Split Q Learning: Reinforcement Learning with Two-Stream Rewards

BAIHAN LIN[1,2], DJALLEL BOUNEFFOUF[2], GUILLERMO CECCHI[2]

[1] Center for Theoretical Neuroscience, Columbia University, New York, NY 10027, USA
[2] IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598, USA

Abstract

Drawing an inspiration from behavioral studies of human decision making, we propose here a general parametric framework for a reinforcement learning problem, which extends the standard Q-learning approach to incorporate a two-stream framework of reward processing with biases biologically associated with several neurological and psychiatric conditions, including Parkinson's and Alzheimer's diseases, attention-deficit/hyperactivity disorder (ADHD), addiction, and chronic pain. For AI community, the development of agents that react differently to different types of rewards can enable us to understand a wide spectrum of multi-agent interactions in complex real-world socioeconomic systems. Empirically, the proposed model outperforms Q-Learning and Double Q-Learning in artificial scenarios with certain reward distributions and real-world human decision making gambling tasks. Moreover, from the behavioral modeling perspective, our parametric framework can be viewed as a first step towards a unifying computational model capturing reward processing abnormalities across multiple mental conditions and user preferences in long-term recommendation systems.

Human Q Learning

Algorithm 1 Human Q-Learning (HQL)

1: **For** each episode t **do**
2: Initialize s
3: **Repeat**
4: $Q(s,a) := \phi_2 Q^+(s,a) + \phi_4 Q^-(s,a)$
5: action $i_t = \arg\max_i Q_i(t)$, observe $s' \in S$, r^+ and $r^- \in R(s)$
6: $Q^+(s,a) := \phi_1 \hat{Q}^+(s,a) + \alpha_t(r^+ + \gamma \quad max_{a'} \hat{Q}^+(s',a') - \hat{Q}^+(s,a))$
7: $Q^-(s,a) := \phi_3 \hat{Q}^-(s,a) + \alpha_t(r^- + \gamma \quad max_{a'} \hat{Q}^-(s',a') - \hat{Q}^-(s,a))$
8: **until** s is terminal

Reward Processing Bias

Table 1: Algorithms Parameters

	ϕ_1	ϕ_2	ϕ_3	ϕ_4
"Addiction" (ADD)	1 ± 0.1	1 ± 0.1	0.5 ± 0.1	1 ± 0.1
"ADHD"	0.2 ± 0.1	1 ± 0.1	0.2 ± 0.1	1 ± 0.1
"Alzheimer's" (AD)	0.1 ± 0.1	1 ± 0.1	0.1 ± 0.1	1 ± 0.1
"Chronic pain" (CP)	0.5 ± 0.1	0.5 ± 0.1	1 ± 0.1	1 ± 0.1
"bvFTD"	0.5 ± 0.1	100 ± 10	0.5 ± 0.1	1 ± 0.1
"Parkinson's" (PD)	0.5 ± 0.1	1 ± 0.1	0.5 ± 0.1	100 ± 10
"moderate" (M)	0.5 ± 0.1	1 ± 0.1	0.5 ± 0.1	1 ± 0.1
Standard HQL (SQL)	1	1	1	1
Positive HQL (PQL)	1	1	0	0
Negative HQL (NQL)	0	0	1	1

Clinical Inspirations

From the perspective of evolutionary psychiatry, various mental disorders, including depression, anxiety, ADHD, addiction and even schizophrenia can be considered as "extreme points" in a continuous spectrum of behaviors and traits developed for various purposes during evolution, and somewhat less extreme versions of those traits can be actually beneficial in specific environments. Thus, modeling decision-making biases and traits associated with various disorders may actually enrich the existing computational decision-making models, leading to potentially more flexible and better-performing algorithms.

Reward-Scaling in RL

To explore the computational advantage of our proposed two-stream parametric extension of Q Learning can learn better than the baseline Q Learning, we tested our agents in nine computer games: Pacman, Catcher, FlappyBird, Pixelcopter, Pong, PuckWorld, Snake, WaterWorld, and Monster Kong. In each game, we tested in both stationary and non-stationary environments by rescaling the size and frequency of the reward signals in two streams. Preliminary results suggest that HQL outperform classical Q Learning in the long term in certain conditions (for example, positive-only and normal reward environments in Pacman). Our results also suggests that HQL behaves differently in the transition of reward environments.

Markov Decision Process (MDP) with not-Gaussian rewards

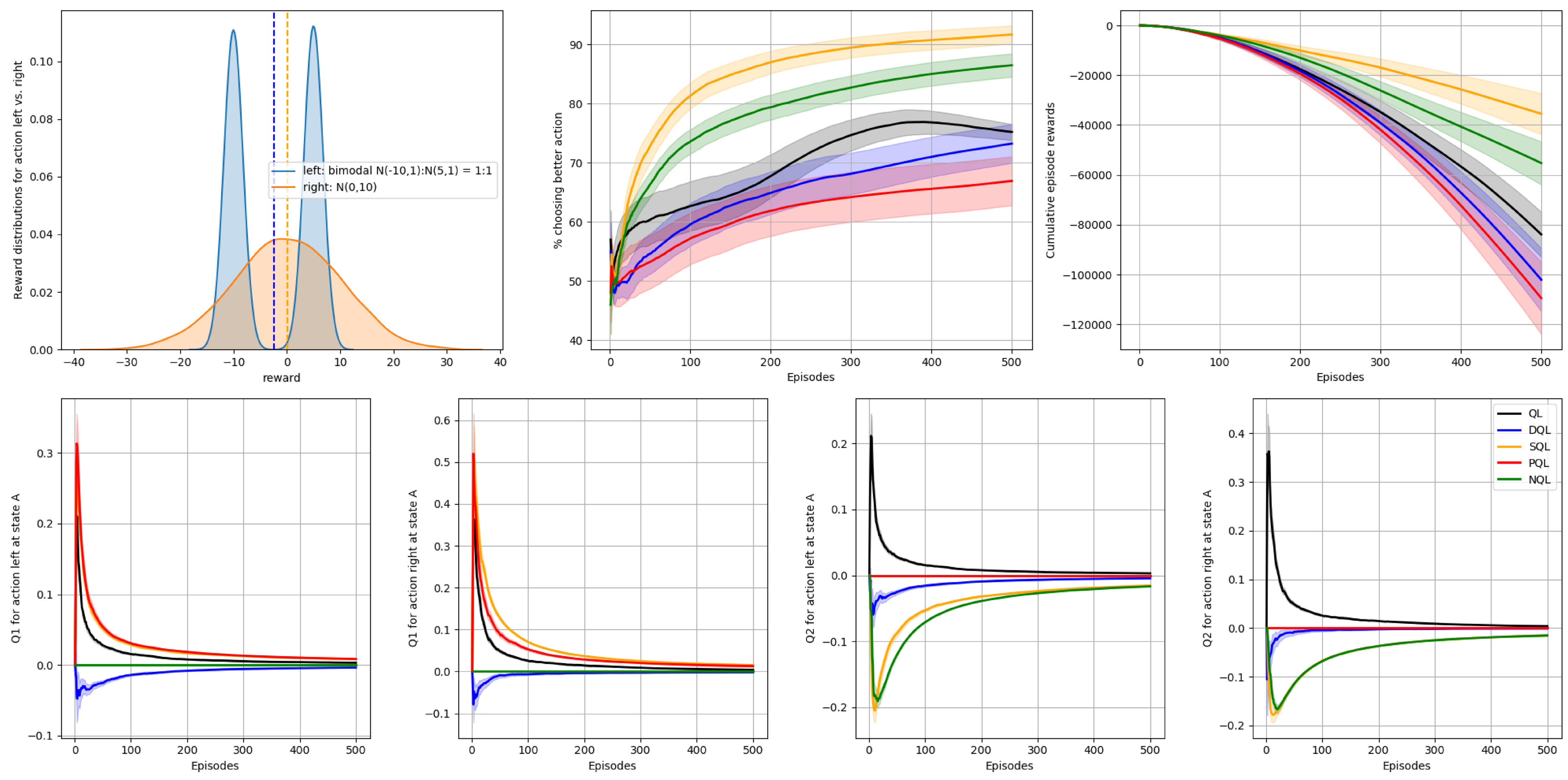

Figure 1: Example bi-modal MDP scenario where HQL performs better than QL and DQL.

	QL	DQL	SQL	PQL	NQL
QL	-	46 : **54**	34:**66**	72 : 28	44 : **56**
DQL	54:46	-	34:**66**	59:41	50:50
SQL	66:34	66:34	-	77:23	62:38
PQL	28:**72**	41:**59**	23:**77**	-	45:**55**
NQL	56:44	50:50	38:**62**	55:45	-
avg wins (%)	0.49	0.49	**0.68**	0.34	0.50

SQL		ADD	ADHD	AD	CP	bvFTD	PD	M	avg wins (%)
29:**71**	QL	60:40	65:35	73:27	43:**57**	75:25	38:**62**	49:**51**	0.58
22:**78**	DQL	54:46	80:20	81:19	61:39	77:23	52:48	53:47	0.65
-	SQL	78:22	94:6	95:5	67:33	89:11	66:34	81:19	**0.81**
-	avg wins (%)	0.36	0.20	0.17	0.40	0.16	0.48	0.39	-

Figure 2: MDP Task with 100 randomly generated scenarios of Bi-modal reward distributions.

Iowa Gambling Task (IGT) with reward-biased mental agents

Table 4: Iowa Gambling Task schemes

Decks	win per card	loss per card	expected value	scheme
A (bad)	+100	Frequent: -150 (p=0.1), -200 (p=0.1), -250 (p=0.1), -300 (p=0.1), -350 (p=0.1)	-25	1
B (bad)	+100	Infrequent: -1250 (p=0.1)	-25	1
C (good)	+50	Frequent: -25 (p=0.1), -75 (p=0.1),-50 (p=0.3)	+25	1
D (good)	+50	Infrequent: -250 (p=0.1)	+25	1
A (bad)	+100	Frequent: -150 (p=0.1), -200 (p=0.1), -250 (p=0.1), -300 (p=0.1), -350 (p=0.1)	-25	2
B (bad)	+100	Infrequent: -1250 (p=0.1)	-25	2
C (good)	+50	Infrequent: -50 (p=0.5)	+25	2
D (good)	+50	Infrequent: -250 (p=0.1)	+25	2

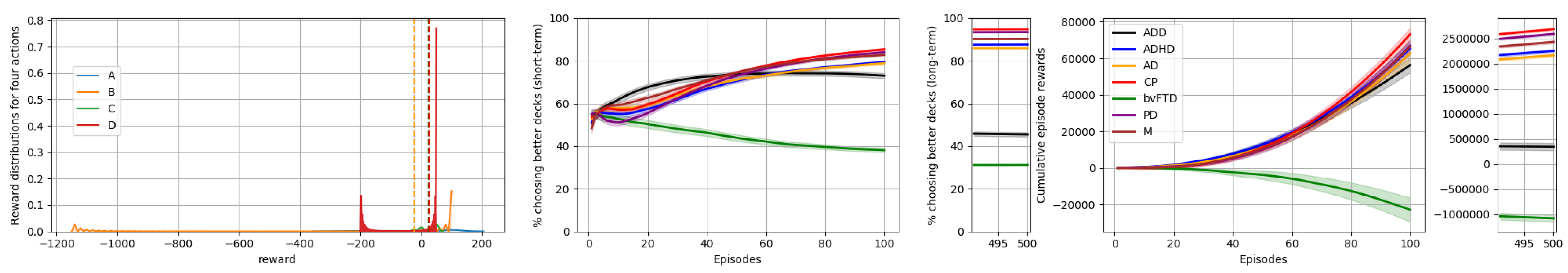

Figure 3: Short-term learning curves of different mental agents in IGT scheme 1.

Ongoing directions

- Investigate the optimal reward bias parameters in a series of computer games evaluated on different criteria, for example, longest survival time vs. highest final score.
- Explore the multi-agent interactions given different reward processing bias.
- Tune and extend the proposed model to better capture observations in literature.
- Learn the parameteric reward bias from actual patient data.
- Test the model on both healthy subjects and patients with specific mental conditions.
- Evaluate the merit in two-stream processing in deep Q networks.