Reinforcement learning in different phases of quantum control

Marin Bukov et al. 2018[1]

Robert Klassert

Universität Heidelberg
Quantum and Neural Networks, summer term 2019
under supervision of Martin Gärttner

[1] source of figures if not stated otherwise

What are we going to learn?

1. The quantum control problem
2. Quantum speed limit
3. Q-learning quantum control
4. Learning from reinforcement learning
5. Phase transitions in protocol space
6. Conclusion & Outlook

Motivation for quantum control

Quantum control
Go from initial to target state by tuning available controls

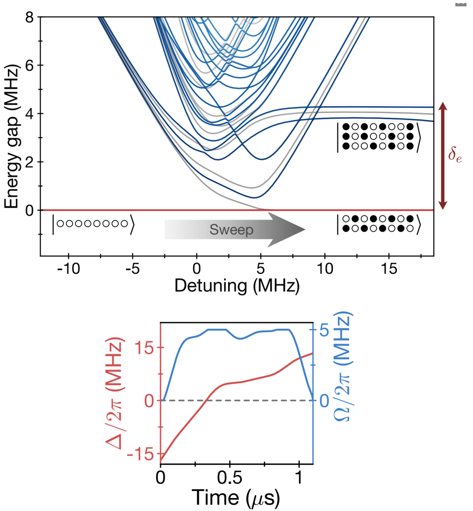
Example: Rydberg cat

Control of detuning Δ and coupling Ω in Rydberg chain: transition from groundstate to GHZ-state

Enables state preparation in

- experiments
- quantum devices

→ fast + high fidelity desired



Omran et al. 2019 [3]

The quantum control problem

Problem:
initial state $|\psi_i\rangle \rightarrow$ final state $|\psi_f\rangle$ under $H(c)$
How to choose $c(t)$ so as to optimize a figure of merit F in time T?
control parameter c, usually: $F = |\langle\psi(T)|\psi_f\rangle|^2$ fidelity

Example: spin flip

$|\psi_i\rangle = |\uparrow\rangle$, $|\psi_f\rangle = |\downarrow\rangle$, $H = c \cdot S^x$
Simple protocol: constant H with $c \cdot T = \pi \rightarrow F = 1$

faster transition → but $|c| < c_{max}$ is bounded by experiment
→ for $T < \pi / c_{max}$ final state $|\psi_f\rangle$ is unreachable

Quantum speed limit (QSL)

Motivation: there is no observable of time! → $\Delta t \geq \frac{\hbar}{\Delta E}$?

Mandelstam-Tamm bound

$$\Delta H \Delta A \geq \frac{\hbar}{2} |\langle \partial_t A \rangle| \text{ with } A = |\psi_i\rangle \langle \psi_i|$$

$$\rightarrow \tau \geq \frac{\hbar \arccos(|\langle \psi_i | \psi(\tau) \rangle|)}{\Delta H} = \tau_{QSL}[2]$$

Interpretation: minimum evolution time between states related to induced energy fluctuations

Example: spin flip

Minimum time $\tau_{QSL} = \frac{\hbar \arccos(|\langle \uparrow | \downarrow \rangle|)}{\hbar c} = \frac{\pi}{c}$ for constant c

- Above the QSL the system is controllable
- Quantum control: time-dependent H → time-averaged ΔH

Reminder: reinforcement learning (RL)

Framework of Markov decision processes (MDP):

- state space $\mathcal{S}$
- action space $\mathcal{A}(s)$
- transition function $p(s', r|s, a)$

RL task:

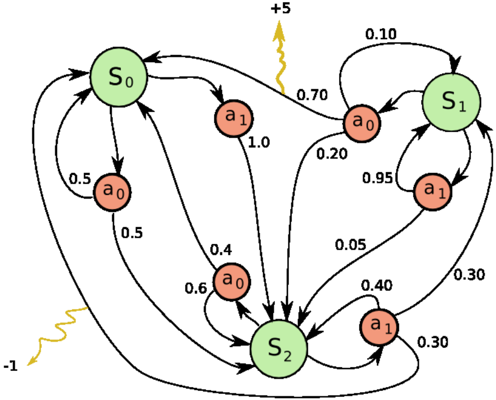
Find $\pi : \mathcal{S} \to \mathcal{A}$ under which

$$R_t = \sum_{k=t+1}^{T,\infty} \gamma^{k-t-1} r_k$$

is maximal from all s_t,

γ: discount factor

graph representation:



towardsdatascience.com, accessed: July 3, 2019

sequential representation:

$s_0,\ a_0,\ r_1,\ s_1,\ a_1,\ r_2, \ldots$

RL setup in the paper

Environment

- Ising model $H = -\sum_{j=1}^{L}[S_{j+1}^z S_j^z + S_j^z + h_x S_j^x]$ with field $h_x \in [-4, 4]$
- $\partial_t |\psi(t)\rangle = H(t)|\psi(t)\rangle$ with $|\psi_i\rangle$, $|\psi_f\rangle$ groundstates at $h_x = \mp 2$

Markov decision process

- episodic ($T =$ finite), undiscounted ($\gamma = 1$) task
- $\mathcal{S} = \{s = [t, h_x(t)]\}$, $\mathcal{A} = \{a = \delta h_x\}$, p is deterministic:

  $$s'(s,a) = [t+1, h_x(t) + \delta h_x] \text{ and } r(s) = \begin{cases} 0 \text{ for } t < T \\ F \text{ for } t = T \end{cases}$$
- initial state $s_0 = [t = 0, h_x = -4] \rightarrow$ protocol depends on history!

Simplification: bang-bang (BB) protocols

$$\mathcal{S} = \{[t, h_x(t) \in \{-4, 4\}]\},\ \mathcal{A} = \{\delta h_x \in \{stay, flip\}\}$$

Reminder: What is Q-learning?

Q-learning is

- a model free (environment is a black box),
- off-policy (learn optimal policy indirectly),
- 1-step time-difference (TD) method
- learning state-action values $Q_\pi(s_t, a_t) = \mathbb{E}[R_t|\pi]$ (control problem)

Trick: learn Q independent of any policy! 1-step approximation:

$$Q(s_t, a_t) \approx r_{t+1} + \gamma \max_{a'} Q(s_{t+1}, a')$$

Iterative update (initial Q's are inaccurate/wrong) with learning rate α:

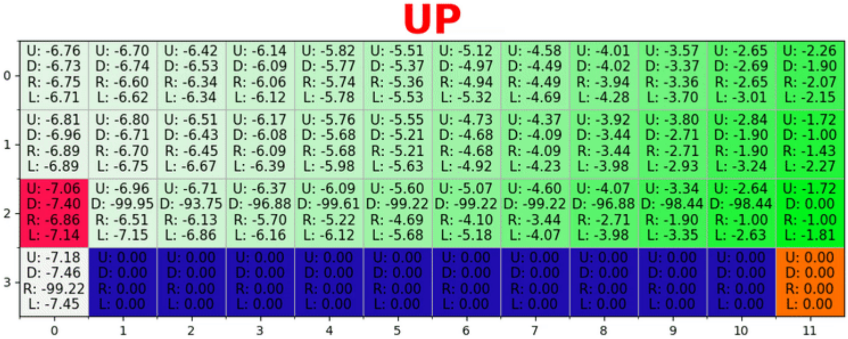
$$Q(s,a) \leftarrow Q(s,a) + \alpha \overbrace{\left[\underbrace{r + \gamma \max_{a'} Q(s', a')}_{\text{target}} - \underbrace{Q(s,a)}_{\text{prediction}}\right]}^{\text{TD error}}$$

Optimal Q's via behaviour policy → exploration/exploitation trade-off

Example: grid world

- agent (red) has to reach orange square (reward 0) without falling off the blue cliff (reward -100)
- all other state-actions yield reward -1

Final Q-value distribution with fixed ϵ-greedy:



	0	1	2	3	4	5	6	7	8	9	10	11
0	U: -6.76 D: -6.73 R: -6.75 L: -6.71	U: -6.70 D: -6.74 R: -6.60 L: -6.62	U: -6.42 D: -6.53 R: -6.34 L: -6.34	U: -6.14 D: -6.09 R: -6.06 L: -6.12	U: -5.82 D: -5.77 R: -5.74 L: -5.78	U: -5.51 D: -5.37 R: -5.36 L: -5.53	U: -5.12 D: -4.97 R: -4.94 L: -5.32	U: -4.58 D: -4.49 R: -4.49 L: -4.69	U: -4.01 D: -4.02 R: -3.94 L: -4.28	U: -3.57 D: -3.37 R: -3.36 L: -3.70	U: -2.65 D: -2.69 R: -2.65 L: -3.01	U: -2.26 D: -1.90 R: -2.07 L: -2.15
1	U: -6.81 D: -6.96 R: -6.89 L: -6.89	U: -6.80 D: -6.71 R: -6.70 L: -6.75	U: -6.51 D: -6.43 R: -6.45 L: -6.67	U: -6.17 D: -6.08 R: -6.09 L: -6.39	U: -5.76 D: -5.68 R: -5.68 L: -5.98	U: -5.55 D: -5.21 R: -5.21 L: -5.63	U: -4.73 D: -4.68 R: -4.68 L: -4.92	U: -4.37 D: -4.09 R: -4.09 L: -4.23	U: -3.92 D: -3.44 R: -3.44 L: -3.98	U: -3.80 D: -2.71 R: -2.71 L: -2.93	U: -2.84 D: -1.90 R: -1.90 L: -3.24	U: -1.72 D: -1.00 R: -1.43 L: -2.27
2	U: -7.06 D: -7.40 R: -6.86 L: -7.14	U: -6.96 D: -99.95 R: -6.51 L: -7.15	U: -6.71 D: -93.75 R: -6.13 L: -6.86	U: -6.37 D: -96.88 R: -5.70 L: -6.16	U: -6.09 D: -99.61 R: -5.22 L: -6.12	U: -5.60 D: -99.22 R: -4.69 L: -5.68	U: -5.07 D: -99.22 R: -4.10 L: -5.18	U: -4.60 D: -99.22 R: -3.44 L: -4.07	U: -4.07 D: -96.88 R: -2.71 L: -3.98	U: -3.34 D: -98.44 R: -1.90 L: -3.35	U: -2.64 D: -98.44 R: -1.00 L: -2.63	U: -1.72 D: 0.00 R: -1.00 L: -1.81
3	U: -7.18 D: -7.46 R: -99.22 L: -7.45	U: 0.00 D: 0.00 R: 0.00 L: 0.00	U: 0.00 D: 0.00 R: 0.00 L: 0.00	U: 0.00 D: 0.00 R: 0.00 L: 0.00	U: 0.00 D: 0.00 R: 0.00 L: 0.00	U: 0.00 D: 0.00 R: 0.00 L: 0.00	U: 0.00 D: 0.00 R: 0.00 L: 0.00	U: 0.00 D: 0.00 R: 0.00 L: 0.00	U: 0.00 D: 0.00 R: 0.00 L: 0.00	U: 0.00 D: 0.00 R: 0.00 L: 0.00	U: 0.00 D: 0.00 R: 0.00 L: 0.00	U: 0.00 D: 0.00 R: 0.00 L: 0.00

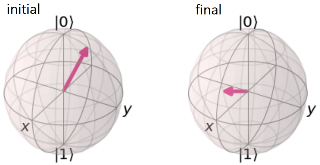
https://medium.com/@lgvaz/understanding-q-learning-the-cliff-walking-problem-80198921abbc, accessed: July 3, 2019

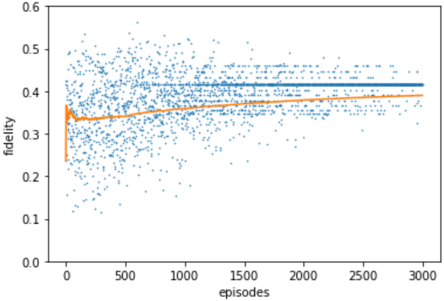
Q-value propagation

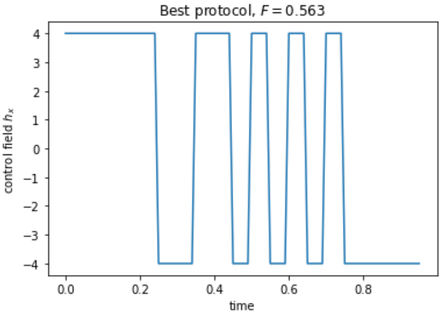
https://github.com/lgvaz/blog/blob/master/rl_intro.ipynb, modified, accessed: July 3, 2019

1-qubit control using a Q-table with ϵ-greedy

1-qubit: $L = 1 \to H = -S^z - h_x S^x$
Q-table with $\alpha = \epsilon = 0.9$, ϵ decay,
duration $T = 1 < T_{QSL}$ with
$\delta t = 0.05$







Q-value evolution over training

linear Q-function with tile coding

linear Q-function approximation:

$$Q(s,a) = \sum_{i=1}^{d} w_i x_i(s,a) \text{ with } w_i \text{ weights, } x_i \text{ features}$$

- allows generalization to unknown protocols
- gradient descent in weights: $w_i \leftarrow w_i + \alpha(r + \max Q - Q)\nabla_{w_i} Q$

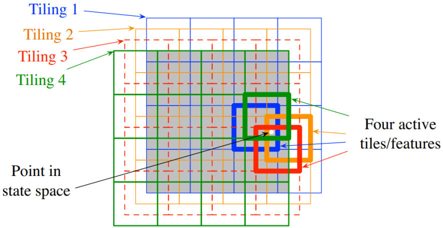
tile coding the features:

$$Q(s,a) = \sum_{i=1}^{n} w_i b_i(s,a)$$

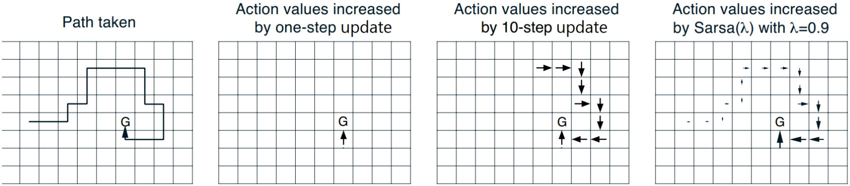
- discretize state-action space in n ways (tilings)
- binary function $b_i \in \{0,1\}$ selects tiles of current state-action (s,a)

RL tricks: generality and efficiency

tile coding: enables interpolation



eligibility traces: value updates in the "backward view"


Path taken
Action values increased by one-step update
Action values increased by 10-step update
Action values increased by Sarsa(λ) with λ=0.9
G
G
G
G

Sutton & Barto [4]

RL tricks: exploration and experience

2 alternating training phases:

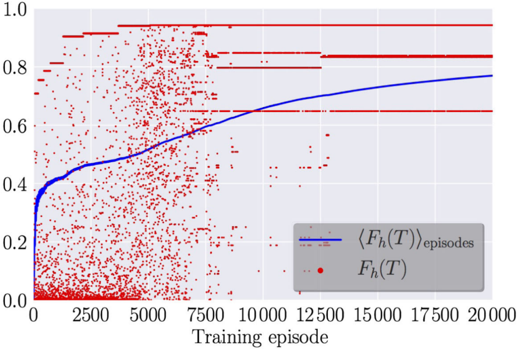
Exploratory

- actions sampled $P(a) \propto \exp(-\beta_{RL} Q(s,a))$
- ramp up of β_{RL}: uniform $\rightarrow$ greedy

Replay

Repeat best encountered protocols → bias agent for next exploration phase

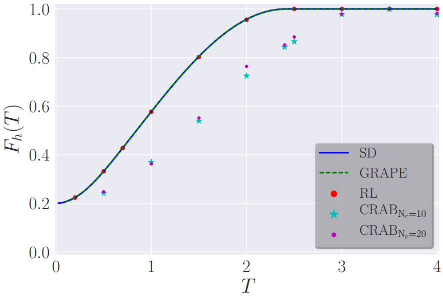
ϵ-**greedy** is used if not overridden by the above



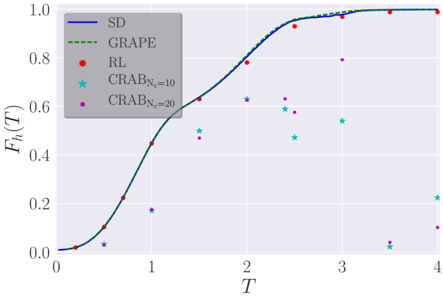
training for 10-qubits with $T = 3$

Comparision with optimal control algorithms

- Stochastic descent (SD), RL and GRAPE[2] find the optimal protocols
- performance drop-off of RL for large T → exponential state space scaling



$L = 1$



$L = 6$

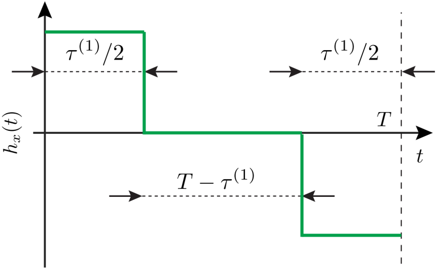
[2] Gradient Ascent Pulse Engineering

Learning from RL

RL protocol for 1 qubit at $T = 1$

An agent inspired protocol

- agent flips the magnetic field → wants h_x to be zero (but not possible in the BB setup)
- idea: positive pulse to reach equator, free evolution, negative pulse to reach target state
- pulse length $\tau/2$ is symmetric due to initial and final state



Learning from RL

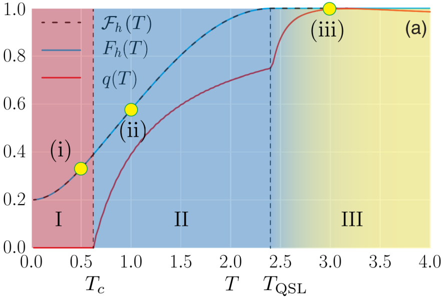
RL inspired protocol for 1 qubit at $T = 1$

Phase transitions in protocol space

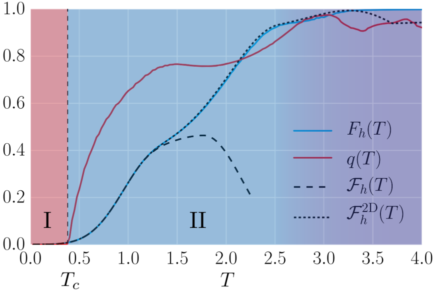
Control phase diagram = fidelity F of best protocol (SD) vs time T

- phase transition at critical time T_c and T_{QSL}

- phase transition at T_c
- "glassy" phase up to high T



1-qubit phase diagram



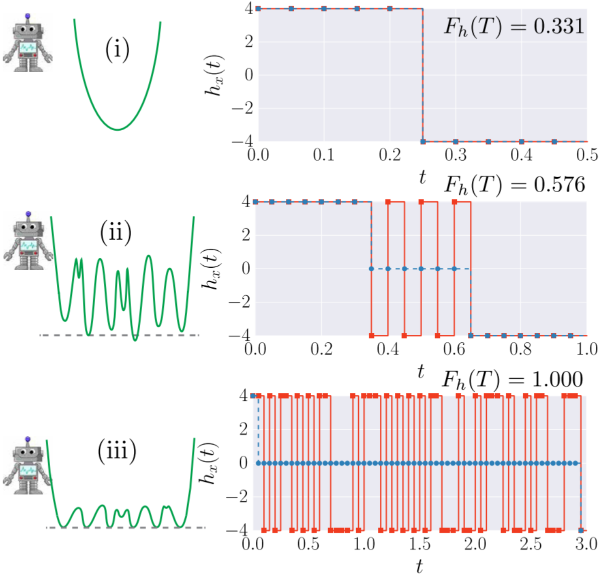
6-qubit phase diagram

Infidelity landscape

Infidelity $I_h = 1 - F_h$

- i Overconstrained phase: One global minimum
- ii Glassy phase: non-degenerate local minima → hard to find best protocol
- iii Controllable phase: degenerate minima with unit fidelity

→ best BB protocol ⇔ ground state of an Ising model



Conclusion & Outlook

Reinforcement learning ..

- .. is a feasible approach to quantum control
- .. offers comparable performance to model-based algorithms
- .. can inspire simple but powerful protocols
- .. may extend our ability to control to noisy and complex systems

Improvements:

- Reduce computational cost by use of matrix product states
- deep RL → deal with state space scaling
- adjust Q-learning to needs of quantum control
- pre-training/combination with model-based methods

References I

M. Bukov, A. Day, D. Sels, P. Weinberg, A. Polkovnikov, and P. Mehta.
Reinforcement learning in different phases of quantum control.
Physical Review X, 8, 09 2018.

S. Deffner and S. Campbell.
Quantum speed limits: from heisenberg's uncertainty principle to optimal quantum control.
Journal of Physics A: Mathematical and Theoretical, 50(45):453001, oct 2017.

References II

A. Omran, H. Levine, A. Keesling, G. Semeghini, T. T. Wang, S. Ebadi, H. Bernien, A. S. Zibrov, H. Pichler, S. Choi, J. Cui, M. Rossignolo, P. Rembold, S. Montangero, T. Calarco, M. Endres, M. Greiner, V. Vuletić, and M. D. Lukin.
Generation and manipulation of Schrödinger cat states in Rydberg atom arrays.

R. S. Sutton, A. Barto, and A. G. Barto.
Reinforcement Learning.
Adaptive computation and machine learning. MIT Press, Cambridge, Mass. [u.a.], 3. printing edition, 2000.

Thank you for your attention!
Questions? Ideas? Comments?