A LIST OF NOTATION

$:=$	defined to be equal
$\mathbb{N}$	the natural numbers, starting with 0
$\Delta\mathcal{Y}$	the set of all probability distributions on $\mathcal{Y}$
$\mathcal{X}^*$	the set of all finite strings over the alphabet $\mathcal{X}$
$\mathcal{X}^\infty$	the set of all infinite strings over the alphabet $\mathcal{X}$
$\mathcal{A}$	the (finite) set of possible actions
$\mathcal{E}$	the (finite) set of possible percepts
α, β	two different actions, $\alpha, \beta \in \mathcal{A}$
a_t	the action in time step t
e_t	the percept in time step t
r_t	the reward in time step t, bounded between 0 and 1
$æ_{<t}$	the history up to time $t-1$, i.e., the first $t-1$ interactions, $a_1e_1a_2e_2 \ldots a_{t-1}e_{t-1}$
ϵ	the history of length 0
ε	a small positive real number
γ	the discount function $\gamma : \mathbb{N} \to \mathbb{R}_{\geq 0}$
Γ_t	a discount normalization factor, $\Gamma_t := \sum_{i=t}^{\infty} \gamma_i$
$H_t(\varepsilon)$	the ε-effective horizon, defined in (1)
π	a (stochastic) policy, i.e., a function $\pi : (\mathcal{A} \times \mathcal{E})^* \to \Delta\mathcal{A}$
π^*_ν	an optimal policy for environment ν
V^π_ν	value of the policy π in environment ν
n, k, i	natural numbers
t	(current) time step
m	time step at the end of an effective horizon
$\mathcal{M}$	a countable class of environments
ν, μ, ρ	environments from $\mathcal{M}$, i.e., functions $\nu : (\mathcal{A} \times \mathcal{E})^* \times \mathcal{A} \to \Delta\mathcal{E}$; μ is the true environment
ξ	Bayesian mixture over all environments in $\mathcal{M}$

B OMITTED PROOFS

Let P and Q be two probability distributions. We say P is *absolutely continuous with respect to* Q ($P \ll Q$) iff $Q(E) = 0$ implies $P(E) = 0$ for all measurable sets E. If $P \ll Q$ then there is a function dP/dQ called *Radon-Nikodym derivative* such that

$$\int f dP = \int f \frac{dP}{dQ} dQ$$

for all measurable functions f. This function dP/dQ can be seen as a density function of P with respect to the background measure Q.

Proof of Lemma 2. Let P, R, and Q be probability measures with $P \ll Q$ and $R \ll Q$ (we can take $Q := P/2 + R/2$), let dP/dQ and dR/dQ denote their Radon-Nikodym derivative with respect to Q, and let X denote a random variable with values in $[0,1]$. Then

$$\begin{aligned} \int X dP - \int X dR &= \int \left(X \frac{dP}{dQ} - X \frac{dR}{dQ} \right) dQ \\ &\leq \int_A X \left(\frac{dP}{dQ} - \frac{dR}{dQ} \right) dQ \end{aligned}$$

with $A := \left\{ x \;\middle|\; \frac{dP}{dQ}(x) - \frac{dR}{dQ}(x) \geq 0 \right\}$

$$\begin{aligned} &\leq \int_A \left(\frac{dP}{dQ} - \frac{dR}{dQ} \right) dQ \\ &= P(A) - R(A) \\ &\leq \sup_A |P(A) - R(A)| = D(P, R) \end{aligned}$$

From this also follows $\int X dR - \int X dP \leq D(R, P)$, and since D is symmetric we get

$$\left| \int X dP - \int X dR \right| \leq D(P, R). \tag{9}$$

According to Definition 1, the value function is the expectation of the random variable $\sum_{k=t}^{m} \gamma_k r_k / \Gamma_t$ that is bounded between 0 and 1. Therefore we can use (9) with $P := \nu^{\pi_1}(\cdot \mid æ_{<t})$ and $R := \rho^{\pi_2}(\cdot \mid æ_{<t})$ on the space $(\mathcal{A} \times \mathcal{E})^m$ of the histories of length $\leq m$ to conclude that $|V^{\pi_1,m}_\nu(æ_{<t}) - V^{\pi_2,m}_\rho(æ_{<t})|$ is bounded by $D_m(\nu^{\pi_1}, \rho^{\pi_2} \mid æ_{<t})$. □

Proof of Lemma 5. From Blackwell-Dubins' theorem [BD62] we get $D_\infty(\mu^\pi, \xi^\pi \mid æ_{<t}) \to 0$ μ^π-almost surely, and since D is bounded, this convergence also occurs in mean. Thus for every environment $\nu \in \mathcal{M}$,

$$\mathbb{E}^\pi_\nu \big[D_\infty(\nu^\pi, \xi^\pi \mid æ_{<t}) \big] \to 0 \text{ as } t \to \infty. \tag{10}$$

Now

$$\begin{aligned} &\mathbb{E}^\pi_\mu [F^\pi_\infty(æ_{<t})] \\ &\leq \frac{1}{w(\mu)} \mathbb{E}^\pi_\xi [F^\pi_\infty(æ_{<t})] \\ &= \frac{1}{w(\mu)} \mathbb{E}^\pi_\xi \left[\sum_{\nu \in \mathcal{M}} w(\nu \mid æ_{<t}) D_\infty(\nu^\pi, \xi^\pi \mid æ_{<t}) \right] \\ &= \frac{1}{w(\mu)} \mathbb{E}^\pi_\xi \left[\sum_{\nu \in \mathcal{M}} w(\nu) \frac{\nu^\pi(æ_{<t})}{\xi^\pi(æ_{<t})} D_\infty(\nu^\pi, \xi^\pi \mid æ_{<t}) \right] \\ &= \frac{1}{w(\mu)} \sum_{\nu \in \mathcal{M}} w(\nu) \mathbb{E}^\pi_\nu \big[D_\infty(\nu^\pi, \xi^\pi \mid æ_{<t}) \big] \to 0 \end{aligned}$$

by [Hut05, Lem. 5.28ii] since total variation distance is bounded. □

Proof of Lemma 12. By Assumption 10a we have $\gamma_t > 0$ for all t and hence $\Gamma_t > 0$ for all t. By Assumption 10b have that γ is monotone decreasing, so we get for all $n \in \mathbb{N}$

$$\Gamma_t = \sum_{k=t}^{\infty} \gamma_k \leq \sum_{k=t}^{t+n-1} \gamma_t + \sum_{k=t+n}^{\infty} \gamma_k = n\gamma_t + \Gamma_{t+n}.$$

And with $n := H_t(\varepsilon)$ this yields

$$\frac{\gamma_t H_t(\varepsilon)}{\Gamma_t} \geq 1 - \frac{\Gamma_{t+H_t(\varepsilon)}}{\Gamma_t} \geq 1 - \varepsilon > 0. \tag{11}$$

In particular, this bound holds for all t and $\varepsilon > 0$.

Next, we define a series of nonnegative weights $(b_t)_{t\geq 1}$ such that

$$\sum_{t=t_0}^{m} d_k = \sum_{t=t_0}^{m} \frac{b_t}{\Gamma_t} \sum_{k=t}^{m} \gamma_k d_k.$$

This yields the constraints

$$\sum_{k=t_0}^{t} \frac{b_k}{\Gamma_k} \gamma_t = 1 \quad \forall t \geq t_0.$$

The solution to these constraints is

$$b_{t_0} = \frac{\Gamma_{t_0}}{\gamma_{t_0}}, \text{ and } b_t = \frac{\Gamma_t}{\gamma_t} - \frac{\Gamma_t}{\gamma_{t-1}} \text{ for } t > t_0. \tag{12}$$

Thus we get

$$\begin{aligned}
\sum_{t=t_0}^{m} b_t &= \frac{\Gamma_{t_0}}{\gamma_{t_0}} + \sum_{t=t_0+1}^{m} \left(\frac{\Gamma_t}{\gamma_t} - \frac{\Gamma_t}{\gamma_{t-1}} \right) \\
&= \frac{\Gamma_{m+1}}{\gamma_m} + \sum_{t=t_0}^{m} \left(\frac{\Gamma_t}{\gamma_t} - \frac{\Gamma_{t+1}}{\gamma_t} \right) \\
&= \frac{\Gamma_{m+1}}{\gamma_m} + m - t_0 + 1 \\
&\leq \frac{H_m(\varepsilon)}{1-\varepsilon} + m - t_0 + 1
\end{aligned}$$

for all $\varepsilon > 0$ according to (11).

Finally,

$$\begin{aligned}
\sum_{t=1}^{m} d_t &\leq \sum_{t=1}^{t_0} d_t + \sum_{t=t_0}^{m} \frac{b_t}{\Gamma_t} \sum_{k=t}^{m} \gamma_k d_k \\
&\leq t_0 + \sum_{t=t_0}^{m} \frac{b_t}{\Gamma_t} \sum_{k=t}^{\infty} \gamma_k d_k - \sum_{t=t_0}^{m} \frac{b_t}{\Gamma_t} \sum_{k=m+1}^{\infty} \gamma_k d_k
\end{aligned}$$

and using the assumption (5) and $d_t \geq -1$,

$$\begin{aligned}
&< t_0 + \sum_{t=t_0}^{m} b_t \varepsilon + \sum_{t=t_0}^{m} \frac{b_t \Gamma_{m+1}}{\Gamma_t} \\
&\leq t_0 + \frac{\varepsilon H_m(\varepsilon)}{1-\varepsilon} + \varepsilon(m - t_0 + 1) + \sum_{t=t_0}^{m} \frac{b_t \Gamma_{m+1}}{\Gamma_t}
\end{aligned}$$

For the latter term we substitute (12) to get

$$\begin{aligned}
\sum_{t=t_0}^{m} \frac{b_t \Gamma_{m+1}}{\Gamma_t} &= \frac{\Gamma_{m+1}}{\gamma_{t_0}} + \sum_{t=t_0+1}^{m} \left(\frac{\Gamma_{m+1}}{\gamma_t} - \frac{\Gamma_{m+1}}{\gamma_{t-1}} \right) \\
&= \frac{\Gamma_{m+1}}{\gamma_m} \leq \frac{H_m(\varepsilon)}{1-\varepsilon}
\end{aligned}$$

with (11). □