A Detailed proof of Theorem 4.1

A.1 Notation and Setting

First, we need to define the value function of an SMDP. In Sutton et al. (1999) it is defined as a formalism for MDP with options, that itself, by the demonstration presented in the same article, is an SMDP.
In our case, however, for the SMDP model, we are considering an additional dependency on $h \in [0, H]$.
Notation used:

- H is the horizon
- μ policy over options $\{\mu : S \times O \times H \to [0, 1]\}$
- $r(s, o, h)$ is the discounted cumulative reward gained by selecting the option o, in state s, in the instant h of the horizon H
- $p(s', h'|s, o, h)$ is a new transition model that characterizes both the state dynamic and the time the option executes.
- $w(s, o, h)$ is the probability of playing option o being in state s at time-step h

The value function is defined as:

$$V^{\mu}(s, h) = \sum_{o \in O_s} \mu(s, o, h) \Big[r(s, o, h) + \sum_{s', h' > h} p(h', s'|s, o, h) V^{\mu}(s', h') \Big] \tag{14}$$

with $V^{\mu}(s, H) = 0$.

A.2 Performance Difference Lemma

Lemma 7.1. *[Performance Difference Lemma for FH-SMDP] Given two FH-SMDPs $\hat{M}$ and $\tilde{M}$ with horizon H, and respectively rewards $\hat{r}$, $\tilde{r}$ and transition probabilities $\hat{p}$, $\tilde{p}$. The difference in the performance of a policy μ_k is:*

$$\begin{aligned}
&\tilde{V}^{\mu_k}(s, 1) - \hat{V}^{\mu_k}(s, 1) \\
&= \hat{\mathbb{E}} \Bigg[\sum_{i=1}^{H} \Big(\big(\tilde{r}(s_i, o_i, h_i) - \hat{r}(s_i, o_i, h_i) \big) \\
&+ \big(\tilde{p}(s_{i+1}, h_{i+1}|s_i, o_i, h_i) - \hat{p}(s_{i+1}, h_{i+1}|s_i, o_i, h_i) \big) \\
&\tilde{V}^{\mu_k}(s_{i+1}, h_{i+1}) \Big) \mathbb{1}\{h_i < H\} \Bigg]
\end{aligned}$$

where $\hat{\mathbb{E}}$ is the expectation taken w.r.t. $\hat{p}$ and μ_k.

Proof. $\hat{\mathbb{E}}$ is the expectation taken w.r.t. the policy μ and the transition probability $\hat{p}(s', h'|s, o, h)$, and can be rewrite as:

$$\prod_{i=1}^{H} \mu_k(s_i, o_i, h_i) \hat{p}_k(s_{i+1}, h_{i+1}|s_i, o_i, h_i) \mathbb{1}\{h_i < H\}$$

This quantity is the distribution of visits for the policy μ_k in the " $\hat{}$ " SMDP and it is equivalent to w_{tk} for the FH-MDP case. The result follows by unrolling equation 14. Lemma E.5 Dann et al. (2017) for an example in FH-MDPs. □

A.3 Confidence Intervals

The confidence sets are defined as:

$$B_k^r(s,o,h) := [\hat{r}_k(s,o,h) - \beta_k^r(s,o,h),\ \hat{r}_k(s,o,h) + \beta_k^r(s,o,h)]$$
$$B_k^p(s,o,h) := \left\{p_k(\cdot,\cdot|s,o,h) \in \nabla(s) : \|\tilde{p}_k(\cdot,\cdot|s,o,h) - \hat{p}_k(\cdot,\cdot|s,o,h)\|_1 \leq \beta_k^p(s,o,h)\right\}$$

and the relative confidence bounds $\beta_k^r(s,o,h)$ and $\beta_k^p(s,o,h)$ using Empirical Bernstein bound (Maurer & Pontil, 2009), Hoeffding (1963) and Weissman et al. (2003).

$$\beta_k^r(s,o,h) \ \propto\ \sqrt{\frac{2\hat{\mathrm{Var}}(r)\ln 2/\delta}{n(s,o,h)}} + \frac{7\ln 2/\delta}{3(n-1)} \tag{15}$$

$$\beta_k^p(s,o,h) \ \propto\ \sqrt{\frac{S\log\left(\frac{n_k(s,o,h)}{\delta}\right)}{n_k(s,o,h)}} \tag{16}$$

with $\hat{\mathrm{Var}}(r)$ be the sample variance of r.

$$\hat{\mathrm{Var}}(r) = \frac{1}{n(n-1)}\sum_{1\leq i\leq j\leq n}(r_i - r_j)^2 \tag{17}$$

A.4 Actual Proof

Theorem 4.1. *Considering a non-stationary Finite Horizon SMDP $\mathcal{SM}$ and a set of options $\mathcal{O}$, with bounded primitive reward $r(s,a) \in [0,1]$. The regret suffered by algorithm FH-SMDP-UCRL, in K episodes of horizon H is bounded as:*

$$\textit{Regret(K)} \leq \tilde{O}\left(\left(\sqrt{SOKd^2}\right)\left(\overline{T} + \sqrt{S}H\right)\right)$$

with probability $1-\delta$.
Where:

$$\begin{aligned}\overline{T} &= \max_{s,o,h}\sqrt{\mathbb{E}[\tau(s,o,h)^2]}\\ &= \max_{s,o,h}\sqrt{\mathbb{E}[\tau(s,o,h)]^2 + \mathrm{Var}[\tau(s,o,h)]},\end{aligned}$$

τ is the holding time, and d describes the expected number of decisions in one episode.

Proof.

$$\begin{aligned}\textit{Regret(K)} &= \sum_{k=1}^{K} V^*(s,1) - \bar{V}^{\mu_k}(s,1)\\ &\overset{Opt.}{\leq} \sum_{k=1}^{K}\tilde{V}^{\mu_k}(s,1) - \bar{V}^{\mu_k}(s,1)\\ &= \sum_{k=1}^{K}\bar{\mathbb{E}}\Bigg[\sum_{i=1}^{H}\Big(\big(\tilde{r}(s_i,o_i,h_i) - \bar{r}(s_i,o_i,h_i)\big) + \big(\tilde{p}(s_{i+1},h_{i+1}|s_i,o_i,h_i) - \bar{p}(s_{i+1},h_{i+1}|s_i,o_i,h_i)\big)\tilde{V}^{\mu_k}(s_{i+1},h_{i+1})\Big)\\ &\quad \mathbb{1}\{h_i < H\}\Bigg]\end{aligned}$$

$$
\begin{aligned}
&\overset{a}{=} \sum_k \sum_{i\in[H]} \sum_{(s,o,h)\in L_k} w_k(s_i,o_i,h_i)\Big(\big(\tilde{r}(s_i,o_i,h_i) - \bar{r}(s_i,o_i,h_i)\big) \\
&+ \big(\tilde{p}(s_{i+1},h_{i+1}|s_i,o_i,h_i) - \bar{p}(s_{i+1},h_{i+1}|s_i,o_i,h_i)\big)^T \tilde{V}^{\mu_k}(s_{i+1},h_{i+1})\Big) \\
&\overset{b}{\leq} \sum_k \sum_{i\in[H]} \sum_{(s,o,h)\in L_k} w_k(s_i,o_i,h_i)\Big(2\beta_k^r(s_i,o_i,h_i) + 2\beta_k^p(s_i,o_i,h_i)^T H\Big) \\
&\overset{c}{\propto} \sum_k \sum_{i\in[H]} \sum_{(s,o,h)\in L_k} w_k(s_i,o_i,h_i)\Bigg(\sqrt{\frac{\hat{\mathrm{Var}}(r)}{n_k(s,o,h)}} + \frac{1}{n_k(s,o,h)-1} + \sqrt{\frac{S}{n_k(s,o,h)}}H\Bigg)
\end{aligned}
$$

$$
\begin{aligned}
&\overset{d}{\leq} \sum_k \sum_{i\in[H]} \sum_{(s,o,h)\in L_k} \frac{w_k(s_i,o_i,h_i)}{\sqrt{n_k(s_i,o_i,h_i)}}\Big(\sqrt{\hat{\mathrm{Var}}(r)} + \sqrt{S}H\Big) + \sum_k \sum_{i\in[H]} \sum_{(s,o,h)\in L_k} \frac{w_k(s_i,o_i,h_i)}{n_k(s,o,h)} \\
&\overset{e}{\leq} \tilde{O}\Bigg(\Big(\sqrt{dSOn}\Big)\Big(\sqrt{\hat{\mathrm{Var}}(r)} + \sqrt{S}H\Big) + dSO\Bigg) \\
&\overset{f}{\leq} \tilde{O}\Bigg(\Big(\sqrt{dSOn}\Big)\Big(R_{max}\overline{T} + \sqrt{S}H\Big) + dSO\Bigg)
\end{aligned}
$$

with

$$
\overline{T} = \max_{s,o,h} \sqrt{\mathbb{E}[\tau(o_i,s_i,h_i)^2]} = \max_{s,o,h} \sqrt{\mathbb{E}[\tau(o_i,s_i,h_i)]^2 + \mathrm{Var}[\tau(s_i,o_i,h_i)]} \tag{18}
$$

with τ representing the average duration of the set of options seen so far.

The first passage is a standard inequality when proving the regret in frameworks adopting optimism in face of uncertainty.

(a) The expectation with respect to the policy μ_k and the transition model $\bar{p}$ can be replaced with a more common formulation used in the Finite Horizon literature (Dann et al., 2017; Zanette & Brunskill, 2018), $\sum_{(s,o,h)\in L_k}$.
Where, L_k is defined as the good set (Dann et al., 2017; Zanette & Brunskill, 2018), which is the number of episodes in which the triple (s,o,h) is seen sufficiently often, and this equation is valid for all the tuples (s,o,h) being part of this set.

(b) We upper bound the difference of rewards and transition probabilities with two times their relative confidence intervals, and, the Value function at the next step with the horizon length H.

(c) We substitute the confidence intervals with their definitions (eq. 16) neglecting logarithmic terms.

(d) We divide the summation in two, to upper bound the terms separately

(e) Using the adaptation of lemma 16 of Zanette & Brunskill (2018) for SMDPs, lemma 7.2, for the first term. Using passage (b) and (c) in the proof of lemma E.1

(f) Upperbounding the sample variance of r, with $R_{max}\overline{T}$. Where $\overline{T}$ is the sample variance of the duration.

□

B Special Case of fixed-length options

Let's consider the same finite horizon MDP with options with fixed length $M :=< S, O, R_h, P_h, H, \bar{\tau} >$ where each option $o :=< I, \pi^o, \beta >$ has a fixed initial set I and fixed termination condition β, $R_h(s,o) \in [0, \bar{\tau} R_{max}]$ is the expectation of the reward function distribution, $P_h(\cdot|s,o)$ is the transition distribution, H is the horizon, and $\bar{\tau} \leq H$ is the options fixed length. The solution of the MDP will be a policy $\pi^H : S \to O$ that maximizes the cumulative return choosing among options' optimal policies π^{o_i}. The reward function over $state - options$ pairs relates to the flat-MDP's reward as:

$$R_h(s,o) = \mathbb{E}_{\substack{s_0=s \\ a_i \sim \pi^o(\cdot|s_i) \\ s_{i+1} \sim p_{h+1}(\cdot|s_i,a_i)}} \left[\sum_{i=0}^{\bar{\tau}-1} r_{h+i}(a_i, s_i) \right]$$

Denote with $V_n^{\pi^H}(s)$ the state value function associated with a hierarchical policy π^H (with Hierarchical policy we define a policy that chooses among options).

$$V_n^{\pi^H}(s) = \mathbb{E}_{\substack{s_0=s \\ o_j \sim \pi^H(\cdot|s_j) \\ s_{j+1} \sim P(\cdot|s_j,o_j)_h}} \left[\sum_{j=0}^{N} R_{h \times \bar{\tau}}(s_j, o_j) \right]$$

with $N = \frac{H}{\bar{\tau}}$ the number of decision steps that occurs during the Horizon.
In this way, we can exploit the same *performance difference lemma* of Dann (Dann et al., 2017) Lemma E.15, where, instead of actions we have fixed length options, and we sum over N decision steps. Hence, we can write:

$$\begin{aligned}
Regret(K) &\overset{Opt.}{\leq} \sum_{k=1}^{K} \tilde{V}_1^{\pi_k^H}(s) - \bar{V}_1^{\pi_k^H}(s) && (19)\\
&\overset{a}{=} \sum_k \sum_{n \in [N]} \sum_{(s,o) \in L_k} w_{nk}(s,o) \Big(\big(\tilde{R}_n(s,o) - \bar{R}_n(s,o)\big) + \big(\tilde{P}_n(s,o) - \bar{P}_n(s,o)\big)^T \tilde{V}_{n+1}^{\pi_k} \Big) && (20)\\
&+ \textit{term considering the state-options pairs inside the failure event} && (21)
\end{aligned}$$

here $w_{nk}(s,o)$ is the probability of visiting state s and choosing option o there at the decision step n in the k-th episode.

Then, we consider a new formulation of the confidence sets:

$$\begin{aligned}
B_{nk}^R(s,o) &:= [\hat{R}_{nk}(s,o) - \beta_{nk}^R(s,o),\ \hat{R}_{nk}(s,o) + \beta_{nk}^R(s,o)] \\
B_{nk}^P(s,o) &:= \big\{ P_{nk}(\cdot|s,o) \in \nabla(s) : \|\tilde{P}_{nk}(\cdot|s,o) - \hat{P}_{nk}(\cdot|s,o)\|_1 \leq \beta_{nk}^P(s,o) \big\}
\end{aligned}$$

using Hoeffding (1963) and Weissman et al. (2003) the confidence bounds are:

$$\beta_{nk}^R(s,o) \ \propto \ R_{max} \bar{\tau} \sqrt{\frac{\log\big(\frac{n_{nk}(s,o)}{\delta}\big)}{n_{nk}(s,o)}} \tag{22}$$

$$\beta_{nk}^P(s,o) \ \propto \ \sqrt{\frac{S \log\big(\frac{n_{nk}(s,o)}{\delta}\big)}{n_{nk}(s,o)}} \tag{23}$$

After the definition of the confidence sets we can bound the previous equation as follow:

$$\begin{aligned}
\sum_{k=1}^{K} \tilde{V}_1^{\pi_k^H}(s) - \bar{V}_1^{\pi_k^H}(s) &= \sum_k \sum_{n\in[N]} \sum_{(s,o)\in L_k} w_{nk}(s,o)\Big(\big(\tilde{R}_{nk}(s,o) - \bar{R}_{nk}(s,o)\big) + \big(\tilde{P}_{nk}(s,o) - \bar{P}_{nk}(s,o)\big)^T \tilde{V}_{h+1}^{\pi_k}\Big) \\
&\quad + \textit{term considering the state-options pairs inside the failure event} \\
&\overset{a}{\leq} \sum_k \sum_{n\in[N]} \sum_{(s,o)\in L_k} w_{nk}(s,o)\Big(2\beta_{nk}^R + 2\beta_{nk}^{P\,T}(H-\bar{\tau})\Big) \\
&\overset{b}{\propto} \sum_k \sum_{n\in[N]} \sum_{(s,o)\in L_k} w_{nk}(s,o)\Big(\frac{R_{max}\bar{\tau}}{\sqrt{n_{nk}}} + \sqrt{\frac{S}{n_{nk}}}(H-\bar{\tau})\Big) \\
&= \sum_k \sum_{n\in[N]} \sum_{(s,o)\in L_k} \frac{w_{nk}(s,o)}{\sqrt{n_{nk}}}\Big(R_{max}\bar{\tau} + \sqrt{S}(H-\bar{\tau})\Big) \\
&\overset{c}{\leq} \tilde{O}\Big(N\sqrt{SOK}\big(R_{max}\bar{\tau} + \sqrt{S}H - \sqrt{S}\bar{\tau}\big)\Big) \\
&\overset{d}{\leq} \tilde{O}\Big(N\sqrt{SOK}\big(R_{max}\bar{\tau} + \sqrt{S}H\big)\Big)
\end{aligned}$$

(a) substituting the $\tilde{R}_n(s,o) - \bar{R}_n(s,o)$ and $\tilde{P}_n(s,o) - \bar{P}_n(s,o)$ with double the relative confidence interval and considering $\tilde{V}_{h+1}^{\pi_k} \leq (H - \bar{\tau})$. The second term will be omitted for ease of notation.

(b) replacing the confidence intervals with their definition

(c) Lemma E.2

(d) considering the worst case, where there isn't the negative term

comparing it with the bound of Fruit & Lazaric (2017) for bounded holding time:

$$\tilde{O}\Big(\big(D_o\sqrt{S} + T_{max} + (T_{max} - T_{min})\big)R_{max}\sqrt{S_O O n}\Big)$$

that having options with fixed duration $\bar{\tau}$, and considering $R_{max} = 1$ reduces to:

$$\tilde{O}\Big(DS\sqrt{On} + \bar{\tau}\sqrt{SOn}\Big)$$

we have the same bound where instead of the diameter we have the Horizon H, and where NK is exactly equal to the number of the decisions up to episode k, which is n in their notation. We have:

$$\tilde{O}\Big(HS\sqrt{ON^2K} + \bar{\tau}\sqrt{SON^2K}\Big)$$

Important: Note that we have an additional $\sqrt{N}$ terms because we considered non-stationary MDP. This is a well-known penalty term when considering non-stationarity in the process.

C Proof of Theorem 6.2

Theorem 6.2. *The regret paid by the two-phase learning algorithm until the episode K is:*

$$Regret(K) \leq \tilde{O}\Big(K^{\frac{2}{3}}\sqrt[3]{H_O^5 S_O^2 A_O O} + HS\sqrt{Od^2K}\Big)$$

with $H_O = \max_{o\in\mathcal{O}} H_o$, and S_O, and A_O, respectively, the upper bounds on the cardinality of the state and action space of the sub-FH-MDPs.

Proof. The regret of the two-phase algorithm can be written in this form

$$\begin{aligned} Regret(K) &= \sum_{k=1}^{K_1} V_*^*(s,1) - V_{(\pi_k)}^{\mu}(s,1) + \sum_{k=k_1} V_*^*(s,1) - V_{\pi_{K_1}}^{\mu_k} \\ &= \underbrace{\sum_{k=1}^{K_1} V_*^*(s,1) - V_{(\pi_k)}^{\mu}(s,1)}_{\text{Options learning Regret}} + \sum_{k=K_1}^{K} \underbrace{V_*^*(s,1) - V_{(\pi_{K_1})}^*}_{\text{Bias}} + \underbrace{V_{(\pi_{K_1})}^* - V_{(\pi_{K_1})}^{\mu_k}}_{\text{Regret SMDP with fixed options}} \end{aligned}$$

The regret is the sum of the regret paid in the first phase and the regret paid in the second one, plus an additional bias term. By assuming all options with equal samples K_o and the options' policies learning as O finite horizon MDPs for which S_o, A_o, H_o are the upper bounds of the option's state space dimension, the option's action space dimension and the option's horizon, and $K_1 = \sum_{o \in O} K_o$ and

$$K = \sum_{o \in O} K_o + K_2 \tag{24}$$

we can write the regret as

$$Regret(K) \leq \sum_{o \in O} K_o H_o + K_2 \max_{o \in O} \frac{1}{K_o} H_o^2 S_o \sqrt{A_o K_o} + HS\sqrt{Od^2 K_2}$$

In which we pay full regret for each option learning, then the maximum average regret considering the option learning as a finite horizon MDP with horizon H_o, and the regret of the SMDP learning with fixed options. However, considering the options with equal samples and with S_O, A_O, H_O the upper bounds of the relative quantities we can get rid of the maximization in the second term, and the regret became

$$Regret(K) \leq OK_o H_O + \frac{K_2}{K_o} H_O^2 S_O \sqrt{A_O K_o} + HS\sqrt{OK_2 d^2}$$

Now, by substituting K_2 with eq. 24, and upper bounding $(K - OK) \leq K$ we can solve in closed form to find K_o to minimize the regret.

$$\begin{aligned} &Regret(K) \leq OK_o H_O + \frac{K}{K_o} H_O^2 S_O \sqrt{A_O K_o} + HS\sqrt{OKd^2} \\ &K_o = \sqrt[3]{\frac{K^2 S_O^2 H_O^2 A_O}{O^2 4}} \end{aligned}$$

Therefore, by substituting K_o in the original equation we have

$$Regret(K) \leq \tilde{O}\Big(K^{\frac{2}{3}} (H_O^5 S_O^2 A_O O)^{\frac{1}{3}} + HS\sqrt{OKd^2}\Big)$$

□

Now we can compare the regret of this algorithm compared to the regret of UCRL2 adapted for non-stationary FH-MDPs (Ghavamzadeh et al., 2020).

$$Regret(UCRL2 - CH) \leq \tilde{O}(H^2 S\sqrt{AK})$$

$$\frac{Regret_{SMDP}}{Regret_{MDP}} \leq \frac{K^{1/6} \alpha^{3/8} O^{1/3}}{(HS)^{1/3} A^{1/6}} \leq 1 \tag{25}$$

$$K \leq \frac{H^2 S^2 A}{\alpha^{16} O^2} \tag{26}$$

D Renewal Processes

Lemma D.1 (Renewal Function Bound)**.** *Considering a Renewal process, $(X_t)_{t\geq 0}$, and a sequence $S_1, S_2 \dots$ of random variables, characterizing the random duration of an event, alternatively defined as holding time, with supp$(S_i) \in \{1, \dots, H\}$. We can bound, with probability $1-\delta$, the expected number of random events that occurred up to time t, X_t, with:*

$$X_t < \sqrt{\frac{\ln 2 - \ln \delta}{cK}} + \frac{t}{\mu}$$

with $c = \frac{\mu^3}{32\sigma^2 T}$ where μ is the mean of the r.v.s and σ^2 the variance.

Proof. Based on the proof presented on Pinelis (2019), which apply DKW type inequalities to renewal processes (Dvoretzky et al., 1956)

$$\Pr\left(\sup_{0\leq t\leq T}\left|\frac{X_{nt}}{n} - \frac{t}{\mu}\right| \geq \epsilon\right) \leq 2e^{-cn\epsilon^2}$$

Now we can equal $2e^{-cn\epsilon^2}$ to δ and find ϵ.

$$\epsilon = \sqrt{\frac{\ln 2 - \ln \delta}{cn}}$$

Thus with probability $1-\delta$

$$X_t \leq \sqrt{\frac{\ln 2 - \ln \delta}{cn}} + \frac{t}{\mu}$$

that completes the proof □

Lemma 5.3. *[Bound on number of options played in one episode] Considering a Finite Horizon SMDP $\mathcal{SM}$ with horizon H and, O options with duration $\tau_{min} \leq \tau \leq \tau_{max}$ and $\min_o(\mathbb{E}[\tau_o])$ the expected duration of the shorter option. The expected number of options played in one episode d can be seen as the renewal function $m(t)$ of a renewal process up to the instant H. With probability $1-\delta$, this quantity is bounded by*

$$d < \sqrt{\frac{32(\tau_{max} - \tau_{min})H(\ln 2 - \ln \delta)}{(\min_o(\mathbb{E}[\tau_o]))^3}} + \frac{H}{\min_o(\mathbb{E}[\tau_o])}$$

Proof. The proof followed the one of lemma D.1 and the fact that we are considering $T = H$, $n = 1$, $t = H$, $\mu = \bar{\tau}$, $\sigma^2 = (\tau_{max} - \tau_{min})$, and $X_t = d$. □

E Useful Lemmas

Lemma E.1 (lemma 16 (Zanette & Brunskill, 2018) for non stationary MDPs)**.** *The following holds true:*

$$\sum_k \sum_{h\in[H]} \sum_{(s,a)\in L_k} w_{hk}(s,a)\sqrt{\frac{1}{n_k(s,a,h)}} = \tilde{O}(\sqrt{HSAT})$$

where the extra $\sqrt{H}$ is due to the non-stationarity of the environment

Proof.

$$\begin{aligned}
&\sum_k \sum_{h\in[H]} \sum_{(s,a)\in L_k} w_{hk}(s,a)\sqrt{\frac{1}{n_k(s,a,h)}} \\
&\overset{a}{\leq} \sqrt{\sum_k \sum_{h\in[H]} \sum_{(s,a)\in L_k} w_{hk}(s,a)} \sqrt{\sum_k \sum_{h\in[H]} \sum_{(s,a)\in L_k} w_{hk}(s,a)\frac{1}{n_k(s,a,h)}} \\
&\overset{b}{=} \sqrt{KH} \sqrt{\sum_k \sum_{h\in[H]} \sum_{(s,a)\in L_k} w_{hk}(s,a)\frac{1}{n_k(s,a,h)}} \\
&\overset{e}{\leq} \tilde{O}(\sqrt{HSAT})
\end{aligned}$$

Then:

$$\begin{aligned}
&\sum_k \sum_{h\in[H]} \sum_{(s,a)\in L_k} \frac{w_{hk}(s,a)}{n_k(s,a,h)} \\
&\overset{c}{\leq} \sum_{h\in[H]} \sum_{(s,a)\in L_k} \sum_k \frac{w_{hk}(s,a)}{\frac{1}{4}\sum_{j\leq k} w_{hj}(s,a)} \\
&\overset{d}{\leq} 4HSA \log\left(\frac{Ke}{w_{min}}\right) \\
&\tilde{\propto} HSA
\end{aligned}$$

(a) by Cauchy-Schwartz

(b) $\sum_{t\in[H]} \sum_{(s,a)\in L_k} w_{tk}(s,a) = H$ lemma 17 (b) Zanette & Brunskill (2018)

(c) lemma 2 Zanette & Brunskill (2018) adapted to the non-stationary case

(d) lemma E.5 Dann et al. (2017) considering that being (s,a) part of the good set L_k, then we are assuming (Appendix E.3 Dann et al. (2017)) that $w_k(s,a) \geq w_{min}$.

(e) substituting (f) we get the upper bound, and we conclude the proof.

□

Lemma 7.2. *Considering a non-stationary MDP M with a set of options as an SMDP $M_{\mathcal{O}}$ (Sutton et al., 1999). In $M_{\mathcal{O}}$ the number of decisions taken in the k^{th}-episode is a random variable d and*

$$\sum_{i\in H} \sum_{(s,o)\in L_k} w_k(s_i,o_i,h_i)\mathbb{1}\{h_i < H\} = d \text{ with } \{\forall k : d \leq H\}$$

Therefore, the following holds true:

$$\sum_k \sum_{i\in H} \sum_{(s,o)\in L_k} w_k(s_i,o_i,h_i)\sqrt{\frac{1}{n_k(s_i,o_i,h_i)}} = \tilde{O}\left(\sqrt{SOKd^2}\right)$$

or, using the same notation used in Fruit & Lazaric (2017), $\tilde{O}(\sqrt{SOKd^2})$, with $n = Kd$ the number of decisions taken up to episode K.

Proof. Due to the stochasticity of the option's duration, d is a random variable expressing the number of decisions taken in a step. Thus, first, we can rewrite passage (b) of the proof of lemma 17 Zanette & Brunskill (2018) then, we change lemma E.1 considering the same notion of good set considered in the appendix of Zanette & Brunskill (2018) and the validity of lemma 2 of Zanette & Brunskill (2018), in the options framework(replacing o with a). If all the aforementioned assumptions hold, thus the derivation of the new lemma follows the derivation of lemma E.1 □

Lemma E.2 (lemma 16 (Zanette & Brunskill, 2018) for MDPs with options of fixed lenght)**.** *For an MDP with O options, with a fixed lenght $\bar{\tau}$, where the horizon is divided in $N = \frac{H}{\bar{\tau}}$ decision steps, the following holds true:*

$$\sum_{k} \sum_{n \in N} \sum_{(s,o) \in L_k} w_{nk}(s,o) \sqrt{\frac{1}{n_k(s,o)}} = \tilde{O}\left(N\sqrt{SOK}\right)$$

Proof. In this MDP the control returns to the hierarchical policy after exactly $\bar{\tau}$ time steps (the length of an option), thus, we can have at most $N = \frac{H}{\bar{\tau}}$ actions in the horizon H. For this reason, passage (b) of the proof of lemma E.1 become

$$\sum_{n \in N} \sum_{(s,o) \in L_k} w_{nk}(s,o) = N$$

The rest results for the same passage of the proof of lemma E.1. □

To have a more complete analysis we need also to consider the triples (s, o, h) which aren't inside the good set. To do that, we can adapt Lemma 3 of Zanette & Brunskill (2018), for the FH-SMDP setting.

Lemma E.3 (Outside the good set)**.** *It holds that:*

$$\sum_{k=1}^{K} \sum_{h=1}^{d} \sum_{(s,o,h) \notin L_k} w_k(s,o,h) = \tilde{O}(SOd)$$

The proof follows from the one of lemma 3 of Zanette & Brunskill (2018).