figure shows the state transition path and the right figure shows the control sequence used (0 for no pulse, −1 for negative pulse and +1 for positive pulse).

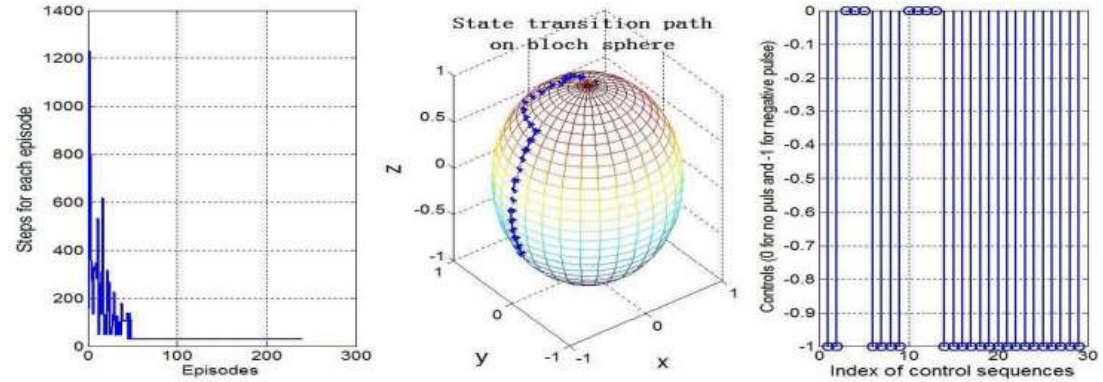


Fig. 8. Learning performance of fidelity-based PQL and the learning results with an optimal control sequence.

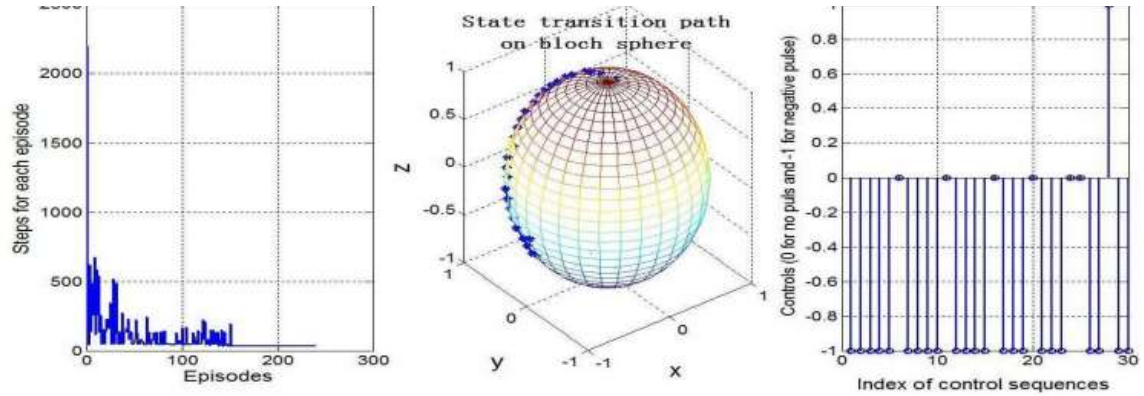


Fig. 9. Learning performance of PQL and the learning results with an optimal control sequence.

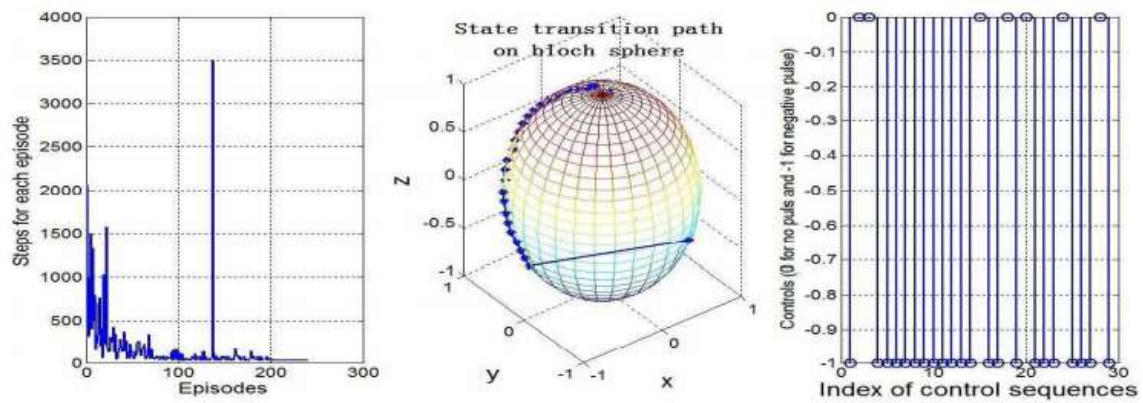


Fig. 10. Learning performance of standard QL with ε-greedy policy and the learning results with an optimal control sequence.

Learning Control of Λ -Type Quantum System

$$|\psi(t)\rangle = c_1(t)|1\rangle + c_2(t)|2\rangle + c_3(t)|3\rangle$$

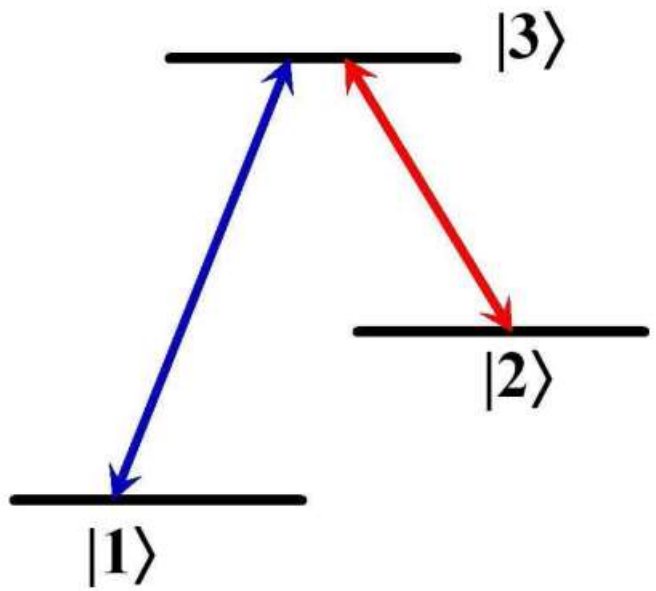
permitted controls are a finite number of (positive or negative) control pulses, i.e., we have the propagators

$$U_E = e^{-i\Delta t(H_0+0.1EH_1)} \tag{30}$$

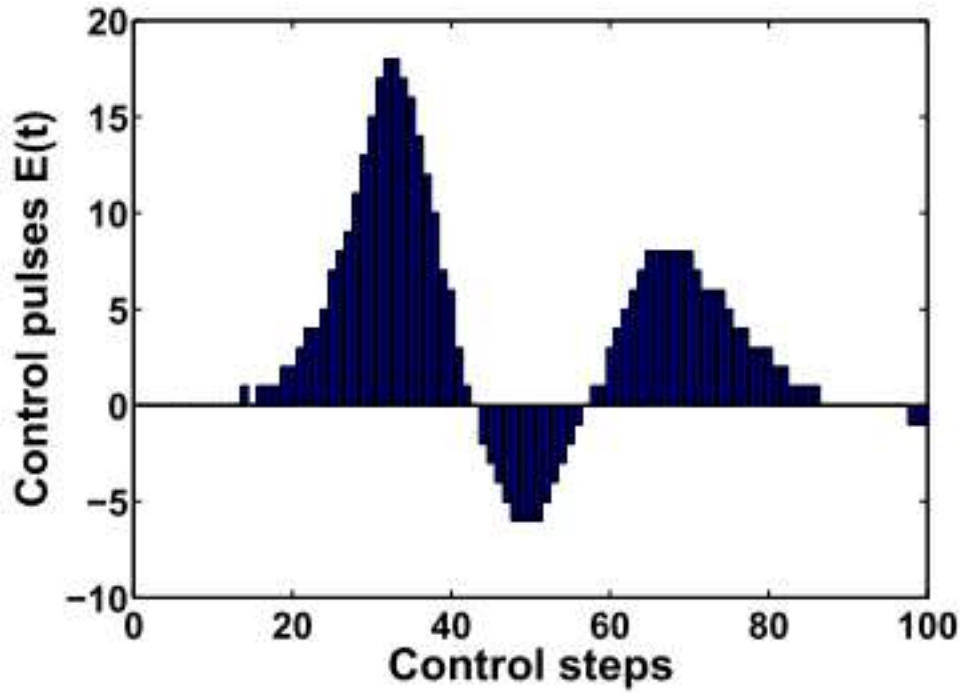
where $\Delta t = 0.1$

$$H_0 = \begin{pmatrix} 1.5 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad H_1 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix} \tag{31}$$

and $E \in \{0, \pm 1, \pm 2, \ldots, \pm 20\}$ is the number of chosen control pulses at a certain control step.



We apply the fidelity-based PQL, PQL, and QL algorithms to this learning control problem, respectively. First, we reformulate the RL problem of controlling a quantum system from an initial state $s_{\text{initial}} = |\psi_{\text{initial}}\rangle$ to a desired target state $s_{\text{target}} = |\psi_{\text{target}}\rangle$ as follows: the number of control steps is fixed as a constant number of 100, so that we can use a virtual state set to construct the state-action space instead of the real state space (with a very high dimension) of the Λ-type system and the state set $S = \{s_i\}, i = 1, 2, \ldots, 101$ and the action set is $A = \{a_j = E_j = j - 21\}, j = 1, 2, \ldots, 41$. The experiment settings for these algorithms are listed as follows: $r = 0$ for each control step until it reaches the target state at the end of the control process where it gets a reward of $r = 1000$; the discount factor $\gamma = 0.99$, the learning rate $\alpha = 0.01$, and the Q-values are all initialized as 0. For PQL and fidelity-based PQL, $k = 0.01$. The ϵ-greedy exploration strategy is used for QL and $\epsilon = 0.1$. The fidelity for a current policy π is defined as $F = |\langle \psi_f^{\pi} | \psi_{\text{target}} \rangle|$.



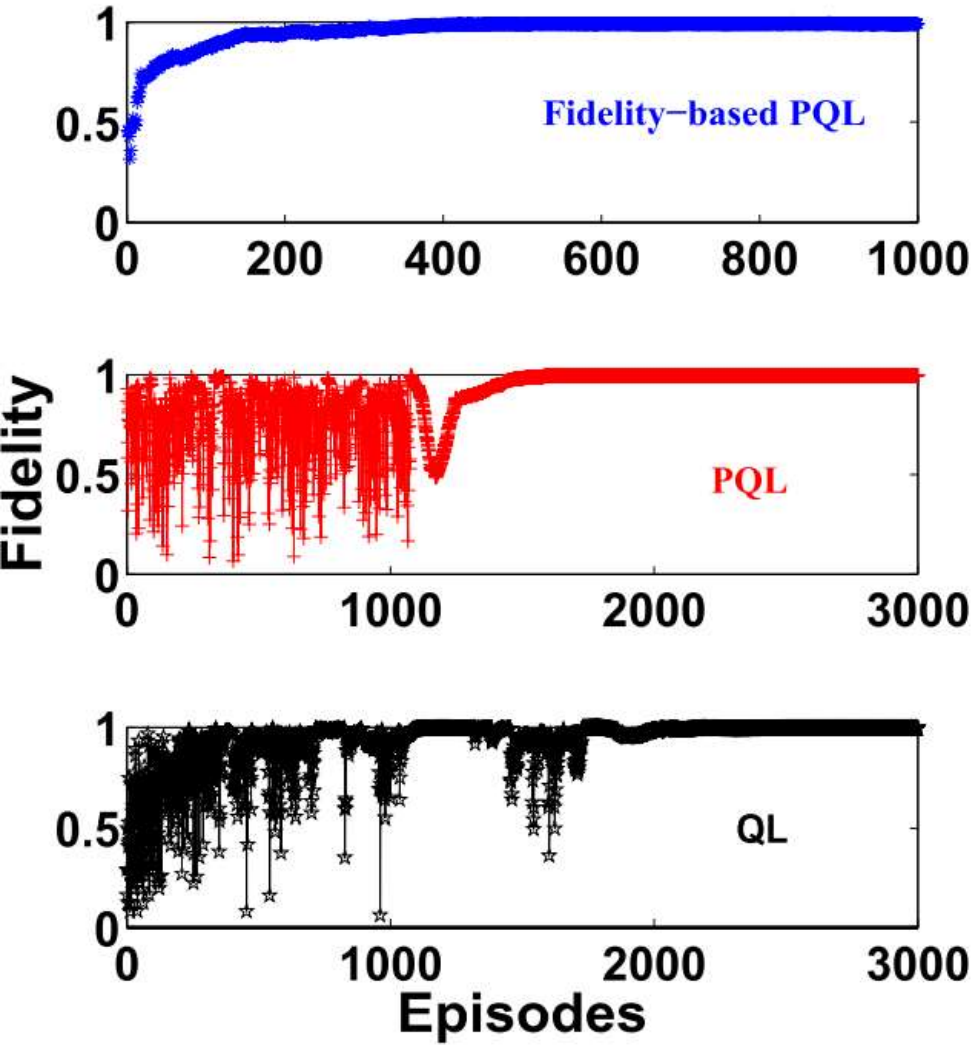


Fig. 15. Learned optimal control pulse sequence.

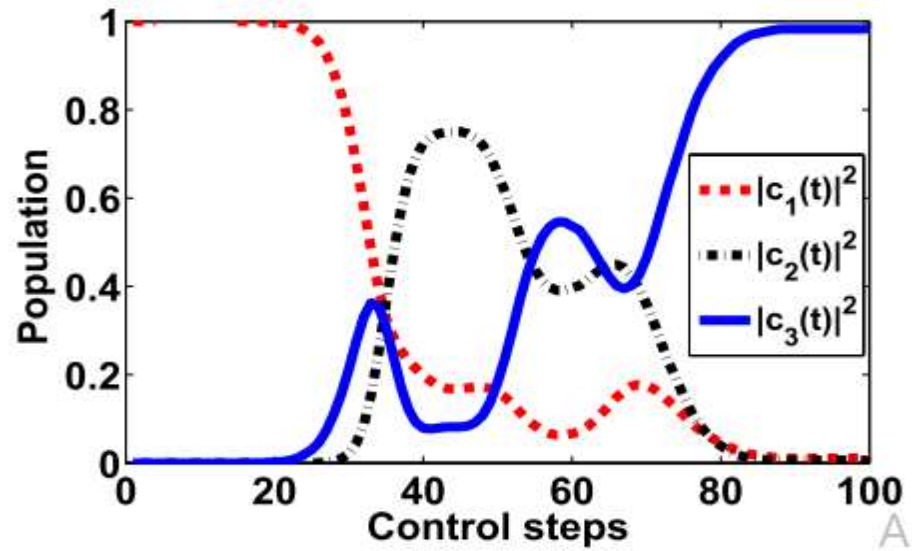


Fig. 16. Population evolution trajectories with the learned optimal control pulse sequence.