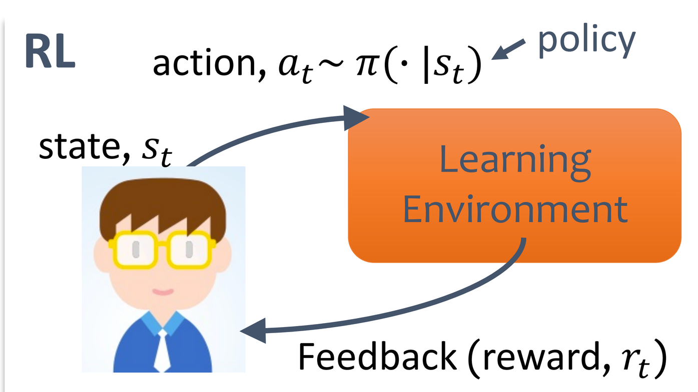
# Bayesian Q-learning with Assumed Density Filtering


Heejin Jeong and Daniel D. Lee

*Department of Electrical and System Engineering, University of Pennsylvania*


## INTRODUCTION

### Bayesian Reinforcement Learning (BRL)




Markov Decision Process : $\mathcal{M} = < S, A, P, R, \gamma >$

**Goal**: To maximize its expected total discounted future reward

**Value** : $V^\pi(s) = \mathbf{E}_\pi[\sum_{t=0}^{\infty} \underbrace{\gamma}_{\text{discount factor} \in [0,1]}{}^t r_t \,|s_0 = s]$

**Action-Value** : $Q^\pi(s,a) = \mathbf{E}_\pi[\sum_{t=0}^{\infty} \gamma^t r_t \,|s_0 = s, a_0 = a]$

**Optimality:** $V^*(s) = \max_a Q^*(s,a)$

$$Q^*(s,a) = \mathbf{E}_{s' \sim P(\cdot|s,a)}[\underbrace{r(s,a)}_{\text{Immediate}} + \underbrace{\gamma V^*(\overset{\text{subsequent state}}{s'})}_{\text{Future}}] = \mathbf{E}_{s' \sim P(\cdot|s,a)}\left[r(s,a) + \gamma \max_{a' \in A} Q^*(s',a')\right]$$

**BRL** leverages methods from Bayesian inference to incorporate information into the learning process.

**Off-policy Temporal Difference (TD) Learning :**

$\underbrace{\text{Off-policy}}_{\text{action policy} \neq \text{target policy}}$ $\underbrace{\text{Temporal Difference (TD)}}_{\text{long term future outcomes} \approx \text{temporally successive predictions}}$

**Kalman Temporal Difference** (Geist et al.), **KTD-Q**: a Bayesian approach to *off-policy TD learning* which approximates the value function using the *Kalman filtering scheme* - $Q^*(s,a) \approx Q^*(s,a;\theta)$, $\theta$: hidden states and $r$: indirect observation – and *Unscented Transform* for the nonlinearity of the max operator.

## Bayesian Q-learning with Assumed Density Filtering

### Q-learning

The most popular *off-policy TD learning* - After observing a reward $r_t$ and the next state $s_{t+1}$,

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha\left(\underbrace{\underbrace{r_t + \gamma \max_a Q(s_{t+1},a)}_{\text{TD target}} - Q(s_t,a_t)}_{\text{TD error}}\right)$$

### Belief Updates on Q-values

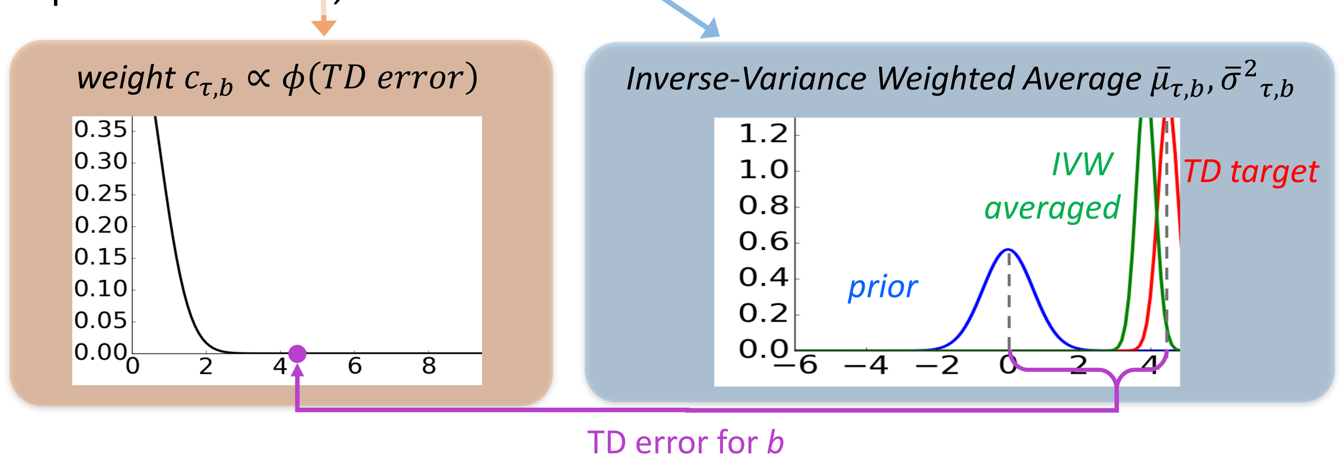
- $Q_{s,a} \sim \mathcal{N}(\mu_{s,a}, \sigma^2_{s,a})$ where $[\mu_{s,a}, \sigma_{s,a}] \neq [\mu_{s',a'}, \sigma_{s',a'}]$ if $s \neq s'$ or $a \neq a'$
- For *One-step Temporal Difference* (TD) Learning, we observe $r, s'$

$$\underbrace{\hat{p}_{Q_{s,a}}(q|\theta,r,s')}_{\text{Posterior}} \propto \underbrace{p(r+\gamma V_{s'}|q,\theta)}_{\text{Likelihood}}\underbrace{p_{Q_{s,a}}(q|\theta)}_{\text{Prior}} \propto \underbrace{p\left(V_{s'} = \frac{q-r}{\gamma}|q,s',\theta\right)}_{\text{Distribution over } V_{s'}} p_{Q_{s,a}}(q|\theta)$$

$$\propto \sum_{b \in A} c_{\tau,b}\,\phi(q;\bar{\mu}_{\tau,b},\bar{\sigma}^2{}_{\tau,b}) \prod_{b' \in A, b' \neq b} \Phi(q; r+\gamma\mu_{s',b'}, \gamma^2\sigma^2_{s',b'})$$

For each next possible action $b$,

$\phi, \Phi$ : Gaussian PDF and CDF
$\tau = < s,a,r,s' >$ : causality tuple




### Assumed Density Filtering Q-learning (ADFQ)

**Assumed Density Filtering (ADF) :** approximating a true posterior to a tractable parametric distribution in Bayesian networks by minimizing the reverse Kullback-Leibler divergence

$$\hat{p}_{Q_{s,a}}(q|\theta,r,s') \neq Gaussian \xrightarrow{\text{ADF}} \approx p_{Q_{s,a}}(q|\theta^{(new)}) = \mathcal{N}\left(q; \underline{\mathbf{E}_{q\sim\hat{p}_{Q_{s,a}}(\cdot)}[q]}\ , \underline{\mathrm{Var}_{q\sim\hat{p}_{Q_{s,a}}(\cdot)}[q]}\right)$$

- Simple analytic solutions for $|A|>2$ are not known/available.
- Algorithm with numerically computed solutions : ***ADFQ-Numeric***

### Approximated ADFQ ( *ADFQ-Approx* )

When $\sigma^2 \ll 1$, $\phi(\cdot) \approx \delta(\cdot)$ (dirac delta function) and $\Phi(\cdot) \approx H(\cdot)$ (Heaviside function).
Define a function $f(\cdot)$ - the approximation of the term inside the summation, $c_{\tau,b}\phi(\cdot)\prod\Phi(\cdot)$ :

$$f(q;\mu,\sigma) = \begin{cases} \frac{1}{\sigma}\phi\left(\frac{q-\mu}{\sigma}\right) & \text{for } q \in [\mu-\epsilon, \mu+\epsilon], \epsilon \ll 1 \\ 0 & \text{otherwise} \end{cases}$$

Then, $\hat{p}_{Q_{s,a}}(q|\theta,r,s') \approx \hat{p}_{Q_{s,a}}(q) = \frac{1}{Z}\sum_{b\in\mathcal{A}} c_{\tau,b} f(q;\bar{\mu}_{\tau,b},\bar{\sigma}_{\tau,b})$ for $q \in (-\infty,+\infty)$

Applying ADF, new mean and variance are:

$$\mathbf{E}_{q\sim\hat{p}_{Q_{s,a}}(\cdot)}[q] = \frac{\sum_b c_{\tau,b}\bar{\mu}_{\tau,b}}{\sum_b c_{\tau,b}} \qquad \mathrm{Var}_{q\sim\hat{p}_{Q_{s,a}}(\cdot)}[q] = \frac{\sum_b c_{\tau,b}\bar{\sigma}^2_{\tau,b}}{\sum_b c_{\tau,b}}$$

Just a **linear combination** of IVW mean/variance!

**Algorithm Complexity**

| Algorithm | Time per step | Space | Algorithm | Time per step | Space |
|---|---|---|---|---|---|
| Q-learning | $O(\|A\|)$ | $O(\|S\|\|A\|)$ | ADFQ-Numeric | $O(m\|A\|)$ | $O(\|S\|\|A\|)$ |
| KTD-Q | $O(\|S\|^2\|A\|^3)$ | $O(\|S\|^2\|A\|^2)$ | ADFQ-Approx | $\mathbf{O(\|A\|)}$ | $\mathbf{O(\|S\|\|A\|)}$ |

**Connection to Q-learning**

Suppose that $c_{\tau,b} = 0\ \forall b \neq \mathrm{argmax}_b\ \mu_{s,b}$, we can correspond the learning rate of Q-learning to the following:

$$\bar{\alpha} \equiv \frac{\bar{\sigma}^2_{b^*}}{\gamma^2\sigma^2_{s',b^*}} = \left(1 + \left(\frac{\gamma\sigma_{s',b^*}}{\sigma_{s,a}}\right)^2\right)^{-1}$$

## EXPERIMENTS

**Algorithms** ($\gamma = 0.9$)

- ADFQ with behavior policies - BS (Bayesian Sampling), semi-BS (performs *BS* with a small probability and greedily selects an action otherwise), $\epsilon$-greedy
- Q-learning with behavior policies - $\epsilon$-greedy and Boltzmann (softmax)
- KTD-Q with behavior policies - $\epsilon$-greedy and its active learning scheme.

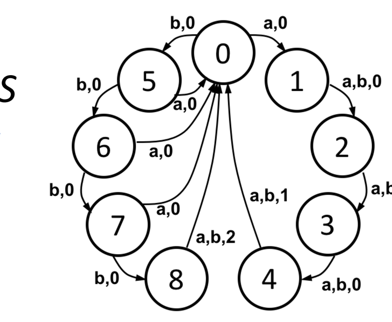



**Fig.1 Loop domain**

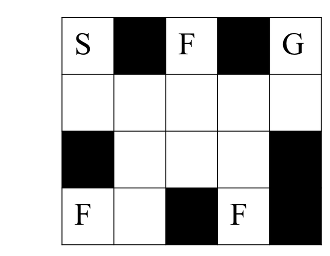


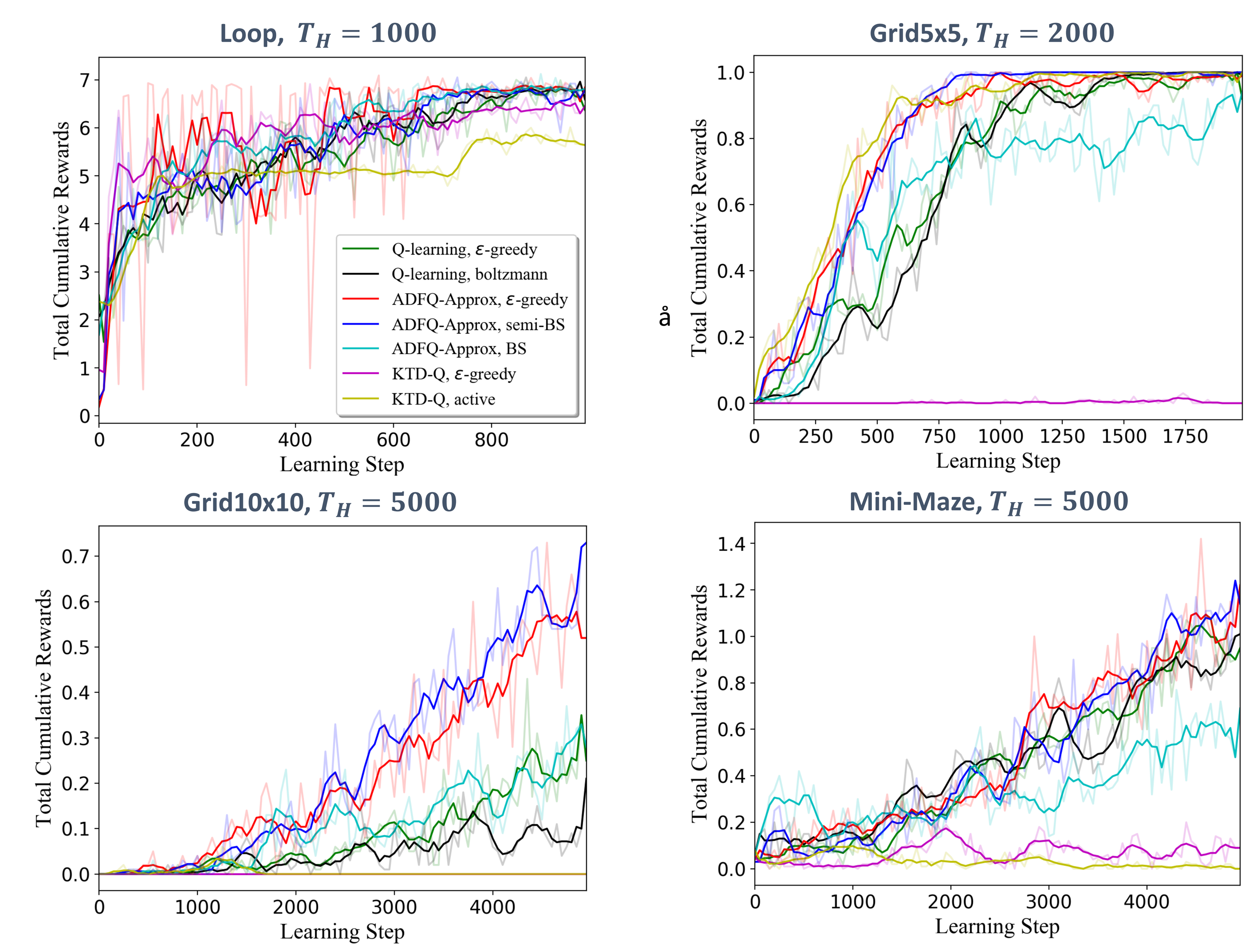

**Fig.2 Mini-Maze domain**

**Domains**

- Loop (Fig.1) : |S|=9, |A|=2, non-episodic, deterministic
- Mini-Maze (Fig.2) : |S|=112, |A|=4, $r$ = # of collected Flags at the Goal (F: Flag locations, S: starting point, G: goal), episodic, stochastic,
- Grid5x5 & Grid10x10 : |S|=25 or 100, |A|=4, $r = 1$ at the Goal (S: starting point, G: goal), episodic, stochastic,

**Semi-greedy Evaluation**

: Learning was paused at every $T_H/100$ step and the current policy was semi-greedily evaluated ($\epsilon$-greedy with $\epsilon = 0.1$). In the evaluation, the maximum # of steps is bounded by $T_H/50$, and for the episodic domains, it is also terminated when G is reached. The results were averaged over 10 trials.




**Total Cumulative Rewards**

: $\sum_{t=1,\cdots,T_I} r_t$, and averaged over 10 trials

| | Loop | Grid 5x5 | Grid 10x10 | Mini-Maze |
|---|---|---|---|---|
| Q-learning, $\epsilon$-greedy | $302.4 \pm 12.1$ | $150.6 \pm 3.8$ | $45.6 \pm 3.9$ | $239.7 \pm 81.4$ |
| Q-learning, Boltzmann | $288.2 \pm 17.4$ | $61.6 \pm 5.5$ | $18.0 \pm 1.9$ | $106.1 \pm 10.4$ |
| ADFQ-Approx, $\epsilon$-greedy | $\mathbf{338.0 \pm 0.0}$ | $178.1 \pm 5.5$ | $\mathbf{82.7 \pm 5.0}$ | $\mathbf{274.8 \pm 80.3}$ |
| ADFQ-Approx, semi-BS | $329.2 \pm 13.8$ | $\mathbf{184.7 \pm 4.5}$ | $80.9 \pm 7.1$ | $264.0 \pm 67.3$ |
| ADFQ-Approx, BS | $333.2 \pm 3.2$ | $135.9 \pm 5.7$ | $51.5 \pm 3.3$ | $180.9 \pm 47.8$ |
| KTD-Q, $\epsilon$-greedy | $281.6 \pm 5.2$ | $0.6 \pm 1.8$ | $0.0 \pm 0.0$ | $20.5 \pm 16.4$ |
| KTD-Q, active learning | $157.4 \pm 7.4$ | $18.8 \pm 2.7$ | $8.0 \pm 1.9$ | $55.4 \pm 8.6$ |

## DISCUSSION

**Contributions**

- **Regularization with Uncertainty Information in the Q update**: Unlike the Q-learning algorithm, the ADFQ algorithms incorporate the information of all possible actions for the next state with weights depending on TD errors and uncertainty measures - $\sigma_{s',b} \uparrow$ then contribution to the update $\downarrow$.
- **Connection to Q-learning** showed Q-learning could be a special case of our algorithm.
- **Computational Efficiency**.
- **No deterministic/stochastic environment assumption**: As the experiment results show, the ADFQ algorithms can work well on stochastic environments.
- **Only two hyperparameters** - initial variance and the discount factor: Other BRL algorithms tend to require many hyperparameters to be chosen.

**Limitations**

- Convergence analysis is not provided in this paper.
- Applied domains are limited to finite state and action spaces. We are currently extending our method to continuous domains

## FEATURED REFERENCE


X. Boyen and D. Koller. Tractable inference for complex stochastic processes. *In Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*, Berkeley, CA, 1998.

P. S. Maybeck. Stochastic models, estimation and control. *Academic Press*, chapter 12.7, 1982.

M. Opper. A bayesian approach to online learning. *On-Line Learning in Neural Networks*, 1999.

R. S. Sutton. Learning to predict by the methods of temporal differences. *Machine Learning*, 3(1):9–44, 1988.

Watkins, C. J., and Dayan, P. 1992. Q-learning. *In Machine Learning*, 279–292.

M. Geist and P. Olivier. Kalman temporal differences. *Journal of artificial intelligence research*, 39: 483–532, 2010.



**CONTACT :** Heejin Jeong ( heejinj@seas.upenn.edu), Daniel Lee ( ddlee@seas.upenn.edu )