Reinforcement Learning with Multiple Experts: A Bayesian Model Combination Approach

Michael Gimelfarb, Scott Sanner, Chi-Guhn Lee

Department of Mechanical and Industrial Engineering, University of Toronto, ON, Canada

UNIVERSITY OF TORONTO

Neural Information Processing Systems Foundation

Key Questions

Model-free RL suffers from sparse rewards and sample inefficiency, so prior knowledge is often used to improve convergence. Here, prior knowledge is given as a fixed set of value function estimates $\Phi_1, \Phi_2, \dots \Phi_N$ of varying reliability. How can we:

1. design an on-line data-driven framework to learn which combination of these experts to trust?
2. make the framework compatible with most standard RL algorithms?
3. benefit from the expert advice while preserving the asymptotic behavior?

Related Work

Several methods have been introduced in order to learn a reward shaping function from data (Grzes and Kudenko, 2009), (Grzes and Kudenko, 2010), either for model-based RL or under specific assumptions.

There are also methods for performing action selection by sampling from multiple MDP models (Asmuth et al., 2009) in model-based RL.

Few work incorporates multiple sources of reward shaping advice in a Bayesian framework in model-free RL (Rosman et al., 2018). Here, prior state space knowledge is required, and it is not clear how to do posterior inference in constant time.

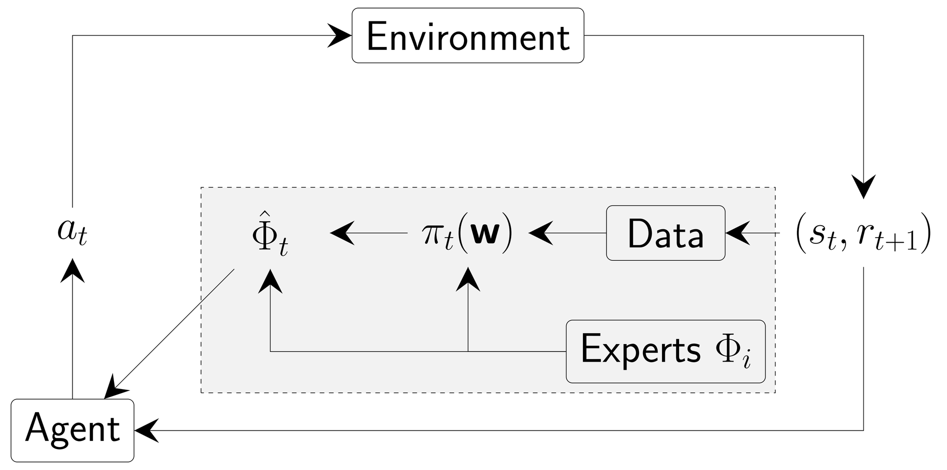
Main Ideas

Potential-based reward shaping (PBRS) transforms a sparse reward function R into a dense one $R + F$, where

$$\boxed{F(s,a,s') = \gamma\Phi(s') - \Phi(s).}$$

Φ is typically an estimate of the value function, called a **potential function**. Most importantly, PBRS does not change the optimal policies of MDPs (Ng et al., 1999).

We want to learn which combination of $\Phi_1, \Phi_2, \dots \Phi_N$ best represents the problem we are trying to solve and weight them accordingly.



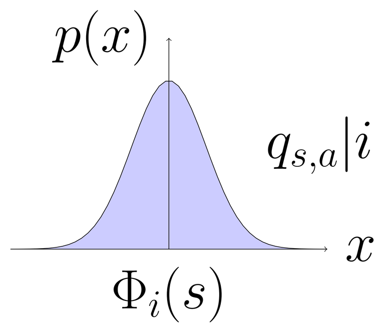
The Agent, Environment, and transitions (s_t, a_t, r_{t+1}) are standard in RL. The data available to the agent is used to form a posterior belief over the given expert models. This belief is used to combine the individual models into a single, and hopefully more informative, potential function for shaping.

Bayesian Model Combination

Q-values are interpreted as random variables $q_{s,a}$, data $\mathcal{D}$ of past returns is accumulated, and a belief $\pi_t : \mathcal{S}^{N-1} \to \mathbb{R}$ over the $(N-1)$-dimensional probability simplex is maintained. It is shown that

$$\boxed{\mathbb{E}[q_{s,a}|\mathcal{D}] = \int_{\mathbb{R}} q\mathbb{P}(q_{s,a}|\mathcal{D})\,\mathrm{d}q = \sum_{i=1}^{N} \mathbb{E}_{\pi_t}[w_i|\mathcal{D}]\ \mathbb{E}[q_{s,a}|i].}$$

Conditioned on Φ_i, Q-values are Gaussian with mean $\Phi_i(s)$ and variance σ^2 (Dearden et al., 1998). Here σ^2 is the sample variance of $\mathcal{D}$.



$$q_{s,a}|i \sim \mathcal{N}(\Phi_i(s), \sigma^2)$$

Starting with prior π_0, the posterior π_t is updated using **Bayes' theorem** as follows:

$$\pi_{t+1}(\mathbf{w}) = \mathbb{P}(\mathbf{w}|\mathcal{D}, q) \propto \mathbb{P}(q|\mathbf{w})\pi_t(\mathbf{w}) \propto \sum_{i=1}^{N} \mathbb{P}(q|i) w_i \pi_t(\mathbf{w}).$$

Unfortunately, exact inference is intractable so approximate inference is necessary.

Approximate Inference using Moment Matching

Starting with $\pi_t \sim \mathrm{Dir}(\alpha_t)$, we would like to approximate π_{t+1} by a Dirichlet with parameter α_{t+1} using **moment matching**.

Let $\mathbf{e} = [\mathbb{P}(q|i)]_{i=1\dots N}$ and let $\alpha_{t,0} = \alpha_{t,1} + \dots + \alpha_{t,N}$. The moments of π_{t+1} are shown to be

$$\boxed{\begin{aligned} m_i &= \mathbb{E}_{\pi_{t+1}}[w_i] = \int_{\mathcal{S}^{N-1}} \frac{\alpha_{t,0}}{\mathbf{e}\cdot\alpha_t} \sum_{j=1}^{N} e_j w_j w_i \pi_t(\mathbf{w})\,\mathrm{d}\mathbf{w} = \frac{\alpha_{t,i}(e_i + \mathbf{e}\cdot\alpha_t)}{(\mathbf{e}\cdot\alpha_t)(\alpha_{t,0}+1)} \\ s_1 &= \mathbb{E}_{\pi_{t+1}}[w_1^2] = \int_{\mathcal{S}^{N-1}} \frac{\alpha_{t,0}}{\mathbf{e}\cdot\alpha_t} \sum_{j=1}^{N} e_j w_j w_1^2 \pi_t(\mathbf{w})\,\mathrm{d}\mathbf{w} = \frac{\alpha_{t,i}(\alpha_{t,1}+1)(2e_1 + \mathbf{e}\cdot\alpha_t)}{(\mathbf{e}\cdot\alpha_t)(\alpha_{t,0}+1)(\alpha_{t,0}+2)}. \end{aligned}}$$

The parameters $\alpha = \alpha_{t+1}$ can now be found by matching the moments of $\mathrm{Dir}(\alpha)$ with those of π_{t+1}, m_i and s_1, and solving the resulting system

$$\left.\begin{aligned} m_i &= \frac{\alpha_i}{\alpha_0}, \quad i = 1, \dots N-1 \\ s_1 &= \frac{\alpha_i(\alpha_i+1)}{\alpha_0(\alpha_0+1)} \end{aligned}\right\} \quad \boxed{\begin{aligned} \alpha_{t+1,0} &= \frac{m_1 - s_1}{s_1 - m_1^2} \\ \alpha_{t+1,i} &= m_i \alpha_{t+1,0} = m_i \frac{m_1 - s_1}{s_1 - m_1^2}, \quad i = 1, \dots N-1. \end{aligned}}$$

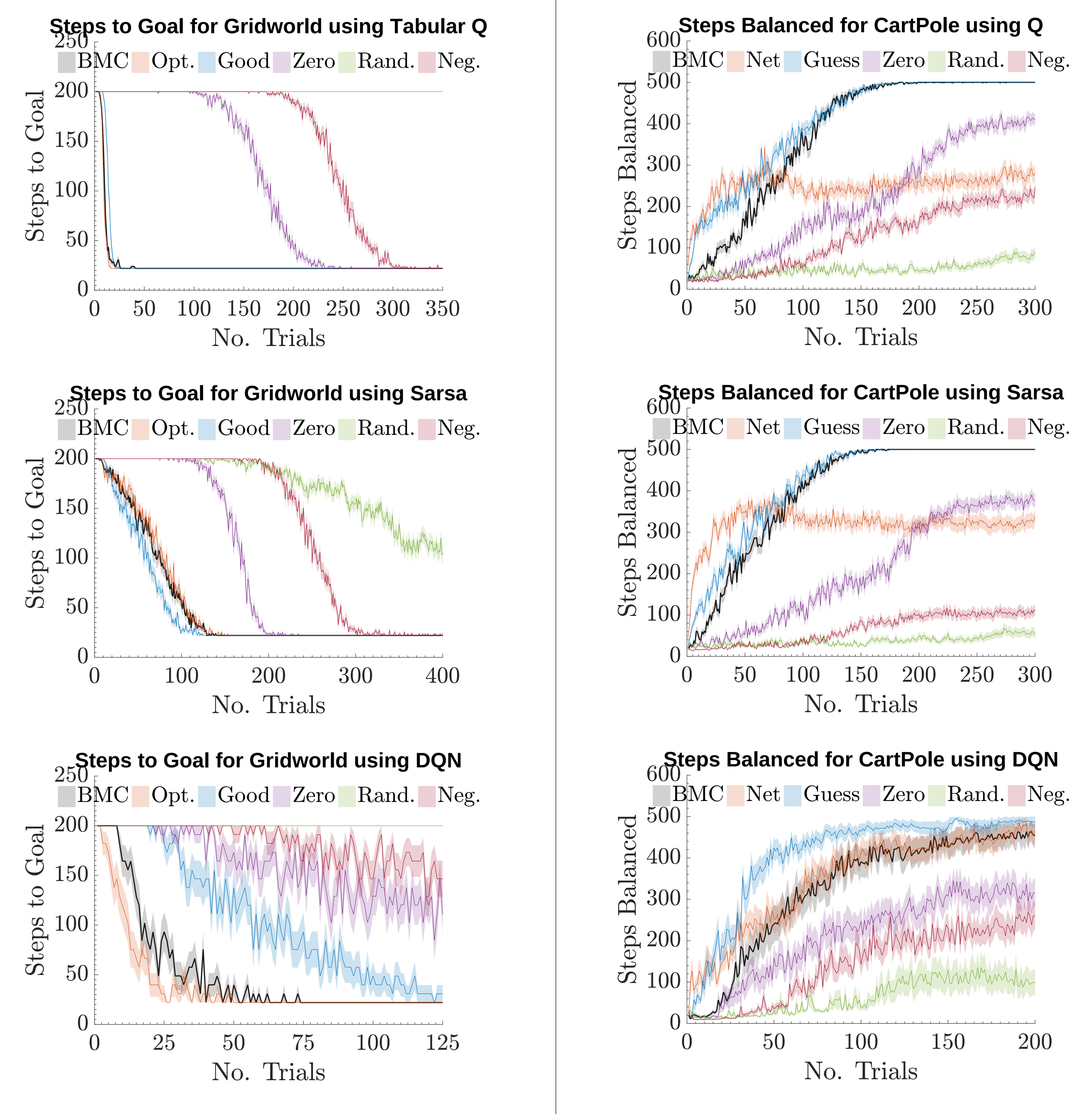
The expected return $\mathbb{E}[q_{s,a}|\mathcal{D}]$ is used as a proxy of the true optimal value function. The data-driven potential function is updated in $O(N)$-time

$$\boxed{\hat{\Phi}_t(s) = \mathbb{E}[q_{s,a}|\mathcal{D}] \approx \sum_{i=1}^{N} \frac{\alpha_{t,i}}{\sum_j \alpha_{t,j}} \Phi_i(s).}$$

Experimental Results

The # of steps required to reach the final goal on the discrete grid-world domain with five flags (Ng et al., 1999).

The # of steps the pole is balanced before falling on the classical continuous Cart-pole experiment.



The weights assigned to each expert for each of the three algorithms on the grid-world domain (top row) and cart-pole domain (bottom row):