Supplemental Material

Supplementary Figures

Figure 1 illustrates that the features proposed in the Main Text are sufficient to capture the VOC. Each data point corresponds to the VOC of deliberating for a different belief state. The cost of computation was $\lambda = 0.02$ and similar fits were obtained for other costs.

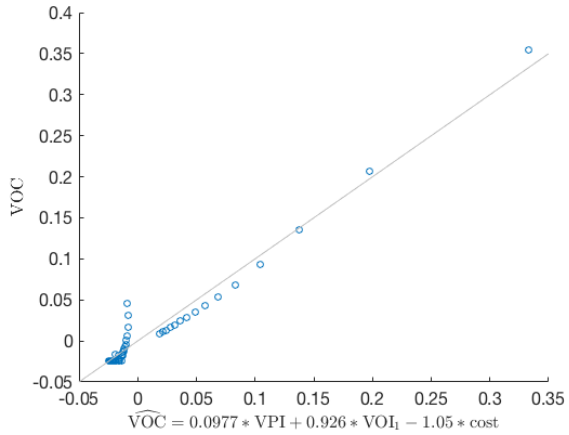

Figure 1: Linear fit of the value of computation for metareasoning about when to stop deliberating.

Figure 2 illustrates that our meta-level RL very quickly converged to near-optimal policies. This particular learning curve was observed for the 4-action meta-level probability model with a horizon of $h = 25$ and a cost of $\lambda = 0.01$.

Accelerating metalevel RL

The effectiveness of model-free RL critically depends on the reward structure. Delayed rewards are a particularly dire problem. This problem is particularly pronounced in meta-level MDPs, because the meta-level reward function is negative until the agent terminates deliberation. Previous research on model-free reinforcement learning has shown that the problem of delayed rewards can be ameliorated by reward shaping (Ng, Harada, and Russell 1999). Future work might therefore explore using reward-shaping to accelerate meta-level reinforcement learning.

One approach could be to use the method proposed in the main text to quickly learn a rough approximation to the meta-level value function and then use it to generate shaping rewards for a method that learns an approximation that is more accurate or can be evaluated more efficiently. This could be done as follows:

1. Apply the method presented in the main text to learn $\pi_{(meta)}(b)$.
2. Approximate $Q_{\pi_{\text{meta}}}$ by regressing the returns of $\pi_{(meta)}$ onto the features VOC_1, VPI_a, VPI_{all}, $\text{cost}(c)$, and $r_{\text{meta}}(b, \perp)$.
3. Apply the shaping theorem (Ng, Harada, and Russell 1999) to translate the resulting approximation $\hat{Q}_{\text{meta}}$ into a modified reward function

$$r'_{\text{meta}}(b,c) = \hat{Q}\text{meta}(b,c) - \max_c \hat{Q}_{\text{meta}}(b,c). \quad (1)$$

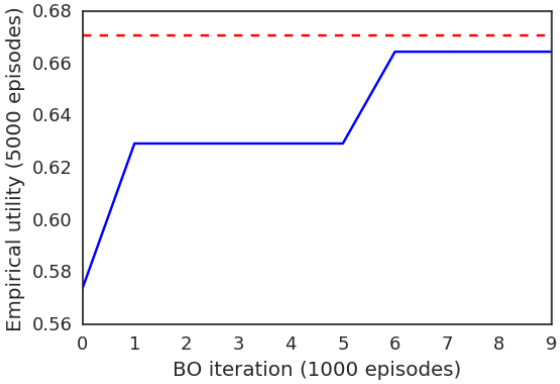

Figure 2: Learning curve of our metalevel RL method. The vertical axis shows the performance of the current best policy as a function of the number of iterations of the Bayesian optimization algorithm (horizontal axis) on an independent test set.

4. Train a different meta-level reinforcement learning algorithm on the modified reward function r'_{meta}.

Approximating the meta-level Q-function using regression

To generate shaping rewards, we can use the policy $\pi(b; \mathbf{w})$ learned with the method described in the main text to approximate the optimal meta-level action-value function $Q^{\star}_{\text{meta}}$. This can be achieved by running the approximate policy starting from randomly selected (b, c)-pairs, recording the resulting returns, and regressing them on features of the (b, c)-pair that the policy started from. Recall that $Q^{\star}_{\text{meta}}$ is the sum of the VOC and the expected reward of acting without deliberation (Equation 4 in the Main Text). Thus, assuming that we can approximate the VOC as a linear combination of the VOI_1, the VPI_{all}, the VPI_A, and $\text{cost}(c)$, it should be possible to approximate $Q^{\star}_{\text{meta}}$ as a linear combination of these three predictors and the expected reward of acting without deliberation ($r_{\text{meta}}(b, \perp)$). We will illustrate this approach by regressing the Monte-Carlo estimate of the returns of $\pi(b; \mathbf{w})$ onto these features. This leads to a simple approximation of $Q^{\star}_{\text{meta}}$ by

$$\begin{aligned}\hat{Q}_{\text{meta}}(b, c; \mathbf{w}) = w_1 \cdot \text{VOI}_1(b, c) + w_2 \cdot \text{VPI}_{\text{all}} \\ + w_2 \cdot \text{VPI}_A(b, c) - w_4 \cdot \text{cost}(c) \\ + w_5 \cdot \mathbb{E}\left[r_{\text{meta}}(b, \perp)|b\right],\end{aligned} \tag{2}$$

where the weights $\mathbf{w}$ are be estimated by linear regression.

Proof-of-Concept Simulations

As a proof-of-concept that our method could be leveraged to accelerate meta-level reinforcement learning, we show that the shaping rewards computed with this approach can accelerate tabular Q-learning on the first two metareasoning problems that we studied in the Main Text. This illustrates that our approximate method could, in principle, be used to accelerate reinforcement learning methods that converge to the optimal solution.

Learning when to terminate deliberation We first evaluate the approach on the learning when to terminate deliberation meta-MDP with a horizon of $h = 30$ and a computational cost of $\lambda = -0.015$. In this problem, the VPI_{all} and VPI_A are identical, so Equation 2 reduces to

$$\begin{aligned}\hat{Q}_{\text{meta}}(b, c; \mathbf{w}) = w_1 \cdot \text{VOI}_1(b, c) + w_2 \cdot \text{VPI} - w_3 \\ \cdot \text{cost}(c) + w_4 \cdot r_{\text{meta}}(b, \perp).\end{aligned} \tag{3}$$

First, we learned π_{meta} using the meta-level reinforcement learning algorithm described in the Main Text. Second, we approximated $Q_{\pi_{\text{meta}}}$ using linear regression. To do so, we ran the policy learned with meta-level reinforcement learning once for each $((\alpha, \beta), c)$-pair with $1 \leq \alpha, \beta \leq 30$ and $\alpha + \beta < 30$ and $c \in c_1, \perp$. We then used the resulting returns to estimate the feature-weights in in Equation 3 by linear regression with the maximum likelihood method (w_{ML}). We found that the resulting predictions $\hat{Q}(b, c; w_{\text{ML}})$ captured the optimal meta-level Q-function very accurately: The root-mean-squared-error between our approximation and the optimal meta-level Q-function was only 0.0103, the correlation between $\hat{Q}_{\text{meta}}$ and $Q^{\star}_{\text{meta}}$ was $r = 0.9995$, and our approximation explained 99.9% of the variance in $Q^{\star}$ across state-action pairs. We found that the same held true when we restricted our analysis to the meta-level Q-values of deliberating (RMSE = 0.0146, $r = 0.9996$, $R^2 = 0.999$) and terminating deliberation respectively (RMSE $< 10^{-15}$, $r = 1.0000$, $R^2 = 1.0000$).

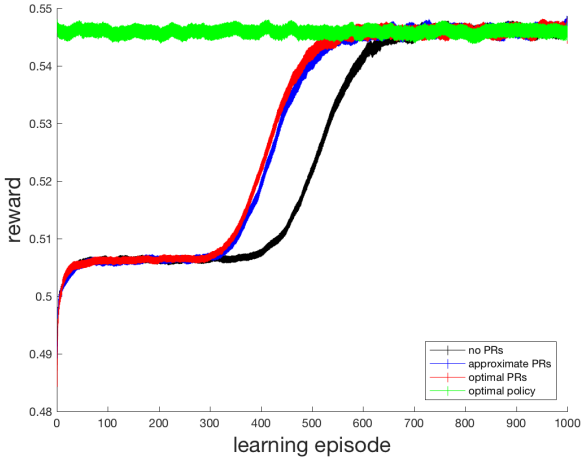

Figure 3: Shaping rewards accelerate learning when to stop deliberating. The amount of relative reward (vertical axis) increases over time at a greater rate when the agent receives shaping rewards.

Third, we translated $\hat{Q}_{\pi_{\text{meta}}}$ into shaping rewards according to Equation 1 and ran 2000 simulations to determine whether they can accelerate tabular Q-learning of the meta-level policy. The agent which was given either optimal shaping rewards based on $Q^{\star}_{\text{meta}}$, pseudorewards based on $\hat{Q}_{\pi_{\text{meta}}}$ (Equation 1), or no pseudorewards. The agent was trained for 1000 learning episodes with Q-values initialized to zero, a constant learning rate of 0.1, and a constant exploration rate of 0.25, and the cost of computation was set to 0.1. Figure 3 shows the performance of these three agents, with the approximate and optimal pseudoreward agents earning comparable rewards, and learning faster than the agent with no pseudorewards. All three agents eventually converge to the performance of an agent following the optimal policy, but for the agents that received pseudorewards this takes only 500 episodes instead of the 600 episodes required without pseudorewards.

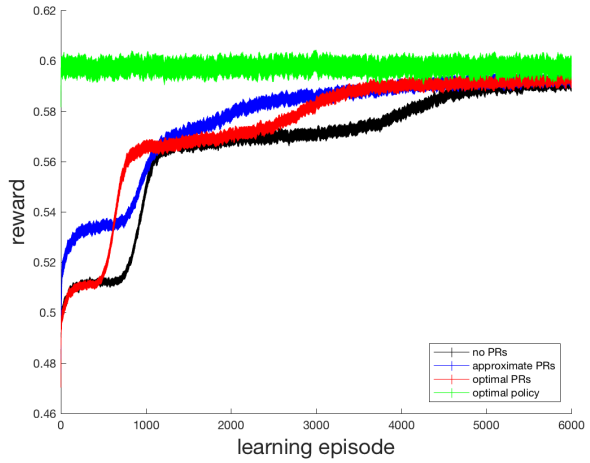

Figure 4: Shaping rewards accelerate learning how to choose between 3 options.

Metareasoning about decision-making Next, we evaluated this approach on the problem of metareasoning about decision-making described in the main text. As shown in Figure 4, the approximate pseudorewards generated with our method significantly accelerated learning compared to the control condition without pseudorewards. In the beginning the learning with approximate pseudorewards was even faster than with optimal pseudorewards, and over all both kinds of pseudoreward were about equally beneficial. This supports the hypothesis that our method can be used to accelerate meta-level RL. The cost of computation was 0.01. All other parameters were the same as in the simulations above.

References

[Ng, Harada, and Russell 1999] Ng, A. Y.; Harada, D.; and Russell, S. 1999. Policy invariance under reward transformations: Theory and application to reward shaping. In Bratko, I., and Dzeroski, S., eds., *Proceedings of the 16th Annual International Conference on Machine Learning*, 278–287. San Francisco: Morgan Kaufmann.