A. Proof of Theorem 1

Let us first introduce some relevant notation and definitions. Let ν be an arbitrary policy over options, and $c : \mathcal{S} \times \mathcal{O} \to [0,1]$ a coefficient function. It will be convenient to define the following:

$$p^{c\pi^o}(s'|s) \stackrel{\text{def}}{=} c(s',o)p^{\pi^o}(s'|s),$$
$$\mathcal{P}^{c\nu}q(s,o) \stackrel{\text{def}}{=} \sum_{s'} p^{c\pi^o}(s'|s) \sum_{o'} \nu(o'|s')q(s',o').$$

It will also be convenient to introduce the following operators for continuation and termination:

$$\mathcal{P}^{(1-\beta)\iota}q(s,o) \stackrel{\text{def}}{=} \sum_{s'} p^{\pi^o}_{ss'}(1-\beta^o(s'))q(s',o),$$
$$\mathcal{P}^{\beta\mu}q(s,o) \stackrel{\text{def}}{=} \sum_{s'} p^{\pi^o}_{ss'}\beta^o(s') \sum_{o'} \mu(o'|s')q(s',o').$$

That is: ι (*"iota"*) denotes the policy over options that maintains the current (argument) option. Let r^π denote the matrix of r^{π^o} for all options. Using these notations we can rewrite the Bellman operator over options for a discount γ:

$$\mathcal{T}^\mu_\gamma q = (I - \gamma\mathcal{P}^{(1-\beta)\iota})^{-1}(r^\pi + \gamma\mathcal{P}^{\beta\mu}q). \tag{5}$$

We are now ready to give our proof.

Proof. In order to preserve the appealing form of the option equations, it will be convenient to renormalize the *one-step* reward model r^{π^o} to be in terms of γ_p. We do this by setting $R^o_{\gamma_r}$ and $R^o_{\gamma_p}$ to be the same:

$$R^o_{\gamma_r} = (I - \gamma_r p^{(1-\beta)\pi^o})^{-1} r^{\pi^o}$$
$$= (I - \gamma_p p^{(1-\beta)\pi^o})^{-1} z^{\pi^o} = R^o_{\gamma_p},$$

and solving for z^{π^o}:

$$z^{\pi^o} = (I - \gamma_p p^{(1-\beta)\pi^o})(I - \gamma_r p^{(1-\beta)\pi^o})^{-1} r^{\pi^o}.$$

The Bellman operator can be rewritten similarly to (5): $\mathcal{T}^\mu_\Gamma q \stackrel{\text{def}}{=} R + \mathcal{P}^\mu_\Gamma q = (I - \gamma_p\mathcal{P}^{(1-\beta)\iota})^{-1}(z^\pi + \gamma_p\mathcal{P}^{\gamma_d\beta\mu}q)$, where we slightly abuse notation by writing $\mathcal{P}^{\Gamma\beta\mu}$ for $\sum_{s'} \gamma_d(s')\beta^o(s')p^{\pi^o}(s'|s)\sum_{o'}\nu(o'|s')q(s',o')$. Let us now derive the fixed point of $\mathcal{T}^\mu_\Gamma$:

$$\begin{aligned}
q &= (I - \gamma_p\mathcal{P}^{(1-\beta)\iota})^{-1}(z^\pi + \gamma_p\mathcal{P}^{\Gamma\beta\mu}q) \\
q - \gamma_p\mathcal{P}^{(1-\beta)\iota}q &= z^\pi + \gamma_p\mathcal{P}^{\Gamma\beta\mu}q \\
q &= z^\pi + \gamma_p\left(\mathcal{P}^{\Gamma\beta\mu}q + \mathcal{P}^{(1-\beta)\iota}q\right) \\
&= \left(I - \gamma_p(\mathcal{P}^{\Gamma\beta\mu} + \mathcal{P}^{(1-\beta)\iota})\right)^{-1} z^\pi \\
&= \sum_{t=0}^{\infty} \gamma^t \left(\mathcal{P}^{\Gamma\beta\mu} + \mathcal{P}^{(1-\beta)\iota}\right)^t z^\pi.
\end{aligned}$$

So $\mathcal{B} = \mathcal{P}^{\Gamma\beta\mu} + \mathcal{P}^{(1-\beta)\iota}$ is the corresponding one-step operator. Let us verify that it induces the new step-discount and termination scheme. Let $\ell^o(s') = \gamma_d(s')\beta^o(s') + 1 - \beta^o(s')$. We have:

$$\begin{aligned}
\mathcal{B}q(s,o) &= (\mathcal{P}^{\Gamma\beta\mu} + \mathcal{P}^{(1-\beta)\iota})q(s,o) \\
&= \sum_{s'} p^{\pi^o}(s'|s)\Big(\sum_{o'}\mu(o'|s')\gamma_d(s') \\
&\quad \times \beta^o(s')q(s',o') + (1-\beta^o(s'))q(s',o)\Big) \\
&= \sum_{s'} p^{\pi^o}(s'|s)\ell^o(s')\Big(\sum_{o'}\mu(o'|s')\frac{\gamma_d(s')\beta^o(s')}{\ell^o(s')} \\
&\quad \times q(s',o') + \frac{1-\beta^o(s')}{\ell^o(s')}q(s',o)\Big) \\
&= \sum_{s'} p^{\pi^o}(s'|s)\ell^o(s')\Big(\sum_{o'}\mu(o'|s') \\
&\quad \times \zeta^o(s')q(s',o') + (1-\zeta^o(s'))q(s',o)\Big).
\end{aligned}$$

Thus, we have a new termination scheme $\zeta^o(s) = \frac{\gamma_d(s)\beta^o(s)}{\ell^o(s)}$, and combining ℓ^o with the step-discount γ_p, we get our new step discount:

$$\begin{aligned}
\kappa(s,o,s') &= \gamma_p\ell^o(s') = \gamma_p\left(\gamma_d(s')\beta^o(s') + 1 - \beta^o(s')\right) \\
&= \gamma_p(1-\beta^o(s'))(1-\gamma_d(s')).
\end{aligned}$$

This implies that we have a state-option-dependent contraction factor.

$$\begin{aligned}
\eta(s,o) &= \mathbb{E}_{S_1,S_2,\ldots\sim p^{\pi^o}}\left[\prod_{i=1}^{\infty}\gamma(S_i,o,S_{i+1})\right] \\
&= \mathbb{E}_{S_1,S_2,\ldots\sim p^{\pi^o}}\left[\prod_{i=1}^{\infty}\gamma_p\left(1-\beta^o(S_i)(1-\gamma_d(S_i))\right)\right] \\
&\le \gamma_p.
\end{aligned}$$

The operator $\mathcal{T}^\mu_\Gamma$ is a contraction if $\eta(s,o) < 1$. This is trivially true if $\gamma_p < 1$. Otherwise, if $\gamma_p = 1$, in order for $\eta(s,o) < 1$, we need $1 - \beta^o(S_i)(1-\gamma_d(S_i)) < 1$ for some S_i along the trajectory, which is the same as $\beta^o(S_i)(1-\gamma_d(S_i)) > 0$ and holds if Assumption 5.2 holds: if such an S_i is reachable by π^o, is terminating in the sense that $\beta^o(S_i) > 0$, and $\gamma_d(S_i) < 1$. □

B. Proof of Lemma 1

Throughout we will refer to the minimum and maximum duration of options by $d^o_{\min}$ and $d^o_{\max}$, where $d^o_{\min}$ denotes the minimum, and $d^o_{\max}$ the maximum duration of o between any s and s' in $\mathcal{I}^o$ We will also write $d_{\min} \stackrel{\text{def}}{=} \min_{o\in\mathcal{O}} d^o_{\min}$, $d_{\max} \stackrel{\text{def}}{=} \max_{o\in\mathcal{O}} d^o_{\max}$ for minimum and maximum durations across options.

Before we proceed, let us show two helper bounds.

Lemma 3. *For each option $o \in \mathcal{O}$:*

$$|P^o_\Gamma(\cdot|s)|_1 \le \|\gamma_d\|\gamma_p^{d^o_{\min}}.$$

Proof. For each s and s', the transition model $P^o_\Gamma(s'|s) \le \gamma_d(s')\gamma_p^{d_{\min}}$, the minimum duration. Taking a max over the states s' yields the result. $\square$

Lemma 4. *The value function is bounded:*

$$\|q^\mu_\Gamma\| \le \frac{r_{\max}}{1-\gamma_r}\frac{1}{1-\|\gamma_d\|\gamma_p^{d_{\min}}}.$$

Proof. From the definition of q^μ_Γ, Lemma 3 and the definition of the reward model R:

$$\begin{aligned}\|q^\mu_\Gamma\|_\infty &= \|(I-\mathcal{P}^\mu_\Gamma)^{-1}R\|_\infty \\ &\le \frac{1}{1-\max_{s\in\mathcal{S},o\in\mathcal{O}}|P^o_\Gamma(\cdot|s)|_1}\|R\|_\infty \\ &\le \frac{1}{1-\|\gamma_d\|\gamma_p^{d_{\min}}}\frac{r_{\max}}{1-\gamma_r},\end{aligned}$$

since $\|\mathcal{P}^\mu_\Gamma q\|_\infty \le \max_{s\in\mathcal{S},o\in\mathcal{O}}|P^o_\Gamma(\cdot|s)|_1\|q\|$. $\square$

Recall the value functions w.r.t. the exact and approximate transition models and some policy μ: $q^\mu_\Gamma \overset{\text{def}}{=} (I-\mathcal{P}^\mu_\Gamma)^{-1}R$ and $q^\mu_{\widehat{\Gamma}} \overset{\text{def}}{=} (I-\widehat{\mathcal{P}^\mu_\Gamma})^{-1}R$. We will show Lemma 1 in two steps.

1) Bounding $\mathcal{E}_{estim}$ in terms of the one-step error. Similarly to Lemma 4 from (Jiang et al., 2015), we can relate the error in the value functions due to the approximate model to the maximum one-step error:

Lemma 5. *For any policy over options μ:*

$$\begin{aligned}\left\|q^\mu_{\widehat{\Gamma}} - q^\mu_\Gamma\right\| \le &\frac{1}{1-\|\gamma_d\|\gamma_p^{d_{\min}}} \\ &\times \max_{s,o}\left|R^o(s) + \widehat{\mathcal{P}^\mu_\Gamma}q^\mu_\Gamma(s,o) - q^\mu_\Gamma(s,o)\right|.\end{aligned}$$

Proof. Consider the evolution

$$q_m(s,o) = R^o(s) + \widehat{\mathcal{P}^\mu_{\mathcal{O}_\Gamma}}q_{m-1}(s,o). \tag{6}$$

We can bound the difference between successive estimates:

$$\begin{aligned}\|q_m - q_{m-1}\|_\infty &= \|\widehat{\mathcal{P}^\mu_{\mathcal{O}_\Gamma}}(q_{m-1}-q_{m-2})\|_\infty \\ &\le \max_{s\in\mathcal{S},o\in\mathcal{O}}\widehat{\mathcal{P}^o_\Gamma}(\cdot|s)|_1\|q_{m-1}-q_{m-2}\|_\infty \\ &= \|\gamma_d\|\gamma_p^{d_{\min}}\|q_{m-1}-q_{m-2}\|_\infty,\end{aligned}$$

due to Lemma 3 (which of course still applies to the approximate model.) Thus

$$\begin{aligned}\|q_m - q_0\| &\le \sum_{k=0}^{m-1}\|q_{k+1}-q_k\|_\infty \\ &\le \|q_1-q_0\|_\infty\sum_{k=1}^{m-1}(\|\gamma_d\|\gamma_p^{d_{\min}})^{k-1}.\end{aligned}$$

Since as $m\to\infty$, $q_m = q^\mu_{\widehat{\Gamma}}$, we have that $\|q^\mu_{\widehat{\Gamma}} - q_0\|_\infty \le \frac{1}{1-\|\gamma_d\|\gamma_p^{d_{\min}}}\|q_1-q_0\|_\infty$. Finally, since q_0 can be initialized to q^μ_Γ, and from Eq. (6) for $m=1$, we have our result. $\square$

2) Bounding the one-step error with the Hoeffding's bound. Now let us bound the one-step error in terms of the number of samples. The following Lemma is similar to Lemma 2 from (Jiang et al., 2015).

Lemma 6. *Let $\widehat{P^o_\Gamma}$ denote the modified transition model of an option o estimated i.i.d. from n samples, and $\widehat{\mathcal{P}^\mu_\Gamma}$ the corresponding operator w.r.t. some policy over options μ. Let $\mathcal{F}^o$ denote the set of possible terminating states of an option o. We have, with probability $1-\delta$:*

$$\begin{aligned}\|R + \widehat{\mathcal{P}^\mu_\Gamma}q^\mu_\Gamma - q^\mu_\Gamma\| \le &((\gamma_p^{d_{\min}} - \gamma_p^{d_{\max}})w + \gamma_p^{d_{\min}}\Delta v) \\ &\times\sqrt{\frac{1}{2n}\log\frac{2|S||\mathcal{O}|}{\delta}},\end{aligned}$$

where $\Delta v = \max_{o\in\mathcal{O}}\Big(\max_{s\in\mathcal{F}^o}v^\mu_\Gamma(s) - \min_{s\in\mathcal{F}^o}v^\mu_\Gamma(s)\Big)$ is the maximum variation of value in terminating states, and $w = \max_{o\in\mathcal{O}}\min_{s\in\mathcal{F}^o}\gamma_d(s)v^\mu_\Gamma(s)$.

Proof. Let us fix s and o, and consider a sample Y of $\sum_{s'}\widehat{P^o_\Gamma}(s'|s)v(s')$, where we write $v(s) = v^\mu_\Gamma(s) = \sum_o \mu(o|s)q^\mu_\Gamma(s,o)$ throughout. Because each $\widehat{P^o_\Gamma}(s'|s)$ is an average of zeros and samples of $\gamma_d(s')\gamma_p^D$, where D is the random variable corresponding to option duration, we have:

$$\gamma_p^{d_{\max}}\underbrace{\min_{s'\in\mathcal{F}^o}\gamma_d(s')v(s')}_{v^o_{\min}} \le Y \le \gamma_p^{d_{\min}}\underbrace{\max_{s'\in\mathcal{F}^o}\gamma_d(s')v(s')}_{v^o_{\max}}.$$

Now let $X_o = R^o(s) + \sum_{s'}\widehat{P^o_\Gamma}(s'|s)v(s')$. We have that:

$$R^o(s) + \gamma_p^{d_{\max}}v^o_{\min} \le X_o \le R^o(s) + \gamma_p^{d_{\min}}v^o_{\max}.$$

Thus the range $a = X_{\min} - X_{\max}$ of X is:

$$\begin{aligned}a &= \gamma_p^{d_{\min}}v^o_{\max} - \gamma_p^{d_{\max}}v^o_{\min} \\ &= (\gamma_p^{d_{\min}} - \gamma_p^{d_{\max}})v^o_{\min} + \gamma_p^{d_{\min}}\underbrace{(v^o_{\max} - v^o_{\min})}_{\Delta v^o}.\end{aligned}$$

Then, since $\widehat{P^o_\Gamma}(s'|s)$ is sampled i.i.d., and $q^\mu_\Gamma(s,o)$ is the average of $R^o(s) + \sum_{s'} \widehat{P^o_\Gamma}(s'|s) v^\mu_\Gamma(s')$, we have by the Hoeffding's bound:

$$\begin{aligned}
\Pr(|X_{s,o} - \mathbb{E}\left[X_{s,o}\right]| \geq t) &\leq 2\exp\left(-\frac{2nt^2}{a^2}\right) \\
&= 2\exp\left(-\frac{2nt^2}{((\gamma_p^{d_{\min}} - \gamma_p^{d_{\max}})v^o_{\min} + \gamma_p^{d_{\min}}\Delta v^o)^2}\right) \\
&= p_o.
\end{aligned}$$

Using the union bound we have:

$$\Pr(|X - \mathbb{E}\left[X\right]| \geq t) \leq |\mathcal{S}| \sum_o^{|\mathcal{O}|} p_o \leq |\mathcal{S}||\mathcal{O}| p_{\max},$$

where

$$\begin{aligned}
p_{\max} &= \max_o p_o \\
&= 2\exp\left(-\frac{2nt^2}{((\gamma_p^{d_{\min}} - \gamma_p^{d_{\max}})w + \gamma_p^{d_{\min}}\Delta v)^2}\right),
\end{aligned}$$

in which Δv is the maximum variation of terminal values over all options, and $w = \max_{o\in\mathcal{O}} \min_{s\in\mathcal{F}^o} \gamma_d(s)v(s)$. Solving for t we have our result:

$$\frac{\delta}{|S||\mathcal{O}|} = 2\exp\left(-\frac{2nt^2}{((\gamma_p^{d_{\min}} - \gamma_p^{d_{\max}})w + \gamma_p^{d_{\min}}\Delta v)^2}\right)$$

$$\log\frac{2|S||\mathcal{O}|}{\delta} = \frac{2nt^2}{((\gamma_p^{d_{\min}} - \gamma_p^{d_{\max}})w + \gamma_p^{d_{\min}}\Delta v)^2}$$

$$t = ((\gamma_p^{d_{\min}} - \gamma_p^{d_{\max}})w + \gamma_p^{d_{\min}}\Delta v)\sqrt{\frac{1}{2n}\log\frac{2|S||\mathcal{O}|}{\delta}}.$$

□

Lemma 1 then follows from combining Lemmas 5 and 6

C. Proof of Lemma 2

Proof. We have

$$\begin{aligned}
q^\mu_\Gamma - q^\mu_{\gamma_r} &= \mathcal{P}^\mu_\Gamma q^\mu_\Gamma - \mathcal{P}^\mu_{\gamma_r} q^\mu_{\gamma_r} \\
&= \mathcal{P}^\mu_\Gamma q^\mu_\Gamma - \mathcal{P}^\mu_\Gamma q^\mu_{\gamma_r} + \mathcal{P}^\mu_\Gamma q^\mu_{\gamma_r} - \mathcal{P}^\mu_{\gamma_r} q^\mu_{\gamma_r} \\
&= \mathcal{P}^\mu_\Gamma (q^\mu_\Gamma - q^\mu_{\gamma_r}) + (\mathcal{P}^\mu_\Gamma - \mathcal{P}^\mu_{\gamma_r}) q^\mu_{\gamma_r} \\
&= (I - \mathcal{P}^\mu_\Gamma)^{-1} \underbrace{(\mathcal{P}^\mu_\Gamma - \mathcal{P}^\mu_{\gamma_r})}_{\mathcal{A}} q^\mu_{\gamma_r}.
\end{aligned} \tag{7}$$

Let us now bound this expression. We will start with the inner term first. Noticing that $\mathcal{P}^{\Gamma\beta\mu} = \mathcal{P}^{\beta\mu}\gamma_d$, and from the definitions of the operators, we can expand:

$$\begin{aligned}
\|\mathcal{A}q^\mu_{\gamma_r}\| &= \|(\mathcal{P}^\mu_\Gamma - \mathcal{P}^\mu_{\gamma_r})q^\mu_{\gamma_r}\| \\
&= \Big\| \gamma_p \mathcal{P}^{\beta\mu}\gamma_d q^\mu_{\gamma_r} + \gamma_p \mathcal{P}^{(1-\beta)\iota}\mathcal{P}^\mu_\Gamma q^\mu_{\gamma_r} - \Big(\gamma_r \mathcal{P}^{\beta\mu} q^\mu_{\gamma_r} \\
&\quad + \gamma_r \mathcal{P}^{(1-\beta)\iota}\mathcal{P}^\mu_{\gamma_r} q^\mu_{\gamma_r}\Big)\Big\| \\
&\leq \Big\| \mathcal{P}^{\beta\mu}(\gamma_p\gamma_d - \gamma_r) q^\mu_{\gamma_r} + \mathcal{P}^{(1-\beta)\iota} \\
&\quad \left(\gamma_p \mathcal{P}^\mu_\Gamma q^\mu_{\gamma_r} - \gamma_r \mathcal{P}^\mu_{\gamma_r} q^\mu_{\gamma_r}\right) \Big\| \\
&= \Big\| \mathcal{P}^{\beta\mu}(\gamma_p\gamma_d - \gamma_r) q^\mu_{\gamma_r} + \mathcal{P}^{(1-\beta)\iota} \\
&\quad \left(\gamma_p \mathcal{P}^\mu_\Gamma q^\mu_{\gamma_r} - \gamma_r \mathcal{P}^\mu_\Gamma q^\mu_{\gamma_r} + \gamma_r \mathcal{P}^\mu_\Gamma q^\mu_{\gamma_r} - \gamma_r \mathcal{P}^\mu_{\gamma_r} q^\mu_{\gamma_r}\right) \Big\| \\
&= \Big\| \mathcal{P}^{\beta\mu}(\gamma_p\gamma_d - \gamma_r) q^\mu_{\gamma_r} + \mathcal{P}^{(1-\beta)\iota} \\
&\quad \left((\gamma_p - \gamma_r) \mathcal{P}^\mu_\Gamma q^\mu_{\gamma_r} + \gamma_r(\mathcal{P}^\mu_\Gamma q^\mu_{\gamma_r} - \mathcal{P}^\mu_{\gamma_r} q^\mu_{\gamma_r})\right) \Big\| \\
&\leq |\|\gamma_d\|\gamma_p - \gamma_r| \|q^\mu_{\gamma_r}\| + (\gamma_p - \gamma_r)\|\mathcal{P}^\mu_\Gamma q^\mu_{\gamma_r}\| \\
&\quad + \gamma_r \|\mathcal{P}^\mu_\Gamma q^\mu_{\gamma_r} - \mathcal{P}^\mu_{\gamma_r} q^\mu_{\gamma_r}\| \\
&\leq |\|\gamma_d\|\gamma_p - \gamma_r| \|q^\mu_{\gamma_r}\| + (\gamma_p - \gamma_r)\gamma_p^{d_{\min}}\|q^\mu_{\gamma_r}\| \\
&\quad + \gamma_r \|(\mathcal{P}^\mu_\Gamma - \mathcal{P}^\mu_{\gamma_r}) q^\mu_{\gamma_r}\|,
\end{aligned}$$

where the last inequality is due to Lemma 3. Let us simplify the first coefficient:

$$\begin{aligned}
|\|\gamma_d\|\gamma_p - \gamma_r| &= |\|\gamma_d\|\gamma_p - \gamma_p + \gamma_p - \gamma_r| \\
&\leq \gamma_p(1 - \|\gamma_d\|) + \gamma_p - \gamma_r.
\end{aligned}$$

Solving for $\|(\mathcal{P}^\mu_\Gamma - \mathcal{P}^\mu_{\gamma_r}) q^\mu_{\gamma_r}\|$, and by Lemma 4, we get:

$$\begin{aligned}
\|(\mathcal{P}^\mu_\Gamma - \mathcal{P}^\mu_{\gamma_r}) q^\mu_{\gamma_r}\| &\leq \frac{1}{1-\gamma_r}(\gamma_p(1 - \|\gamma_d\|) + \gamma_p - \gamma_r \\
&\quad + (\gamma_p - \gamma_r)\gamma_p^{d_{\min}})\|q^\mu_{\gamma_r}\| \\
&\leq \frac{r_{\max}}{(1-\gamma_r)^2}((\gamma_p - \gamma_r)(\gamma_p^{d_{\min}} + 1) + \\
&\quad \gamma_p(1 - \|\gamma_d\|)).
\end{aligned}$$

Finally, from Eq. (7) and using the bound on $\mathcal{P}^\mu_\Gamma$ from Lemma 3:

$$\mathcal{E}_{targ} \leq \frac{r_{\max}}{(1-\gamma_r)^2} \frac{(\gamma_p - \gamma_r)(\gamma_p^{d_{\min}} + 1) + \gamma_p(1 - \|\gamma_d\|)}{1 - \|\gamma_d\|\gamma_p^{d_{\min}}}.$$

□

D. Proof of Proposition 1

Proof. From Eq. (7):

$$\mathcal{E}_{estim} = (I - \mathcal{P}^\mu_\Gamma)^{-1}(\mathcal{P}^\mu_\Gamma - \mathcal{P}^\mu_{\gamma_r}) q^\mu_{\gamma_r}.$$

If the inner term is zero, the bias will be zero as well:

$$\begin{aligned}
(\mathcal{P}^\mu_\Gamma - \mathcal{P}^\mu_{\gamma_r}) q^\mu_{\gamma_r}(s,o) &= 0, \\
\mathcal{P}^\mu_\Gamma q^\mu_{\gamma_r}(s,o) &= \mathcal{P}^\mu_{\gamma_r} q^\mu_{\gamma_r}(s,o),
\end{aligned}$$

and hence:

$$\sum_{s'} P^{o}_{\gamma_p}(s'|s)\gamma_d(s,s')\sum_{o'}\mu(o'|s')q^{\mu}_{\gamma_r}(s',o') = \sum_{s'} P^{o}_{\gamma_r}(s'|s)\sum_{o'}\mu(o'|s')q^{\mu}_{\gamma_r}(s',o').$$

This equality can be achieved, if for each option o:

$$P^{o}_{\gamma_p}(s'|s)\gamma_d(s,s')\sum_{o'}\mu(o'|s')q^{\mu}_{\gamma_r}(s',o') = P^{o}_{\gamma_r}(s'|s)\sum_{o'}\mu(o'|s')q^{\mu}_{\gamma_r}(s',o'),$$

$$P^{o}_{\gamma_p}(s'|s)\gamma_d(s,s') = P^{o}_{\gamma_r}(s'|s),$$

$$\gamma_d(s,s') = \frac{P^{o}_{\gamma_r}(s'|s)}{P^{o}_{\gamma_p}(s'|s)}.$$

$\square$