Appendices

A. Definitions

We repeat the relevant definitions in our paper.

A1. Safe Space: For more details, see Turchetta et al. (2016).

Set of the states identified as safe up to some confidence level of ϵ_g:

$$R^{\text{safe}}_{\epsilon_g}(X) = X \cup \{\boldsymbol{s} \in \mathcal{S} \mid \exists \boldsymbol{s}' \in X : g(\boldsymbol{s}') - \epsilon_g - Ld(\boldsymbol{s}, \boldsymbol{s}') \geq h\}.$$

Set of states with reachability from X:

$$R_{\text{reach}}(X) = X \cup \{\boldsymbol{s} \in \mathcal{S} \mid \exists \boldsymbol{s}' \in X, a \in \mathcal{A}(s') : \boldsymbol{s} = f(\boldsymbol{s}', a)\}.$$

Set of states with returnability to X:

$$\begin{aligned}
&R_{\text{ret}}(X, \bar{X}) = \bar{X} \cup \{\boldsymbol{s} \in X \mid \exists a \in \mathcal{A} : f(\boldsymbol{s}, a) \in \bar{X}\},\\
&R^n_{\text{ret}}(X, \bar{X}) = R_{\text{ret}}(X, R^{n-1}_{\text{ret}}(X, \bar{X})), \text{with } R^1_{\text{ret}}(X, \bar{X}) = R_{\text{ret}}(X, \bar{X}),\\
&\bar{R}_{\text{ret}}(X, \bar{X}) = \lim_{n\to\infty} R^n_{\text{ret}}(X, \bar{X}).
\end{aligned}$$

Set of safe states with reachability and returnability:

$$\begin{aligned}
&R_{\epsilon_g}(X) = R^{\text{safe}}_{\epsilon_g}(X) \cap R_{\text{reach}}(X) \cap R_{\text{ret}}(R^{\text{safe}}_{\epsilon_g}(X), X),\\
&R_{\epsilon_g}(X) = R_{\epsilon_g}(R^{n-1}_{\epsilon_g}(X)), \text{with } R^1_{\epsilon_g}(X) = R_{\epsilon_g}(X),\\
&\bar{R}_{\epsilon_g}(X) = \lim_{n\to\infty} R^n_{\epsilon_g}(X).
\end{aligned}$$

Pessimistic safe space:

$$\begin{aligned}
S^-_t &= \{\boldsymbol{s} \in \mathcal{S} \mid \exists \boldsymbol{s}' \in \mathcal{X}^-_{t-1} : l_t(\boldsymbol{s}') - L \cdot d(\boldsymbol{s}, \boldsymbol{s}') \geq h\},\\
\mathcal{X}^-_t &= \{\boldsymbol{s} \in S^-_t \mid \boldsymbol{s} \in R_{\text{reach}}(\mathcal{X}^-_{t-1}) \cap \bar{R}_{\text{ret}}(S^-_t, \mathcal{X}^-_{t-1})\}.
\end{aligned}$$

Optimistic safe space:

$$\begin{aligned}
S^+_t &= \{\boldsymbol{s} \in \mathcal{S} \mid \exists \boldsymbol{s}' \in \mathcal{X}^+_{t-1} : u_t(\boldsymbol{s}') - L \cdot d(\boldsymbol{s}, \boldsymbol{s}') \geq h\},\\
\mathcal{X}^+_t &= \{\boldsymbol{s} \in S^+_t \mid \boldsymbol{s} \in R_{\text{reach}}(\mathcal{X}^+_{t-1}) \cap \bar{R}_{\text{ret}}(S^+_t, \mathcal{X}^+_{t-1})\}.
\end{aligned}$$

A2. Optimization of Cumulative Reward

For optimal policy:

$$V^*_{\mathcal{M}}(\boldsymbol{s}_t) = \max_{s_{t+1} \in R_{\epsilon_g}(S_0)} \left[\, r(\boldsymbol{s}_{t+1}) + \gamma V^*_{\mathcal{M}}(\boldsymbol{s}_{t+1}) \,\right].$$

For balancing exploration and exploitation (neither ES^2 nor $\mathsf{P\text{-}ES}^2$ is used):

$$\begin{aligned}
&U_t(\boldsymbol{s}) = \mu^r_t(\boldsymbol{s}) + \alpha^{1/2}_{t+1} \cdot \sigma^r_t(\boldsymbol{s}),\\
&J^*_{\mathcal{X}}(\boldsymbol{s}_t, \boldsymbol{b}^r_t, \boldsymbol{b}^g_t) = \max_{s_{t+1} \in \mathcal{X}^-_{t*}} \left[\, U_t(\boldsymbol{s}_{t+1}) + \gamma J^*_{\mathcal{X}}(\boldsymbol{s}_{t+1}, \boldsymbol{b}^r_t, \boldsymbol{b}^g_t) \,\right].
\end{aligned}$$

A3. ES^2 Algorithm

For checking whether the termination condition is satisfied:

$$\begin{aligned}
&V_{\mathcal{M}_y}(\boldsymbol{s}_t) = \max_{s_{t+1} \in \mathcal{X}_t^+} [\, r'(\boldsymbol{s}_{t+1}) + \gamma V_{\mathcal{M}_y}(\boldsymbol{s}_{t+1}) \,], \\
&\mathcal{Y}_t = \{\boldsymbol{s}' \in \mathcal{S}^+ \mid \forall \boldsymbol{s} \in \mathcal{X}_t^- : \boldsymbol{s}' = f(\boldsymbol{s}, \pi_y^*(a \mid \boldsymbol{s}))\}, \\
&\mathcal{Y}_t \subseteq \mathcal{X}_t^-.
\end{aligned}$$

For balancing exploration and exploitation in terms of reward:

$$J_{\mathcal{Y}}^*(\boldsymbol{s}_t, \boldsymbol{b}_t^r, \boldsymbol{b}_t^g) = \max_{s_{t+1} \in \mathcal{Y}_t} \big[\, U_t(\boldsymbol{s}_{t+1}) + \gamma J_{\mathcal{Y}}^*(\boldsymbol{s}_{t+1}, \boldsymbol{b}_t^r, \boldsymbol{b}_t^g) \,\big].$$

A4. P-ES^2 Algorithm

For checking whether the termination condition is satisfied:

$$\begin{aligned}
&V_{\mathcal{M}_z}(\boldsymbol{s}_t) = \max_{s_{t+1} \in \mathcal{X}_t^+} [\, P^z \cdot \{r'(\boldsymbol{s}_{t+1}) + \gamma V_{\mathcal{M}_z}(\boldsymbol{s}_{t+1})\} \,], \\
&\mathcal{Z}_t = \{\boldsymbol{s}' \in \mathcal{S}^+ \mid \forall \boldsymbol{s} \in \mathcal{X}_t^- : \boldsymbol{s}' = f(\boldsymbol{s}, \pi_z^*(a \mid \boldsymbol{s}))\}, \\
&\mathcal{Z}_t \subseteq \mathcal{X}_t^-.
\end{aligned}$$

For balancing exploration and exploitation in terms of the reward:

$$J_{\mathcal{Z}}^*(\boldsymbol{s}_t, \boldsymbol{b}_t^r, \boldsymbol{b}_t^g) = \max_{s_{t+1} \in \mathcal{Z}_t} \big[\, U_t(\boldsymbol{s}_{t+1}) + \gamma J_{\mathcal{Z}}^*(\boldsymbol{s}_{t+1}, \boldsymbol{b}_t^r, \boldsymbol{b}_t^g) \,\big].$$

B. Preliminary Lemma

Lemma 3. *For two arbitrary functions $f_1(x)$ and $f_2(x)$, the following inequality holds:*

$$\max_x f_1(x) - \max_x f_2(x) \geq \min_x (f_1(x) - f_2(x)).$$

Proof. For two arbitrary functions $f_4(x)$ and $f_5(x)$, the following inequality holds:

$$\max_x f_4(x) + \max_x f_5(x) \;\geq \max_x \{f_4(x) + f_5(x)\}.$$

Let $f_2(x) = f_4(x) + f_5(x)$ and $f_3(x) = -f_4(x)$. Then,

$$\begin{aligned}
\max_x \{-f_3(x)\} + \max_x \{f_2(x) + f_3(x)\} \;&\geq \max_x f_2(x), \\
\max_x \{f_2(x) + f_3(x)\} - \max_x f_2(x) \;&\geq -\max_x \{-f_3(x)\}, \\
\max_x \{f_2(x) + f_3(x)\} - \max_x f_2(x) \;&= \min_x f_3(x).
\end{aligned}$$

Finally, let $f_1(x) = f_2(x) + f_3(x)$. Then, the desired lemma is obtained. □

C. Near-optimality

Lemma 4. *Let $J_{\mathcal{X}}^*(\boldsymbol{s}_t, \boldsymbol{b}_t^r, \boldsymbol{b}_t^g)$ be the value function calculated by* SNO-MDP *without the* ES^2 *algorithm. Then, $J_{\mathcal{X}}^*(\boldsymbol{s}_t, \boldsymbol{b}_t^r, \boldsymbol{b}_t^g)$ satisfies the following inequality:*

$$J_{\mathcal{X}}^*(\boldsymbol{s}_t, \boldsymbol{b}_t^r, \boldsymbol{b}_t^g) \geq V^*(\boldsymbol{s}_t).$$

Proof. Consider a state $\boldsymbol{s}_t$ and beliefs $\boldsymbol{b}_t^r$ and $\boldsymbol{b}_t^g$. Also, let I denote the following safety indicator function:

$$I(\boldsymbol{s}) := \begin{cases} 1 & \text{if } \; \boldsymbol{s} \in \bar{R}_{\epsilon_g}(S_0), \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

Then, the following chain of equations and inequalities holds:

$$\begin{aligned}
&J^*_{\mathcal{X}}(\boldsymbol{s}_t, \boldsymbol{b}^r_t, \boldsymbol{b}^g_t) - V^*(\boldsymbol{s}_t) \\
&= \max_{s_{t+1} \in \mathcal{X}^-_{t^*}} [\, U_t(\boldsymbol{s}_{t+1}) + \gamma J^*_{\mathcal{X}}(\boldsymbol{s}_{t+1}, \boldsymbol{b}^r_t, \boldsymbol{b}^g_t)\,] - \max_{s_{t+1} \in \bar{R}_{\epsilon_g}(S_0)} [\, r(\boldsymbol{s}_{t+1}) + \gamma V^*_{\mathcal{M}}(\boldsymbol{s}_{t+1})\,] \\
&\geq \max_{s_{t+1} \in \bar{R}_{\epsilon_g}(S_0)} [\, U_t(\boldsymbol{s}_{t+1}) + \gamma J^*_{\mathcal{X}}(\boldsymbol{s}_{t+1}, \boldsymbol{b}^r_t, \boldsymbol{b}^g_t)\,] - \max_{s_{t+1} \in \bar{R}_{\epsilon_g}(S_0)} [\, r(\boldsymbol{s}_{t+1}) + \gamma V^*_{\mathcal{M}}(\boldsymbol{s}_{t+1})\,] \\
&= \max_{a_t} [\, I(\boldsymbol{s}_{t+1}) \cdot \{U_t(\boldsymbol{s}_{t+1}) + \gamma J^*_{\mathcal{X}}(\boldsymbol{s}_{t+1}, \boldsymbol{b}^r_t, \boldsymbol{b}^g_t)\}\,] - \max_{a_t} [\, I(\boldsymbol{s}_{t+1}) \cdot \{r(\boldsymbol{s}_{t+1}) + \gamma V^*_{\mathcal{M}}(\boldsymbol{s}_{t+1})\}\,] \\
&\geq \min_{a_t} [\, I(\boldsymbol{s}_{t+1}) \cdot \{U_t(\boldsymbol{s}_{t+1}) - r(\boldsymbol{s}_{t+1})\} + \gamma I(\boldsymbol{s}_{t+1}) J^*_{\mathcal{X}}(\boldsymbol{s}_{t+1}, \boldsymbol{b}^r_t, \boldsymbol{b}^g_t) - \gamma I(\boldsymbol{s}_{t+1}) V^*(\boldsymbol{s}_{t+1})\,] \\
&= \min_{a_t} [\, I(\boldsymbol{s}_{t+1}) \cdot \{U_t(\boldsymbol{s}_{t+1}) - r(\boldsymbol{s}_{t+1})\} + \gamma I(\boldsymbol{s}_{t+1}) \{J^*_{\mathcal{X}}(\boldsymbol{s}_{t+1}, \boldsymbol{b}^r_t, \boldsymbol{b}^g_t) - V^*(\boldsymbol{s}_{t+1})\}\,].
\end{aligned}$$

The third line follows from $\mathcal{X}^-_{t^*} \supseteq \bar{R}_{\epsilon_g}(S_0)$ in Theorem 1. Also, the fourth line follows from the definition of I, and the fifth line follows from Lemma 3. Because $\boldsymbol{s}$ is arbitrary in the above derivation, we have

$$\min_{\boldsymbol{s}_t} [\, J^*_{\mathcal{X}}(\boldsymbol{s}_t, \boldsymbol{b}^r_t, \boldsymbol{b}^g_t) - V^*(\boldsymbol{s}_t)\,] \geq \min_{\boldsymbol{s}_{t+1}} [\, I(\boldsymbol{s}_{t+1})\{U_t(\boldsymbol{s}_{t+1}) - r(\boldsymbol{s}_{t+1})\} + \gamma I(\boldsymbol{s}_{t+1}) \{J^*(\boldsymbol{s}_{t+1}, \boldsymbol{b}^r_t, \boldsymbol{b}^g_t) - V^*(\boldsymbol{s}_{t+1})\}\,].$$

By Lemma 2, the following equation holds with probability at least $1 - \Delta^r$:

$$\min_{\boldsymbol{s}_t} [\, J^*_{\mathcal{X}}(\boldsymbol{s}_t, \boldsymbol{b}^r_t, \boldsymbol{b}^g_t) - V^*(\boldsymbol{s}_t, \boldsymbol{b}^r_t, \boldsymbol{b}^g_t)\,] \geq \gamma \cdot \min_{\boldsymbol{s}_{t+1}} [I(\boldsymbol{s}_{t+1}) \{J^*_{\mathcal{X}}(\boldsymbol{s}_{t+1}, \boldsymbol{b}^r_t, \boldsymbol{b}^g_t) - V^*(\boldsymbol{s}_{t+1})\}\,]$$

Repeatedly applying this equation proves the desired lemma. Therefore, we have

$$J^*_{\mathcal{X}}(\boldsymbol{s}_t, \boldsymbol{b}^r_t, \boldsymbol{b}^g_t) \geq V^*(\boldsymbol{s}_t)$$

with high probability. □

Lemma 5. (Generalized induced inequality) *Let $\boldsymbol{b}^r, \boldsymbol{b}^g, r$ and $\hat{\boldsymbol{b}}^r, \hat{\boldsymbol{b}}^g, \hat{r}$ be the beliefs (over reward and safety, respectively) and reward functions (including the exploration bonus) that are identical on some set of states Ω — i.e., $\boldsymbol{b}^r = \hat{\boldsymbol{b}}^r$, $\boldsymbol{b}^g = \hat{\boldsymbol{b}}^g$, and $r = \hat{r}$ for all $\boldsymbol{s} \in \Omega$. Let $P(A_\Omega)$ be the probability that a state not in Ω is generated when starting from state $\boldsymbol{s}$ and following a policy π. If the value is bound in $[0, V_{\max}]$, then*

$$V^\pi(\boldsymbol{s}, \boldsymbol{b}^r, \boldsymbol{b}^g, r) \geq V^\pi(\boldsymbol{s}, \hat{\boldsymbol{b}}_r, \hat{\boldsymbol{b}}_g, \hat{r}) - V_{\max} P(A_\Omega),$$

where we now make explicit the dependence of the value function on the reward.

Proof. The lemma follows from Lemma 8 in Strehl & Littman (2005). □

Lemma 6. *Assume that the reward function r satisfies $\|r\|^2_k \leq B^r$, and that the noise n^r_t is σ_r-sub-Gaussian. If $\alpha_t = B^r + \sigma_r \sqrt{2(\Gamma^r_{t-1} + 1 + \log(1/\Delta^r))}$ and $C_r = 8/\log(1 + \sigma_r^{-2})$, then the following holds:*

$$\frac{1}{2}\sqrt{\frac{C_r \alpha_{t^*} \Gamma^r_{t^*}}{t^*}} \geq \alpha^{1/2}_{t^*} \sigma^r_{t^*}(\boldsymbol{s}),$$

with probability at least $1 - \Delta^r$.

Proof. The lemma follows from Lemma 4 in Chowdhury & Gopalan (2017). □

D. $\textsf{ES}^2$ algorithm

Lemma 7. *Assume that $\mathcal{Y}_t \subseteq \mathcal{X}^-_t$ holds. Suppose that we obtain the optimal policy, π^*_y on the basis of $J^*_{\mathcal{Y}}(\boldsymbol{s}_t, \boldsymbol{b}^r_t, \boldsymbol{b}^g_t) = \max_{s_{t+1} \in \mathcal{Y}_t} \left[U_t(\boldsymbol{s}_{t+1}) + \gamma J^*_{\mathcal{Y}}(\boldsymbol{s}_{t+1}, \boldsymbol{b}^r_t, \boldsymbol{b}^g_t)\right]$. Then, for all t, the following holds:*

$$\boldsymbol{s}_t \in \mathcal{Y}_t \Longrightarrow \boldsymbol{s}_{t+1} \in \mathcal{Y}_t.$$

Proof. When $\mathcal{Y}_t \subseteq \mathcal{X}_t^-$ holds, we have

$$\begin{aligned}\{\boldsymbol{s}' \in \mathcal{S}^+ \mid \forall \boldsymbol{s} \in \mathcal{Y}_t : \boldsymbol{s}' = f(\boldsymbol{s}, \pi_y^*(a \mid \boldsymbol{s}))\} &\subseteq \{\boldsymbol{s}' \in \mathcal{S}^+ \mid \forall \boldsymbol{s} \in \mathcal{X}_t^- : \boldsymbol{s}' = f(\boldsymbol{s}, \pi_y^*(a \mid \boldsymbol{s}))\} \\ &= \mathcal{Y}_t.\end{aligned}$$

This means that the next state $\boldsymbol{s}_{t+1}$ will be within $\mathcal{Y}_t$ if the agent is in $\mathcal{Y}_t$ and decides the action based on π_y^*. Therefore, we have the desired lemma. □

Lemma 8. *Assume that $\mathcal{Y}_t \subseteq \mathcal{X}_t^-$ holds, and let $J_{\mathcal{Y}}^*(\boldsymbol{s}_t, \boldsymbol{b}_t^r, \boldsymbol{b}_t^g)$ be the value function calculated by* SNO-MDP *with the* ES^2 *algorithm. Then, for all $\boldsymbol{s}_t \in \mathcal{X}_t^-$, $J_{\mathcal{Y}}^*(\boldsymbol{s}_t, \boldsymbol{b}_t^r, \boldsymbol{b}_t^g)$ satisfies the following equation:*

$$J_{\mathcal{Y}}^*(\boldsymbol{s}_t, \boldsymbol{b}_t^r, \boldsymbol{b}_t^g) \geq V^*(\boldsymbol{s}_t).$$

Proof. Consider a state $\boldsymbol{s}_t \in \mathcal{X}_t^-$ and beliefs $\boldsymbol{b}^r$ and $\boldsymbol{b}^g$. Also, we define the function I as in (5). Then, the following chain of the equations and inequalities holds:

$$\begin{aligned}
& J_{\mathcal{Y}}^*(\boldsymbol{s}_t, \boldsymbol{b}_t^r, \boldsymbol{b}_t^g) - V^*(\boldsymbol{s}_t) \\
= & \max_{\boldsymbol{s}_{t+1} \in \mathcal{Y}_t} \left[U_t(\boldsymbol{s}_{t+1}) + \gamma J_{\mathcal{Y}}^*(\boldsymbol{s}_{t+1}, \boldsymbol{b}_t^r, \boldsymbol{b}_t^g) \right] - \max_{a_t} \left[I(\boldsymbol{s}_{t+1}) \cdot \{ r(\boldsymbol{s}_{t+1}) + \gamma V_{\mathcal{M}}^*(\boldsymbol{s}_{t+1}) \} \right] \\
= & \max_{\boldsymbol{s}_{t+1} \in \mathcal{Y}_t} \left[U_t(\boldsymbol{s}_{t+1}) + \gamma J_{\mathcal{Y}}^*(\boldsymbol{s}_{t+1}, \boldsymbol{b}_t^r, \boldsymbol{b}_t^g) \right] - \max_{\boldsymbol{s}_{t+1} \in \mathcal{X}_t^+} \left[I(\boldsymbol{s}_{t+1}) \cdot \{ r(\boldsymbol{s}_{t+1}) + \gamma V_{\mathcal{M}}^*(\boldsymbol{s}_{t+1}) \} \right] \\
= & \max_{\boldsymbol{s}_{t+1} \in \mathcal{Y}_t} \left[U_t(\boldsymbol{s}_{t+1}) + \gamma J_{\mathcal{Y}}^*(\boldsymbol{s}_{t+1}, \boldsymbol{b}_t^r, \boldsymbol{b}_t^g) \right] - \max_{\boldsymbol{s}_{t+1} \in \mathcal{Y}_t} \left[I(\boldsymbol{s}_{t+1}) \cdot \{ r(\boldsymbol{s}_{t+1}) + \gamma V_{\mathcal{M}}^*(\boldsymbol{s}_{t+1}) \} \right] \\
\geq & \min_{\boldsymbol{s}_{t+1} \in \mathcal{Y}_t} \left[U_t(\boldsymbol{s}_{t+1}) + \gamma J_{\mathcal{Y}}^*(\boldsymbol{s}_{t+1}, \boldsymbol{b}_t^r, \boldsymbol{b}_t^g) - I(\boldsymbol{s}_{t+1}) \cdot \{ r(\boldsymbol{s}_{t+1}) + \gamma V_{\mathcal{M}}^*(\boldsymbol{s}_{t+1}) \} \right] \\
\geq & \min_{\boldsymbol{s}_{t+1} \in \mathcal{Y}_t} \left[U_t(\boldsymbol{s}_{t+1}) + \gamma J_{\mathcal{Y}}^*(\boldsymbol{s}_{t+1}, \boldsymbol{b}_t^r, \boldsymbol{b}_t^g) - \{ r(\boldsymbol{s}_{t+1}) + \gamma V_{\mathcal{M}}^*(\boldsymbol{s}_{t+1}) \} \right] \\
= & \min_{\boldsymbol{s}_{t+1} \in \mathcal{Y}_t} \left[U_t(\boldsymbol{s}_{t+1}) - r(\boldsymbol{s}_{t+1}) + \gamma J_{\mathcal{Y}}^*(\boldsymbol{s}_{t+1}, \boldsymbol{b}_t^r, \boldsymbol{b}_t^g) - \gamma V_{\mathcal{M}}^*(\boldsymbol{s}_{t+1}) \right].
\end{aligned}$$

The second and third lines follow from the definitions of I and $V_{\mathcal{M}}^*$. The forth line follows from the definition of $\mathcal{Y}$ and the assumption of $\mathcal{Y}_t \subseteq \mathcal{X}_t^-$. The fifth line follows from Lemma 3.

Then, by Lemma 2, the following equation holds with probability at least $1 - \Delta^r$:

$$\begin{aligned}
\min_{s_t \in \mathcal{X}_t^-} \left[J_{\mathcal{Y}}^*(\boldsymbol{s}_t, \boldsymbol{b}_t^r, \boldsymbol{b}_t^g) - V^*(\boldsymbol{s}_t) \} \right] &\geq \gamma \cdot \min_{\boldsymbol{s}_{t+1} \in \mathcal{Y}_t} \left[J_{\mathcal{Y}}^*(\boldsymbol{s}_{t+1}, \boldsymbol{b}_t^r, \boldsymbol{b}_t^g) - V_{\mathcal{M}}^*(\boldsymbol{s}_{t+1}) \right] \\
&\geq \gamma^2 \cdot \min_{\boldsymbol{s}_{t+2} \in \mathcal{Y}_t} \left[J_{\mathcal{Y}}^*(\boldsymbol{s}_{t+2}, \boldsymbol{b}_t^r, \boldsymbol{b}_t^g) - V_{\mathcal{M}}^*(\boldsymbol{s}_{t+2}) \right].
\end{aligned}$$

The second line follows from Lemma 7. Repeatedly applying this equation proves the desired lemma. Therefore, for all $\boldsymbol{s}_t \in \mathcal{X}_t^-$, we have

$$J_{\mathcal{Y}}^*(\boldsymbol{s}_t, \boldsymbol{b}_t^r, \boldsymbol{b}_t^g) \geq V^*(\boldsymbol{s}_t).$$

□

E. Main Theoretical Results

Theorem 1. *Assume that the safety function g satisfies $\|g\|_k^2 \leq B^g$ and is L-Lipschitz continuous. Also, assume that $S_0 \neq \emptyset$ and $g(\boldsymbol{s}) \geq h$ for all $\boldsymbol{s} \in S_0$. Fix any $\epsilon_g > 0$ and $\Delta^g \in (0, 1)$. Suppose that we conduct the stage of "exploration of safety" with the noise n_t^g being σ_g-sub-Gaussian, and that $\beta_t = B^g + \sigma_g \sqrt{2(\Gamma_{t-1}^g + 1 + \log(1/\Delta^g))}$ until $\max_{\boldsymbol{s} \in G_t} w_t(\boldsymbol{s}) < \epsilon_g$ is achieved. Finally, let t^* be the smallest integer satisfying*

$$\frac{t^*}{\beta_{t^*} \Gamma_{t^*}^g} \geq \frac{C_g |\bar{R}_0(S_0)|}{\epsilon_g^2} \cdot D(\mathcal{M}),$$

with $C_g = 8/\log(1 + \sigma_g^{-2})$. Then, the following statements jointly hold with probability at least $1 - \Delta^g$:

- $\forall t \geq 1$, $g(\boldsymbol{s}_t) \geq h$,
- $\exists t_0 \leq t^*$, $\bar{R}_{\epsilon_g}(S_0) \subseteq \mathcal{X}^-_{t_0} \subseteq \bar{R}_0(S_0)$.

Proof. This is an extension of Theorem 1 in Turchetta et al. (2016) to our settings, where t represents not the number of samples but the number of actions. □

Theorem 2. *Assume that the reward function r satisfies $\|r\|_k^2 \leq B^r$, and that the noise is σ_r-sub-Gaussian. Let π_t denote the policy followed by* SNO-MDP *at time t, and let $\boldsymbol{s}_t$ and $\boldsymbol{b}_t^r, \boldsymbol{b}_t^g$ be the corresponding state and beliefs, respectively. Let t^* be the smallest integer satisfying $\frac{t^*}{\beta_{t^*}\Gamma^g_{t^*}} \geq \frac{C_g|\bar{R}_0(S_0)|}{\epsilon_g^2} D(\mathcal{M})$, and fix any $\Delta^r \in (0,1)$. Finally, set $\alpha_t = B^r + \sigma_r\sqrt{2(\Gamma^r_{t-1} + 1 + \log(1/\Delta^r))}$ and*

$$\epsilon_V^* = V_{\max} \cdot (\Delta^g + \Sigma^r_{t^*}/R_{\max}),$$

with $\Sigma^r_{t^*} = \frac{1}{2}\sqrt{\frac{C_r \alpha_{t^*}\Gamma^r_{t^*}}{t^*}}$. Then, with high probability,

$$V^{\pi_t}(\boldsymbol{s}_t, \boldsymbol{b}_t^r, \boldsymbol{b}_t^g) \geq V^*(\boldsymbol{s}_t) - \epsilon_V^*$$

— i.e., the algorithm is ϵ_V^*-close to the optimal policy — for all but t^* time steps, while guaranteeing safety with probability at least $1 - \Delta^g$.

Proof. Define $\tilde{r}$ as the reward function (including the exploration bonus) that is used by SNO-MDP. Let $\hat{r}$ be a reward function equal to r on Ω and equal to $\tilde{r}$ elsewhere. Furthermore, let $\tilde{\pi}$ be the policy followed by SNO-MDP at time t, that is, the policy calculated on the basis of the current beliefs, (i.e., $\boldsymbol{b}_t^r$ and $\boldsymbol{b}_t^g$) and the reward $\tilde{r}$. Finally, let A_Ω be the event in which $\tilde{\pi}$ escapes from Ω. Then,

$$V^{\pi_t}(r, \boldsymbol{s}_t, \boldsymbol{b}_t^r, \boldsymbol{b}_t^g) \geq V^{\tilde{\pi}}(\hat{r}, \boldsymbol{s}_t, \boldsymbol{b}_t^r, \boldsymbol{b}_t^g) - V_{\max} P(A_\Omega)$$

by Lemma 5. In addition, note that, for all $t \geq t^*$, because $\hat{r}$ and $\tilde{r}$ differ by at most $\alpha_{t^*}^{1/2}\sigma^r_{t^*}$ at each state,

$$\begin{aligned} |V^{\tilde{\pi}}(\hat{r}, \boldsymbol{s}_t, \boldsymbol{b}_t^r, \boldsymbol{b}_t^g) - V^{\tilde{\pi}}(\tilde{r}, \boldsymbol{s}_t, \boldsymbol{b}_t^r, \boldsymbol{b}_t^g)| &\leq \frac{1}{1-\gamma} \cdot \alpha_{t^*}^{1/2}\sigma^r_{t^*}(\boldsymbol{s}) \\ &\leq V_{\max}/R_{\max} \cdot \Sigma^r_{t^*}. \end{aligned} \tag{6}$$

For the above inequality, we used Lemma 6. Here, consider the case of $\Omega = \mathcal{X}^-_{t^*}$. Once the safe region is fully explored, $P(A_\Omega) \leq \Delta^g$ holds after t^* time steps. Then, the following chain of equations and inequalities holds:

$$\begin{aligned} V^{\pi_t}(R, \boldsymbol{s}, \boldsymbol{b}) &\geq V^{\tilde{\pi}}(\hat{R}, \boldsymbol{s}, \boldsymbol{b}) - V_{\max} \cdot P(A_\Omega) \\ &= V^{\tilde{\pi}}(\hat{R}, \boldsymbol{s}, \boldsymbol{b}) - V_{\max} \cdot P(A_{\mathcal{X}^-}) \\ &\geq V^{\tilde{\pi}}(\hat{R}, \boldsymbol{s}, \boldsymbol{b}) - V_{\max} \cdot \Delta^g \\ &\geq V^{\tilde{\pi}}(\tilde{R}, \boldsymbol{s}, \boldsymbol{b}) - V_{\max} \cdot (\Delta^g + \Sigma^r_{t^*}/R_{\max}) \\ &= J^*_{\mathcal{X}}(\tilde{R}, \boldsymbol{s}, \boldsymbol{b}) - V_{\max} \cdot (\Delta^g + \Sigma^r_{t^*}/R_{\max}) \\ &\geq V^*(R, \boldsymbol{s}) - V_{\max} \cdot (\Delta^g + \Sigma^r_{t^*}/R_{\max}). \end{aligned}$$

In this derivation, the second line follows from the assumption of $\Omega = \mathcal{X}^-$, the third line follows from $P(A_{\mathcal{X}^-}) \leq \Delta^g$, the fourth line follows from (6), the fifth line follows from the fact that $\tilde{\pi}$ is precisely the optimal policy for $\tilde{R}$ and $\boldsymbol{b}$, and the final line follows from Lemma 4. □

Theorem 3. *Assume that the reward function r satisfies $\|r\|_k^2 \leq B^r$, and that the noise is σ_r-sub-Gaussian. Let π_t denote the policy followed by* SNO-MDP *with the the* ES^2 *algorithm at time t, and let $\boldsymbol{s}_t$ and $\boldsymbol{b}_t^r, \boldsymbol{b}_t^g$ be the corresponding state and beliefs, respectively. Let $\tilde{t}$ be the smallest integer for which (4) holds, and fix any $\Delta^r \in (0,1)$. Finally, set $\alpha_t = B^r + \sigma_r\sqrt{2(\Gamma^r_{t-1} + 1 + \log(1/\Delta^r))}$ and*

$$\tilde{\epsilon}_V = V_{\max} \cdot (\Delta^g + \Sigma^r_{\tilde{t}}/R_{\max}),$$

with $\Sigma_{\tilde{t}}^r = \frac{1}{2}\sqrt{\frac{C_r \alpha_{\tilde{t}} \Gamma_{\tilde{t}}^r}{\tilde{t}}}$. *Then, with high probability,*

$$V^{\pi_t}(\boldsymbol{s}_t, \boldsymbol{b}_t^r, \boldsymbol{b}_t^g) \geq V^*(\boldsymbol{s}_t) - \tilde{\epsilon}_V$$

— i.e., the algorithm is $\tilde{\epsilon}_V$*-close to the optimal policy — for all but* $\tilde{t}$ *time steps while guaranteeing safety with probability at least* $1 - \Delta^g$.

Proof. The proof of Theorem 3 is analogous to that of Theorem 2. Define $\tilde{r}$ as the reward function (including the exploration bonus) that is used by SNO-MDP. Let $\hat{r}$ be a reward function equal to r on $\mathcal{Y}$ and equal to $\tilde{r}$ elsewhere. Furthermore, let $\tilde{\pi}$ be the policy followed by SNO-MDP with the ES^2 algorithm at time t, that is, the policy calculated on the basis of the current beliefs, (i.e., $\boldsymbol{b}_t^r$ and $\boldsymbol{b}_t^g$) and the reward $\tilde{r}$. Finally, let $A_{\mathcal{Y}}$ be the event in which $\tilde{\pi}$ escapes from $\mathcal{Y}$. Then,

$$V^{\pi_t}(r, \boldsymbol{s}_t, \boldsymbol{b}_t^r, \boldsymbol{b}_t^g) \geq V^{\tilde{\pi}}(\hat{r}, \boldsymbol{s}_t, \boldsymbol{b}_t^r, \boldsymbol{b}_t^g) - V_{\max} P(A_{\mathcal{Y}})$$

by Lemma 5. In addition, note that, for all $t \geq \tilde{t}$, because $\hat{r}$ and $\tilde{r}$ differ by at most $\alpha_{\tilde{t}}^{1/2} \sigma_{\tilde{t}}^r$ at each state,

$$\begin{aligned} |V^{\tilde{\pi}}(\hat{r}, \boldsymbol{s}_t, \boldsymbol{b}_t^r, \boldsymbol{b}_t^g) - V^{\tilde{\pi}}(\tilde{r}, \boldsymbol{s}_t, \boldsymbol{b}_t^r, \boldsymbol{b}_t^g)| &\leq \frac{1}{1-\gamma} \cdot \alpha_{\tilde{t}}^{1/2} \sigma_{\tilde{t}}^r(\boldsymbol{s}) \\ &\leq V_{\max}/R_{\max} \cdot \Sigma_{\tilde{t}}^r. \end{aligned} \tag{7}$$

For the above inequalities, we used Lemma 6. Then, the following chain of equations and inequalities holds:

$$\begin{aligned} V^{\pi_t}(R, \boldsymbol{s}, \boldsymbol{b}) &= V^{\tilde{\pi}}(\hat{R}, \boldsymbol{s}, \boldsymbol{b}) - V_{\max} \cdot P(A_{\mathcal{Y}}) \\ &\geq V^{\tilde{\pi}}(\hat{R}, \boldsymbol{s}, \boldsymbol{b}) - V_{\max} \cdot \Delta^g \\ &\geq V^{\tilde{\pi}}(\tilde{R}, \boldsymbol{s}, \boldsymbol{b}) - V_{\max} \cdot (\Delta^g + \Sigma_{\tilde{t}}^r / R_{\max}) \\ &= J_{\mathcal{Y}}^*(\tilde{R}, \boldsymbol{s}, \boldsymbol{b}) - V_{\max} \cdot (\Delta^g + \Sigma_{\tilde{t}}^r / R_{\max}) \\ &\geq V^*(R, \boldsymbol{s}) - V_{\max} \cdot (\Delta^g + \Sigma_{\tilde{t}}^r / R_{\max}). \end{aligned}$$

In this derivation, the second line follows from $P(A_{\mathcal{Y}}) \leq \Delta^g$, the third line follows from (7), the fourth line follows from the fact that $\tilde{\pi}$ is precisely the optimal policy for $\tilde{R}$ and $\boldsymbol{b}$, and the final line follows from Lemma 8. □