A The Option Transition Process

It will be convenient to consider the option transition process:

$$\begin{aligned}
P^{(0)}(x_f|x_s) &= \Pr(x_t = x_f|x_t = x_s) = \mathbb{I}_{x_s=x_f} \\
P^{(1)}(x_f|x_s) &= \Pr(x_{t+1} = x_f|x_t = x_s) \\
&= (1-\beta^o(x_s))p^{\pi^o}(x_f|x_s) \\
&\dots \\
P^{(k)}(x_f|x_s) &= \Pr(x_{t+k} = x_f|x_t = x_s) \\
&= \sum_x P^{(1)}(x|x_s)P^{(k-1)}(x_f|x)
\end{aligned}$$

We can then rewrite P^o from (3) as:

$$\begin{aligned}
P^o(x_f|x_s) &= \beta^o(x_f)\Big(P^{(0)}(x_f|x_s) + P^{(1)}(x_f|x_s) + \dots\Big) \\
&= \beta^o(x_f)\sum_{k=0}^{\infty} P^{(k)}(x_f|x_s) \qquad (6)
\end{aligned}$$

B Omitted Proofs

B.1 Proof of Theorem 1

Proof. We have:

$$\begin{aligned}
&\nabla_{\theta_\beta}P^o(x_f|x_s) \\
&= \nabla_{\theta_\beta}\beta^o(x_s)\mathbb{I}_{x_f=x_s} + \nabla_{\theta_\beta}(1-\beta^o(x_s))\sum_x p^{\pi^o}(x|x_s)P^o(x_f|x) \\
&= \nabla_{\theta_\beta}\beta^o(x_s)\mathbb{I}_{x_f=x_s} + \sum_x p^{\pi^o}(x|x_s)\Big(\nabla_{\theta_\beta}P^o(x_f|x) \\
&\qquad - \nabla_{\theta_\beta}\Big(\beta^o(x_s)P^o(x_f|x)\Big)\Big) \\
&= \nabla_{\theta_\beta}\beta^o(x_s)\mathbb{I}_{x_f=x_s} + \sum_x p^{\pi^o}(x|x_s)\Big(\nabla_{\theta_\beta}P^o(x_f|x) \\
&\qquad - \nabla_{\theta_\beta}\beta^o(x_s)P^o(x_f|x) - \beta^o(x_s)\nabla_{\theta_\beta}P^o(x_f|x)\Big) \\
&= \nabla_{\theta_\beta}\beta^o(x_s)\Big(\mathbb{I}_{x_f=x_s} - \sum_x p^{\pi^o}(x|x_s)P^o(x_f|x)\Big) \\
&\qquad + (1-\beta^o(x_s))\sum_x p^{\pi^o}(x|x_s)\nabla_{\theta_\beta}P^o(x_f|x). \qquad (7)
\end{aligned}$$

And so what we have is a $(1-\beta^o(x_i))$-discounted value function, whose reward is $\nabla_{\theta_\beta}\beta^o(x_i)r^o_{x_f}(x_i)$, where

$$r^o_{x_f}(x_i) = \mathbb{I}_{x_f=x_i} - \sum_{x_{i+1}} p^{\pi^o}(x_{i+1}|x_i)P^o(x_f|x_{i+1})$$

Now, from Eq. (3) and if $\beta^o(x) \neq 1$, we have:

$$\begin{aligned}
\sum_x p^{\pi^o}(x|x_s)P^o(x_f|x) &= \frac{P^o(x_f|x_s) - \beta^o(x_s)\mathbb{I}_{x_f=x_s}}{1-\beta^o(x_s)} \\
r^o_{x_f}(x_s) = \mathbb{I}_{x_f=x_s} &- \frac{P^o(x_f|x_s) - \beta^o(x_s)\mathbb{I}_{x_f=x_s}}{1-\beta^o(x_s)} \\
&= \frac{\mathbb{I}_{x_f=x_s} - P^o(x_f|x_s)}{1-\beta^o(x_s)} \qquad (8)
\end{aligned}$$

Using this notation, and recalling the transition process from Eq. (6), we can rewrite (7) as:

$$\begin{aligned}
&\nabla_{\theta_\beta}P^o(x_f|x_s) \\
&= \nabla_{\theta_\beta}\beta^o(x_s)r^o_{x_f}(x_s) + \sum_x P^{(1)}(x|x_s)\nabla_{\theta_\beta}P^o(x_f|x) \\
&= \sum_x\sum_{k=0}^{\infty} P^{(k)}(x|x_s)\nabla_{\theta_\beta}\beta^o(x)r^o_{x_f}(x) \\
&= \sum_x \frac{P^o(x|x_s)}{\beta^o(x)}\nabla_{\theta_\beta}\beta^o(x)r^o_{x_f}(x) \\
&= \sum_x P^o(x|x_s)\nabla_{\theta_\beta}\log\beta^o(x)r^o_{x_f}(x)
\end{aligned}$$

Where the third equality follows from (6) and requires for $\beta^o(x)$ to not be 0.

□

B.2 Proof of Proposition 1

Proof. Let $\Pr(x|o)$ denote the probability of a state x being terminal for an option o. By definition of entropy we have:

$$\begin{aligned}
H(X_f|o) &= -\sum_{x_f}\Pr(x_f|o)\log\Pr(x_f|o) \\
&= -\sum_{x_f}\sum_{x_s}\Pr(x_s|o)\Pr(x_f|x_s,o) \\
&\qquad \times \log\sum_{x_s}\Pr(x_s|o)\Pr(x_f|x_s,o) \\
&= -\sum_{x_f}\sum_{x_s} d^\mu(x_s|o)P^o(x_f|x_s) \\
&\qquad \times \log\underbrace{\sum_{y_s} d^\mu(y_s|o)P^o(x_f|y_s)}_{\text{marginal } P^o_\mu(x_f)} \\
&= -\sum_{x_s} d^\mu(x_s|o)\sum_{x_f}P^o(x_f|x_s)\log P^o_\mu(x_f)
\end{aligned}$$

□

B.3 Proof of Theorem 2

Proof.

$$\nabla_{\theta_\beta}J(P^o) = -\nabla_{\theta_\beta}\underbrace{\sum_{x_s} d^\mu(x_s|o)}_{\mathbb{E}_{x_s}}\sum_{x_f}P^o(x_f|x_s)\log P^o_\mu(x_f)$$

$$= -\mathbb{E}_{x_s}\Big[\sum_{x_f}\Big(\nabla_{\theta_\beta}P^o(x_f|x_s)\log P^o_\mu(x_f)$$

$$
\begin{aligned}
&\qquad + P^o(x_f|x_s)\frac{\nabla_{\theta_\beta} P^o_\mu(x_f)}{P^o_\mu(x_f)}\Big)\Big]\\
&= -\mathbb{E}_{x_s}\Big[\sum_{x_f}\Big(\sum_{x} P^o(x|x_s) r^o_{x_f}(x)\nabla_{\theta_\beta}\log\beta^o(x)\log P^o_\mu(x_f)\\
&\quad + \frac{P^o(x_f|x_s)}{P^o_\mu(x_f)}\underbrace{\sum_{y_s} d^\mu(y_s|o)\sum_{x} P^o(x|y_s)}_{\sum_x P^o_\mu(x)}\, r^o_{x_f}(x)\nabla_{\theta_\beta}\log\beta^o(x)\Big)\Big]\\
&= -\mathbb{E}_{x_s}\Big[\sum_{x} P^o(x|x_s)\nabla_{\theta_\beta}\log\beta^o(x)\sum_{x_f} r^o_{x_f}(x)\\
&\quad\times\Big(\log P^o_\mu(x_f) + \frac{P^o(x_f|x_s)}{P^o_\mu(x_f)}\frac{P^o_\mu(x)}{P^o(x|x_s)}\Big)\Big]\\
&= -\mathbb{E}_{x_s}\Big[\sum_{x} P^o(x|x_s)\frac{\nabla_{\theta_\beta}\beta^o(x)}{\beta^o(x)}\sum_{x_f}\frac{\mathbb{I}_{x_f=x} - P^o(x_f|x)}{1-\beta^o(x)}\\
&\quad\times\Big(\log P^o_\mu(x_f) + \frac{P^o(x_f|x_s)}{P^o_\mu(x_f)}\frac{P^o_\mu(x)}{P^o(x|x_s)}\Big)\Big]\\
&= -\underbrace{\sum_{x_s} d^\mu(x_s|o)}_{\text{sample}}\quad\underbrace{\sum_{x}\frac{P^o(x|x_s)}{\beta^o(x)}}_{\text{sample (continuation)}}\quad\frac{\nabla_{\theta_\beta}\beta^o(x)}{1-\beta^o(x)}\\
&\quad\times\Big[\Big(\log P^o_\mu(x) + 1\Big)\\
&\quad - \underbrace{\sum_{x_f} P^o(x_f|x)}_{\text{sample}}\Big(\log P^o_\mu(x_f) + \frac{P^o(x_f|x_s)P^o_\mu(x)}{P^o_\mu(x_f)P^o(x|x_s)}\Big)\Big]
\end{aligned}
$$

Sampling the highlighted expectations, and noting that if ℓ are the logits of β^o,

$$\nabla_{\theta_\beta}\ell_{\beta^o}(x) = \frac{\nabla_{\theta_\beta}\beta^o(x)}{\beta^o(x)(1-\beta^o(x))},$$

we have our result.

□

C Correlation with Planning Performance

The policies considered in these experiments consist of some set of four options combined with the set of primitive actions. Planning performance, for a single goal-directed task, is evaluated as the average policy value over all states at the end of each of ten iterations of value iteration. Consider Figure 7 which shows the value iteration performance curve for a single task, comparing policies of primitive actions, options, and their combination. The planning performance is the average of this curve for ten iterations, further averaged over all possible goal-directed tasks in Four Rooms. This measures how quickly value iteration, using this set of option policies and terminations, is able to plan.

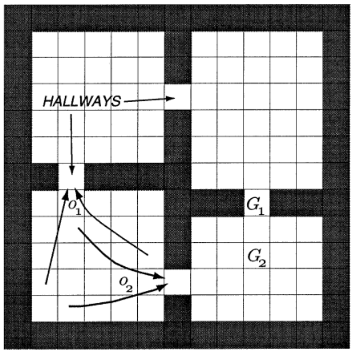

Figure 6: The Four Rooms domain map.

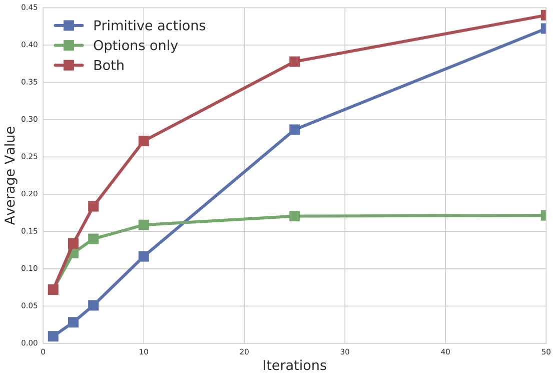

Figure 7: Example of planning performance using options, primitive actions, and both options and primitive actions.

D Learning Dynamics

Fig. 8 further studies the learning dynamics induced by the different components of the algorithm. We compare the previous two variants from Fig. 2 with only including the reachability advantage term (Row 3), and only including the trajectory advantage term (Row 4). The former does not focus on a single state, while the latter does not concentrate at all for many values of β.

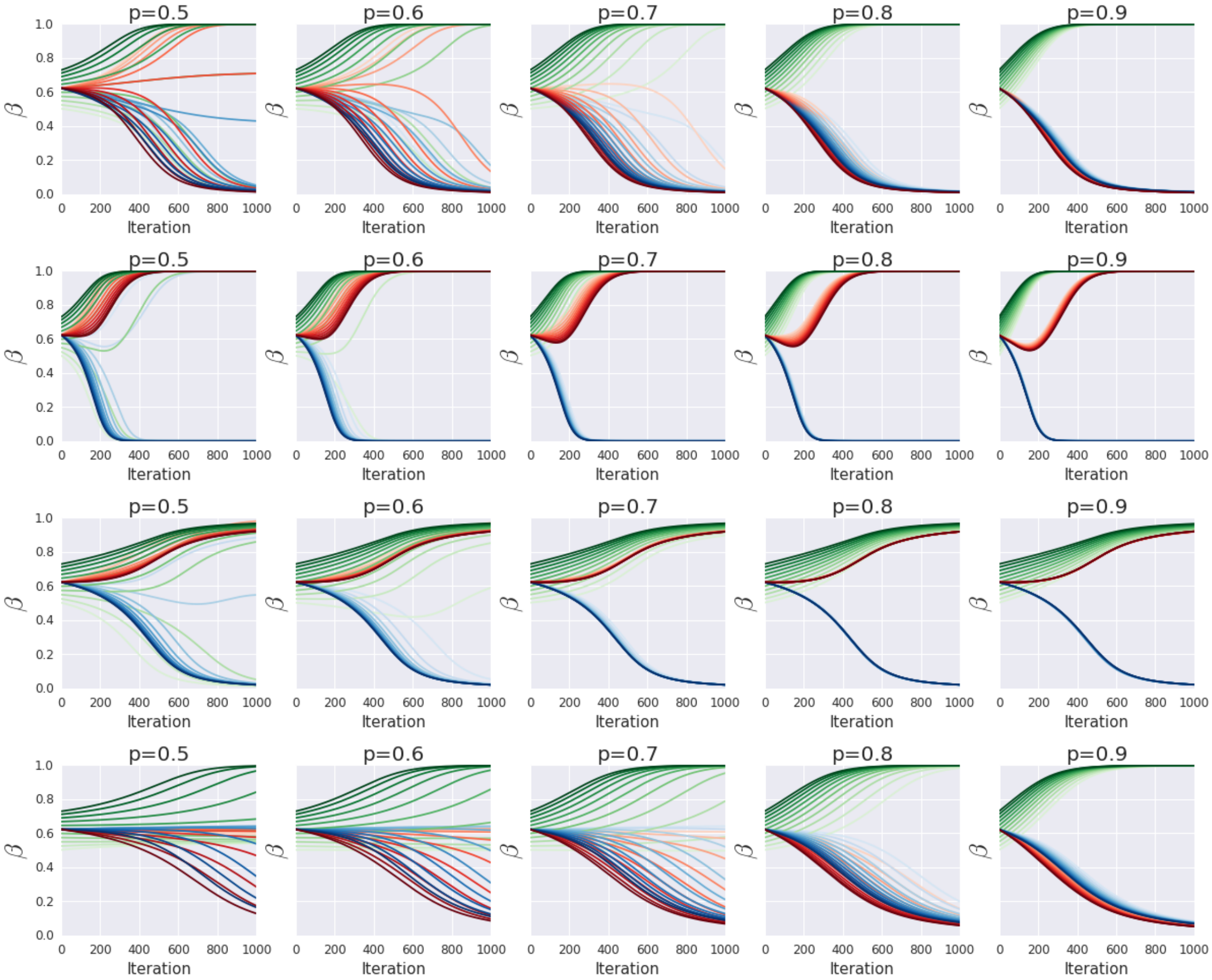

Figure 8: Learning dynamics. The color groups correspond with the states of the MDP from Fig. 1, while different lines correspond to different initial values of $\beta(green)$ (a lighter color depicts a lower value), **First row:** Termination-critic. **Second row:** Naive reachability. **Third row:** Only termination score advantage. **Fourth row:** Only relative termination advantage. We see that when the β-initialization is not too low, termination critic correctly concentrates termination on the attractor state and that state only, while the naive version saturates two of the states. The two ablations show the reachability advantage having similar behavior to naive reachability, while the trajectory advantage is not concentrating enough when the attraction values are low.