A. Proof of Theorem 1

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

\[
p^\pi q(s'|s) \overset{\text{def}}{=} c(s', o) p^\pi(s'|s),
\]

\[
P^{\nu} q(s,o) \overset{\text{def}}{=} \sum_{s'} p^\pi q(s'|s) \sum_{o'} \nu(o'|s') q(s', o').
\]

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

\[
P^{(1-\beta)} q(s,o) \overset{\text{def}}{=} \sum_{s'} p^\pi q(s'|s) (1-\beta^o(s')) q(s', o),
\]

\[
P^{\beta} q(s,o) \overset{\text{def}}{=} \sum_{s'} p^\pi q(s'|s) \sum_{o'} \mu(o'|s') q(s', o').
\]

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

\[
T^n_\gamma q = (I - \gamma P^{(1-\beta)})(r^\pi + \gamma P^{\beta} q).
\]

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 \( r^{\pi^\nu} \) to be in terms of \( \gamma_p \). We do this by setting \( R_{\gamma r}^o \) and \( R_{\gamma p}^o \) to be the same:

\[
R_{\gamma r}^o = (I - \gamma r P^{(1-\beta)} p^{\pi})(1 - r^{\pi^\nu})
\]

\[
= (I - \gamma r P^{(1-\beta)} p^{\pi})(1 - r^{\pi^\nu}) = R_{\gamma p}^o,
\]

and solving for \( z^{\pi^\nu} \):

\[
z^{\pi^\nu} = (I - \gamma r P^{(1-\beta)} p^{\pi})(I - \gamma r P^{(1-\beta)} p^{\pi})^{-1}(1 - r^{\pi^\nu}) z^{\pi^\nu}.
\]

The Bellman operator can be rewritten similarly to (5):

\[
T^n_\gamma q \overset{\text{def}}{=} \frac{R + P^{\beta} q}{1 - P^{(1-\beta)} \gamma} = (I - \gamma P^{(1-\beta)})^{-1}(z^{\pi^\nu} + \gamma P^{\beta} q),
\]

where we slightly abuse notation by writing \( P^{\beta} q \) for \( \sum_{s'} \gamma d(s') \beta^o(s') p^{\pi}(s'|s) \sum_{o'} \nu(o'|s') q(s', o') \). Let us now derive the fixed point of \( T^n_\gamma \):

\[
q = (I - \gamma P^{(1-\beta)})(z^{\pi^\nu} + \gamma P^{\beta} q)
\]

\[
q - \gamma P^{(1-\beta)} q = z^{\pi^\nu} + \gamma P^{\beta} q,
\]

\[
z = z^{\pi^\nu} + \gamma P^{(1-\beta)} q.
\]

Thus, we have a new termination scheme \( \zeta^\nu(s) = \frac{\gamma d(s) z^{\pi^\nu}(s)}{\gamma o q^{(1-\beta)}(s)} \), and combining \( o q^{(1-\beta)} \) with the step-discount \( \gamma_p \), we get our new step discount:

\[
\kappa(s,o,s') = \gamma_p o q^{(1-\beta)}(s') = \gamma_p (\gamma_d(s') \beta^o(s') + 1 - \beta^o(s')) = \gamma_p (1 - \beta^o(s'))(1 - \gamma d(s')).
\]

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

\[
\eta(s,o) = E_{S_1,S_2,...\sim p^{\pi^\nu}} \left[ \prod_{i=1}^\infty \gamma(S_i, o, S_{i+1}) \right]
\]

\[
= E_{S_1,S_2,...\sim p^{\pi^\nu}} \left[ \prod_{i=1}^\infty \gamma_p (1 - \beta^o(S_i)(1 - \gamma d(S_i))) \right]
\]

\[
\leq \gamma_p.
\]

The operator \( T^n_\gamma \) 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 \( \pi^\nu \), is terminating in the sense that \( \beta^o(S_i) > 0 \), and \( \gamma d(S_i) < 1 \).

\[\square\]

B. Proof of Lemma 1

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

Before we proceed, let us show two helper bounds.
Lemma 3. For each option $o \in \mathcal{O}$:
$$|P_{T}^{\mu}(s) - |s|| \leq \|\gamma d\| \gamma_{p}^{d_{\min}}.$$ 

Proof. For each $s$ and $s'$, the transition model $P_{T}^{\mu}(s'|s) \leq \gamma d(s') \gamma_{p}^{d_{\min}}$, the minimum duration. Taking a max over the states $s'$ yields the result.

Lemma 4. The value function is bounded:
$$\|q^{\mu}_{T}\| \leq \frac{r_{\max}}{1 - \gamma \gamma_{r}} \frac{1}{\gamma_{r}} \|\gamma_{d}\| \gamma_{p}^{d_{\min}}.$$ 

Proof. From the definition of $q^{\mu}_{T}$, Lemma 3 and the definition of the reward model $R$:
$$\|q^{\mu}_{T}\| = \|\gamma_{d}|(I - P_{T}^{\mu})^{-1}R\| \leq \frac{1}{1 - \gamma_{d}} \|\gamma_{d}\| \gamma_{p}^{d_{\min}} \|q^{\mu}_{T}\|$$

since $\|P_{T}^{\mu}q\| \leq \|q\|$. We will show Lemma 1 in two steps.

1) Bounding $E_{\text{estim}}$ 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 $\mu$:
$$\|q^{\mu}_{T} - q^{\mu}_{T}\| \leq \frac{1}{1 - \|\gamma_{d}\| \gamma_{p}^{d_{\min}}} \times \max_{o} \left|R^{\mu}(s) + \gamma_{p}^{d_{\min}} q^{\mu}_{T}(s, o) - q^{\mu}_{T}(s, o)\right|.$$ 

Proof. Consider the evolution
$$q_{m}(s, o) = R^{\mu}(s) + \gamma_{p}^{d_{\min}} q_{m-1}(s, o). \quad (6)$$

We can bound the difference between successive estimates:
$$\|q_{m} - q_{m-1}\| \leq \|\gamma_{p}^{d_{\min}} q_{m-1} - q_{m-2}\| \leq \|\gamma_{d}\| \gamma_{p}^{d_{\min}} \|q_{m-1} - q_{m-2}\|$$

due to Lemma 3 (which of course still applies to the approximate model.) Thus
$$\|q_{m} - q_{0}\| \leq \sum_{k=0}^{m-1} \|q_{k+1} - q_{k}\| \leq \|q_{1} - q_{0}\| \prod_{k=1}^{m-1} (\|\gamma_{d}\| \gamma_{p}^{d_{\min}})^{k-1}.$$ 

Since as $m \to \infty$, $q_{m} = q^{\mu}_{T}$, we have that $\|q^{\mu}_{T} - q_{0}\| \leq 0.5 \log 2 \|\gamma_{d}\| \gamma_{p}^{d_{\min}} \|q_{1} - q_{0}\|$. Finally, since $q_{0}$ can be initialized to $q^{\mu}_{T}$, and from Eq. (6) for $m = 1$, we have our result.

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 $\hat{P}_{T}^{\mu}$ denote the modified transition model of an option $o$ estimated i.i.d. from $n$ samples, and $\hat{P}_{T}^{\mu}$ the corresponding operator w.r.t. some policy over options $\mu$. Let $\mathcal{F}_{o}$ denote the set of possible terminating states of an option $o$. We have, with probability $1 - \delta$:
$$\|R + \hat{P}_{T}^{\mu} q^{\mu}_{T} - q^{\mu}_{T}\| \leq ((\gamma_{p}^{d_{\min}} - \gamma_{p}^{d_{\max}}) w + \gamma_{p}^{d_{\min}} \Delta v) \times \sqrt{\frac{1}{2n} \log \frac{2|\mathcal{F}_{o}||\mathbb{S}_{o}|}{\delta}},$$

where $\Delta v = \max_{o \in \mathcal{O}} \left(\max_{s \in \mathcal{F}_{o}} v^{\mu}_{T}(s) \|\min_{s \in \mathcal{F}_{o}} v^{\mu}_{T}(s)\right)$ is the maximum variation of value in terminating states, and $w = \max_{o \in \mathcal{O}} \max_{s \in \mathcal{F}_{o}} \gamma_{d}(s) v^{\mu}_{T}(s)$.

Proof. Let us fix $s$ and $o$, and consider a sample $Y$ of $\sum_{s'} \hat{P}_{T}^{\mu}(s'|s) v(s')$, where we write $v(s) = v^{\mu}_{T}(s) = \sum_{o} \mu(o|s) q^{\mu}_{T}(s, o)$ throughout. Because each $\hat{P}_{T}^{\mu}(s'|s)$ is an average of zeros and samples of $\gamma_{d}(s') \gamma_{p}^{d_{\max}}$, where $D$ is the random variable corresponding to option duration, we have:
$$\gamma_{p}^{d_{\max}} \min_{s \in \mathcal{F}_{o}} \gamma_{d}(s) v(s') \leq Y \leq \gamma_{p}^{d_{\min}} \max_{s \in \mathcal{F}_{o}} \gamma_{d}(s) v(s').$$

Now let $X_{o} = R^{\mu}(s) + \sum_{s'} \hat{P}_{T}^{\mu}(s'|s) v(s')$. We have that:
$$R^{\mu}(s) + \gamma_{p}^{d_{\max}} v_{\min} \leq X_{o} \leq R^{\mu}(s) + \gamma_{p}^{d_{\min}} v_{\max}.$$ 

Thus the range $a = X_{\min} - X_{\max}$ of $X$ is:
$$a = \gamma_{p}^{d_{\max}} v_{\max} - \gamma_{p}^{d_{\min}} v_{\min} = (\gamma_{p}^{d_{\min}} - \gamma_{p}^{d_{\max}}) v_{\min} + \gamma_{p}^{d_{\min}} (v_{\max} - v_{\min}) \Delta v.$$
Then, since $\hat{P}^\alpha(s' | s)$ is sampled i.i.d. and $q^\mu_t(s,o)$ is the average of $R^\alpha(s) + \sum_a \hat{P}^\alpha(s' | s)v^\mu_t(s')$, we have by the Hoeffding’s bound:

$$\Pr(|X_{s,o} - \mathbb{E}[X_{s,o}]| \geq t) \leq 2 \exp \left(- \frac{2nt^2}{\Delta v^2} \right),$$

where $\Delta v$ is the maximum variation of terminal values over all options, and $w = \max_{a \in \mathcal{A}} \min_{s \in \mathcal{S}} \gamma_d(s)v(s)$. Solving for $t$ we have our result:

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

Lemma 1 then follows from combining Lemmas 5 and 6

### C. Proof of Lemma 2

**Proof.** We have

$$q^\mu_t - q^\mu_{\gamma_p} = \mathcal{P}^\mu_{\gamma_p} q^\mu_t - \mathcal{P}^\mu_{\gamma_p} q^\mu_{\gamma_p} = \mathcal{P}^\mu_{\gamma_p} (q^\mu_t - q^\mu_{\gamma_p}) + (\mathcal{P}^\mu_{\gamma_p} - \mathcal{P}^\mu_{\gamma_p}) q^\mu_{\gamma_p} = (I - \mathcal{P}^\mu_{\gamma_p})^{-1}(\mathcal{P}^\mu_{\gamma_p} - \mathcal{P}^\mu_{\gamma_p}) q^\mu_{\gamma_p}. \quad (7)$$

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

$$\|A q^\mu_{\gamma_p}\| = \|(\mathcal{P}^\mu_{\gamma_p} - \mathcal{P}^\mu_{\gamma_p}) q^\mu_{\gamma_p}\|$$

$$= \|\gamma_p^\beta \gamma_d q^\mu_{\gamma_p} + \gamma_p^\beta (1 - \beta^\gamma_p) \mathcal{P}^\mu_{\gamma_p} q^\mu_{\gamma_p} - \gamma_p^\beta \gamma_d q^\mu_{\gamma_p} + \gamma_p^\beta (1 - \beta^\gamma_p) \mathcal{P}^\mu_{\gamma_p} q^\mu_{\gamma_p}\|$$

$$= \|\gamma_p^\beta (\gamma_p - \gamma_p^\gamma_r) q^\mu_{\gamma_p} + \mathcal{P}^\mu (1 - \beta^\gamma_p) \mathcal{P}^\mu_{\gamma_p} q^\mu_{\gamma_p}\|$$

$$= \|\gamma_p^\beta (\gamma_p - \gamma_p^\gamma_r) q^\mu_{\gamma_p} + \mathcal{P}^\mu (1 - \beta^\gamma_p) \mathcal{P}^\mu_{\gamma_p} q^\mu_{\gamma_p}\|$$

$$= \|\gamma_p^\beta (\gamma_p - \gamma_p^\gamma_r) q^\mu_{\gamma_p} + \mathcal{P}^\mu (1 - \beta^\gamma_p) \mathcal{P}^\mu_{\gamma_p} q^\mu_{\gamma_p}\|$$

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

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

**D. Proof of Proposition 1**

**Proof.** From Eq. (7):

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

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

$$(\mathcal{P}^\mu_{\gamma_p} - \mathcal{P}^\mu_{\gamma_p}) q^\mu_{\gamma_p} = 0,$$

$$\mathcal{P}^\mu_{\gamma_p} q^\mu_{\gamma_p} = \mathcal{P}^\gamma_{\gamma_p} q^\mu_{\gamma_p}.$$
Per-Decision Option Discounting

and hence:

$$
\sum_{s'} P_{o \gamma_p}(s'|s) \gamma_d(s, s') \sum_{o'} \mu(o'|s') q_{\gamma_r}^\mu(s', o')
= \sum_{s'} P_{o \gamma_r}(s'|s) \sum_{o'} \mu(o'|s') q_{\gamma_r}^\mu(s', o').
$$

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

$$
P_{o \gamma_p}(s'|s) \gamma_d(s, s') \sum_{o'} \mu(o'|s') q_{\gamma_r}^\mu(s', o')
= P_{o \gamma_r}(s'|s) \sum_{o'} \mu(o'|s') q_{\gamma_r}^\mu(s', o'),
$$

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