Propagating Uncertainty in Reinforcement Learning via Wasserstein Barycenters

Alberto Maria Metelli*, Amarildo Likmeta*, and Marcello Restelli
{albertomaria.metelli, amarildo.likmeta, marcello.restelli}@polimi.it * Equal Contribution
POLITECNICO MILANO 1863

Problem and Motivation

- **Reinforcement Lerning** (RL, Sutton and Barto, 2018): find optimal policy π^* maximizing the *value function* v^π from each state s:

$$v_\pi(s) = \mathop{\mathbb{E}}_{\substack{A_t \sim \pi(\cdot|S_t) \\ S_{t+1}\sim \mathcal{P}(\cdot|S_t,A_t)}} \left[\sum_{t=0}^{+\infty} \gamma^t r(S_t, A_t) | S_0 = s \right]$$

- **Value-Based RL**
  1. estimate the optimal *action-value function* q^* for each state-action pair (s,a):

$$q^*(s,a) = r(s,a) + \gamma \mathop{\mathbb{E}}_{S' \sim \mathcal{P}(\cdot|s,a)} \left[\max_{a' \in \mathcal{A}} q^*(S', a') \right]$$

  2. The optimal policy π^* is any **greedy** policy w.r.t. q^*

$$\pi^*(s) \in \arg\max_{a\in\mathcal{A}} q^*(s,a)$$

- **Trade-Off** between

exploring new portions of the state-action space to reduce **uncertainty**

exploiting current (**uncertain**) information to decide the best action

Contributions

- We need a way of **quantifying the uncertainty** on the estimated optimal action-value function
  - $\Longrightarrow$ We propose to employ **posterior distributions** to model the uncertainty on the action-value function estimate
- We need a way to effectively **propagate the uncertainty** across the state-action space when updating the action-value function estimate
  - $\Longrightarrow$ We propose to use **Wasserstein barycenters** to combine the uncertainty of the state-action pairs

Wasserstein Barycenters

- **Wasserstein Metric**: distance between probability measures μ and ν (Villani, 2008)

$$W_2(\mu,\nu)^2 = \inf_{\rho \in \Gamma(\mu,\nu)} \mathop{\mathbb{E}}_{X,Y\sim\rho} \left[\|X - Y\|_2^2 \right]$$

  - $\Gamma(\mu,\nu)$ is the set of joint measures having μ and ν) as marginals
  - Cost in L^2-norm of "moving" probability mass to turn μ into ν

Wasserstein Barycenter: a way of "averaging" a set of probability measures $\{\nu_i\}_{i=1}^n$ based on Wasserstein metric (Agueh and Carlier, 2011)

$$\overline{\nu} \in \arg\inf_{\nu\in\mathcal{N}} \sum_{i=1}^n \xi_i W_2(\nu_i, \nu)^2$$

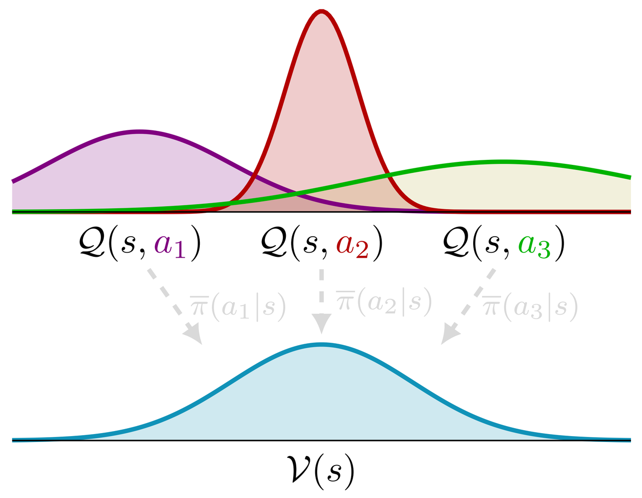
Modeling and Propagating Uncertainty

Modelling Uncertainty

- **Problem**: How to model the uncertainty on the action-value function estimate?
- **Idea**: maintain a probability distribution for each (s,a) (Dearden et al., 1998) $\Longrightarrow$ **Q-posterior** $\mathcal{Q}(s,a)$
- Employ a class of approximating probability distributions $\mathscr{Q}$
- Define the **V-posterior** $\mathcal{V}(s)$ as the Wasserstein barycenter of the Q-posteriors:

$$\mathcal{V}(s) \in \arg\inf_{\mathcal{V}\in\mathscr{Q}} \mathop{\mathbb{E}}_{A\sim\overline{\pi}(\cdot|s)} \left[W_2(\mathcal{V}, \mathcal{Q}(s,A))^2 \right]$$

  - In *prediction* problems $\overline{\pi}$ is the policy we want to evaluate
  - In *control* problems $\overline{\pi}$ aims at selecting the best action in state s
- It is the "Wasserstein version" of $v_{\overline{\pi}}(s) = \mathop{\mathbb{E}}_{A\sim\overline{\pi}} [q_{\overline{\pi}}(s,A)]$

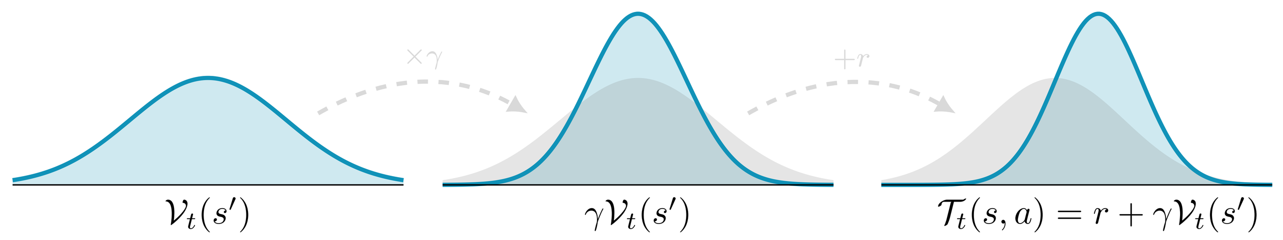


Propagating Uncertainty

- **Problem**: How to propagate uncertainty through a transition (s,a,s',r)?
  - Standard Bayesian updates assumes **independence** of samples!
- **Idea**: combine the Q-posterior $\mathcal{Q}_t(s,a)$ and the $\mathcal{V}_t(s')$ using Wasserstein barycenters

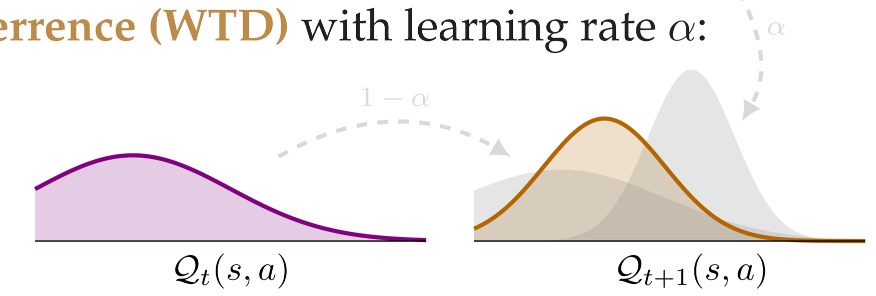
1. Compute the **Temporal Difference Target**

$$\mathcal{T}_t(s,a) = r + \gamma \mathcal{V}_t(s')$$



2. Combine the $\mathcal{T}_t(s,a)$ with $\mathcal{Q}_t(s,a)$ using the **Wasserstein Temporal Diference (WTD)** with learning rate α:

$$\mathcal{Q}_{t+1}(s,a) \in \arg\inf_{\mathcal{Q}\in\mathscr{Q}} (1-\alpha) W_2(\mathcal{Q}, \mathcal{Q}_t(s,a))^2 + \alpha W_2(\mathcal{Q}, \mathcal{T}_t(s,a))^2$$



  - $W_2(\mathcal{Q}, \mathcal{Q}_t(s,a))$ avoids moving too far from current estimate
  - $W_2(\mathcal{Q}, \mathcal{T}_t(s,a))$ allows propagating the V-posterior $\mathcal{V}_t(s')$, including its uncertainty
- It is the "Wasserstein version" of $q_{t+1}(s,a) = (1-\alpha)q_t(s,a) + \alpha(r + \gamma v_t(s'))$

Estimating the Maximum and Exploring

- **Problem**: How to select policy $\overline{\pi}$ in a control problem? How to define a proper *exploration policy*?
- **Idea**: exploit the Q-posteriors to define suitable policies $\overline{\pi}$

Mean Estimator (ME)
- Select the action(s) with the highest estimated mean

$$\arg\max_{a\in\mathcal{A}} \mathbb{E}[\mathcal{Q}(s,a)]$$

Optimistic Estimator/Exploration
- Select the action(s) that maximize an upper bound of the Q-posterior $u^\delta(s,a)$

$$\arg\max_{a\in\mathcal{A}} u^\delta(s,a)$$

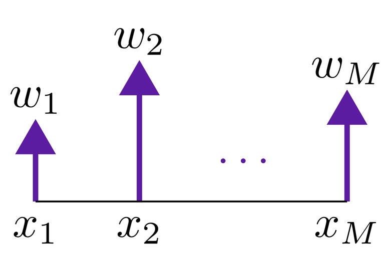
Posterior Estimator/Exploration
- Weight each action with the probability of being optimal

$$\Pr\left(a \in \arg\max_{a'\in\mathcal{A}} \mathcal{Q}(s,a) \right)$$

Particle Model

$$\mathcal{Q}(x;s,a) = \sum_{j=1}^M w_j \delta(x - x_j(s,a))$$

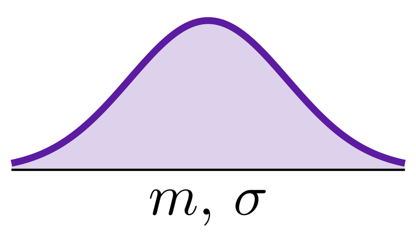
- Parameters: $\{x_j(s,a), w_j\}_{j=1}^M$
- Closed-form V-posterior, WTD
- Extension to function approximation $\Longrightarrow$ **PDQN** (Particle DQN)



Gaussian Model

$$\mathcal{Q}(x;s,a) = \frac{1}{\sqrt{2\pi\sigma^2(s,a)}} \exp\left\{ -\frac{1}{2}\left(\frac{x - m(s,a)}{\sigma(s,a)}\right)^2 \right\}$$

- Parameters: $m(s,a), \sigma(s,a)$
- Closed-form V-posterior, WTD
- We can prove **PAC-MDP** in the *average loss* setting, in the tabular case

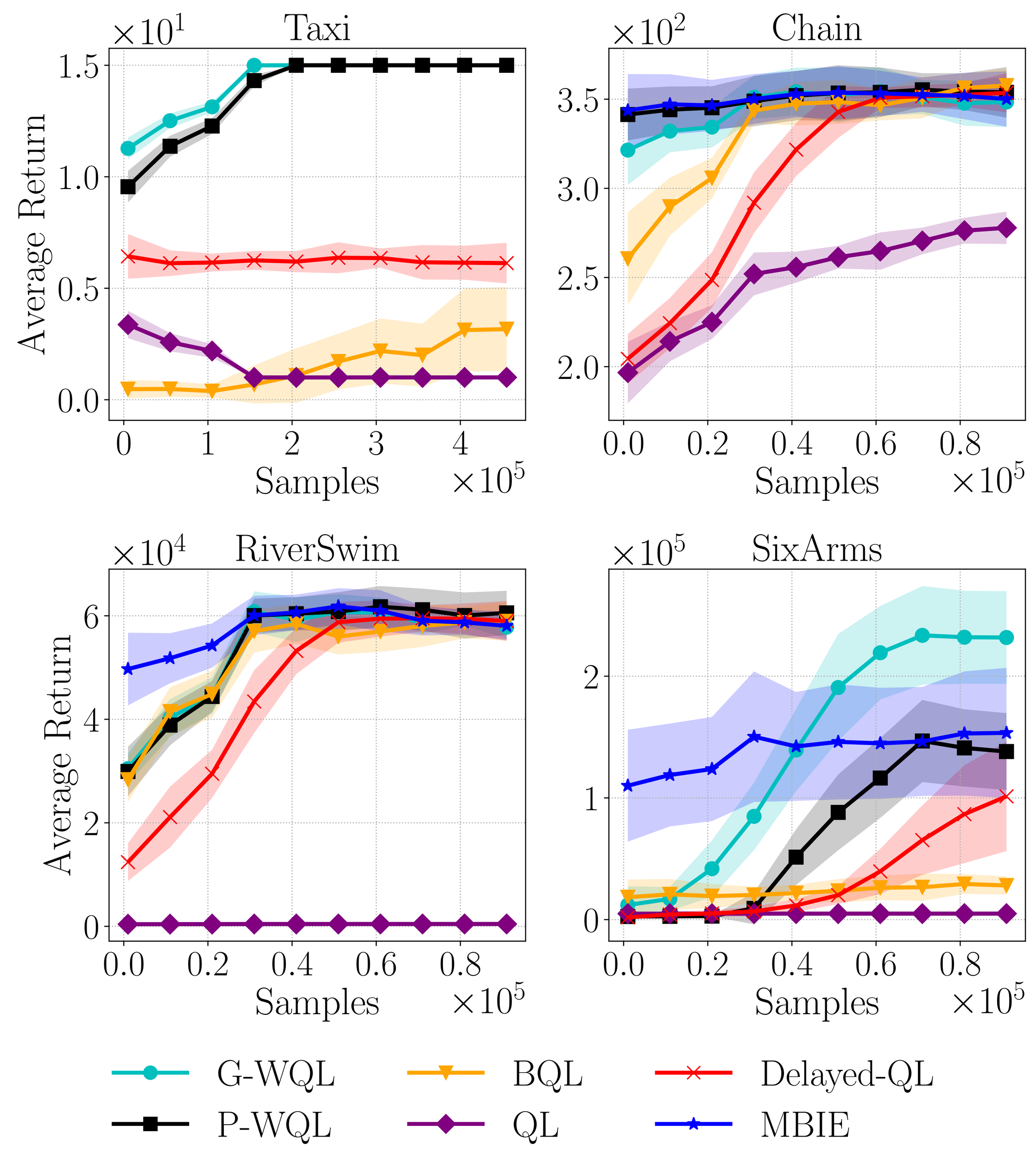


WQL

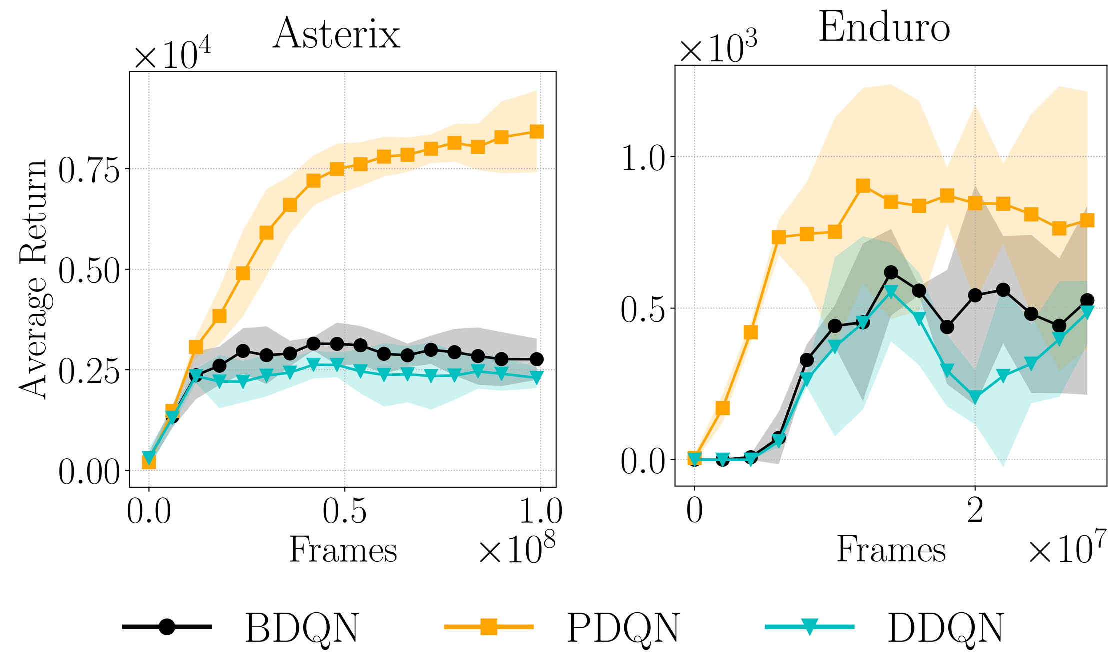
1: Initialize $\mathcal{Q}(s,a)$ with the prior $\mathcal{Q}_0$
2: **for** $t = 1, 2, \dots$ **do**
3: Take action $A_t \sim \overline{\pi}_t(\cdot|S_t)$
4: Observe S_{t+1} and R_{t+1}
5: Compute $\mathcal{V}_t(S_{t+1})$
6: Compute Update $\mathcal{Q}_{t+1}(S_t, A_t)$
7: **end for**

Experiments

Tabular Domains



Atari Games



References

M. Agueh and G. Carlier. Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.
R. Dearden, N. Friedman, and S. J. Russell. Bayesian q-learning. In J. Mostow and C. Rich, editors, *AAAI*, pages 761–768, 1998.
R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.
C. Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.