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 $\epsilon_g$:

$$R_{\epsilon_g}^{\text{safe}}(X) = X \cup \{ s \in S | \exists s' \in X : g(s') - \epsilon_g - Ld(s, s') \geq h \}.$$  

Set of states with reachability from $X$:

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

Set of states with returnability to $X$:

$$R_{\text{ret}}(X, \bar{X}) = \bar{X} \cup \{ s \in X | \exists a \in A : f(s, a) \in \bar{X} \},$$

$$R_{\text{ret}}^n(X, \bar{X}) = R_{\text{ret}}(X, R_{\text{ret}}^{n-1}(X, \bar{X})), \text{ with } R_{\text{ret}}^1(X, \bar{X}) = R_{\text{ret}}(X, \bar{X}),$$

$$\bar{R}_{\text{ret}}(X, \bar{X}) = \lim_{n \to \infty} R_{\text{ret}}^n(X, \bar{X}).$$

Set of safe states with reachability and returnability:

$$R_{\epsilon_g}(X) = R_{\epsilon_g}^{\text{safe}}(X) \cap R_{\text{reach}}(X) \cap R_{\text{ret}}(R_{\epsilon_g}^{\text{safe}}(X), X),$$

$$\bar{R}_{\epsilon_g}(X) = R_{\epsilon_g}(R_{\epsilon_g}^{n-1}(X)), \text{ with } R_{\epsilon_g}^1(X) = R_{\epsilon_g}(X),$$

$$\bar{R}_{\epsilon_g}(X) = \lim_{n \to \infty} R_{\epsilon_g}^n(X).$$

Pessimistic safe space:

$$S^-_t = \{ s \in S | \exists s' \in X_{t-1}^- : t_\epsilon(s') - L \cdot d(s, s') \geq h \},$$

$$X_{t^-} = \{ s \in S^-_t | s \in R_{\text{reach}}(X_{t-1}^-) \cap \bar{R}_{\text{ret}}(S^-_t, X_{t-1}^-) \}.$$  

Optimistic safe space:

$$S^+_t = \{ s \in S | \exists s' \in X_{t-1}^+ : u_\epsilon(s') - L \cdot d(s, s') \geq h \},$$

$$X_{t^+} = \{ s \in S^+_t | s \in R_{\text{reach}}(X_{t-1}^+) \cap \bar{R}_{\text{ret}}(S^+_t, X_{t-1}^+) \}.$$  

A2. Optimization of Cumulative Reward

For optimal policy:

$$V^{*}_{\lambda}(s_t) = \max_{s_{t+1} \in R_{\lambda}(s_t)} [ r(s_{t+1}) + \gamma V^{*}_{\lambda}(s_{t+1}) ].$$

For balancing exploration and exploitation (neither $ES^2$ nor $P-ES^2$ is used):

$$U_t(s) = \mu_t(s) + \alpha_t \frac{1}{2} \sigma_t^2(s),$$

$$J_{\lambda}(s_t, b^*_t, b^*_t) = \max_{s_{t+1} \in X_{t^+}^*} \left[ U_t(s_{t+1}) + \gamma J_{\lambda}(s_{t+1}, b^*_t, b^*_t) \right].$$
A3. ES² Algorithm

For checking whether the termination condition is satisfied:

\[ V_{\mathcal{M}_t}(s_t) = \max_{s_{t+1} \in \mathcal{X}_t^+} \left[ r'(s_{t+1}) + \gamma V_{\mathcal{M}_t}(s_{t+1}) \right], \]

\[ \mathcal{Y}_t = \{ s' \in \mathcal{S}^+ \mid \forall s \in \mathcal{X}_t^- : s' = f(s, \pi^*_t(a \mid s)) \}, \]

\[ \mathcal{Y}_t \subseteq \mathcal{X}_t^- . \]

For balancing exploration and exploitation in terms of reward:

\[ J^*_x(s_t, b^*_t, b^*_t) = \max_{s_{t+1} \in \mathcal{Y}_t} \left[ U_t(s_{t+1}) + \gamma J^*_x(s_{t+1}, b^*_t, b^*_t) \right] . \]

A4. P-ES² Algorithm

For checking whether the termination condition is satisfied:

\[ V_{\mathcal{M}_t}(s_t) = \max_{s_{t+1} \in \mathcal{X}_t^+} \left[ P^z \cdot \{ r'(s_{t+1}) + \gamma V_{\mathcal{M}_t}(s_{t+1}) \} \right], \]

\[ \mathcal{Z}_t = \{ s' \in \mathcal{S}^+ \mid \forall s \in \mathcal{X}_t^- : s' = f(s, \pi^*_t(a \mid s)) \}, \]

\[ \mathcal{Z}_t \subseteq \mathcal{X}_t^- . \]

For balancing exploration and exploitation in terms of the reward:

\[ J^*_z(s_t, b^*_t, b^*_t) = \max_{s_{t+1} \in \mathcal{Z}_t} \left[ U_t(s_{t+1}) + \gamma J^*_z(s_{t+1}, b^*_t, b^*_t) \right] . \]

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,

\[ \max_x \{-f_3(x)\} + \max_x \{ f_2(x) + f_3(x) \} \geq \max_x f_2(x) , \]

\[ \max_x \{ f_2(x) + f_3(x) \} - \max_x f_2(x) \geq -\max_x \{-f_3(x)\} , \]

\[ \max_x \{ f_2(x) + f_3(x) \} - \max_x f_2(x) = \min_x f_3(x) . \]

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

C. Near-optimality

Lemma 4. Let \( J^*_x(s_t, b^*_t, b^*_t) \) be the value function calculated by SNO-MDP without the ES² algorithm. Then, \( J^*_x(s_t, b^*_t, b^*_t) \) satisfies the following inequality:

\[ J^*_x(s_t, b^*_t, b^*_t) \geq V^*(s_t) . \]

Proof. Consider a state \( s_t \) and beliefs \( b^*_t \) and \( b^*_t \). Also, let \( I \) denote the following safety indicator function:

\[ I(s) := \begin{cases} 1 & \text{if } s \in \bar{R}_{\epsilon_*}(S_0), \\ 0 & \text{otherwise}. \end{cases} \]
Then, the following chain of equations and inequalities holds:

\[
J^*_X(s_t, b^r_t, b^g_t) - V^*(s_t) \\
= \max_{s_{t+1} \in X^*_t} \left[ U_t(s_{t+1}) + \gamma J^*_X(s_{t+1}, b^r_{t+1}, b^g_{t+1}) \right] - \max_{s_{t+1} \in R_g(S_0)} \left[ r(s_{t+1}) + \gamma V^*_M(s_{t+1}) \right] \\
\geq \max_{s_{t+1} \in R_g(S_0)} \left[ U_t(s_{t+1}) + \gamma J^*_X(s_{t+1}, b^r_{t+1}, b^g_{t+1}) \right] - \max_{s_{t+1} \in R_g(S_0)} \left[ r(s_{t+1}) + \gamma V^*_M(s_{t+1}) \right] \\
= \max_{a_t} \left[ I(s_{t+1}) \cdot \{ U_t(s_{t+1}) + \gamma J^*_X(s_{t+1}, b^r_{t+1}, b^g_{t+1}) \} \right] - \max_{a_t} \left[ I(s_{t+1}) \cdot \{ r(s_{t+1}) + \gamma V^*_M(s_{t+1}) \} \right] \\
\geq \min_{a_t} \left[ I(s_{t+1}) \cdot \{ U_t(s_{t+1}) - r(s_{t+1}) \} + \gamma I(s_{t+1}) J^*_X(s_{t+1}, b^r_{t+1}, b^g_{t+1}) - \gamma I(s_{t+1}) V^*(s_{t+1}) \right] \\
= \min_{a_t} \left[ I(s_{t+1}) \cdot \{ U_t(s_{t+1}) - r(s_{t+1}) \} + \gamma I(s_{t+1}) \{ J^*_X(s_{t+1}, b^r_{t+1}, b^g_{t+1}) - V^*(s_{t+1}) \} \right].
\]

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

\[
\min_{a_t} \left[ J^*_X(s_t, b^r_t, b^g_t) - V^*(s_t) \right] \geq \min_{a_t} \left[ I(s_{t+1}) \{ U_t(s_{t+1}) - r(s_{t+1}) \} + \gamma I(s_{t+1}) \{ J^*(s_{t+1}, b^r_{t+1}, b^g_{t+1}) - V^*(s_{t+1}) \} \right].
\]

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

\[
\min_{a_t} \left[ J^*_X(s_t, b^r_t, b^g_t) - V^*(s_t, b^r_t, b^g_t) \right] \geq \gamma \cdot \min_{a_t} \left[ I(s_{t+1}) \{ J^*(s_{t+1}, b^r_{t+1}, b^g_{t+1}) - V^*(s_{t+1}) \} \right].
\]

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

\[
J^*_X(s_t, b^r_t, b^g_t) \geq V^*(s_t)
\]

with high probability.

\[\text{Lemma 5. (Generalized induced inequality)}\]

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

\[V^*(s, b^r, b^g, r) \geq V^*(s, \hat{b}^r, \hat{b}^g, \hat{r}) - V_{\text{max}} P(A_{\Omega}),\]

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

\[\text{Proof.}\] The lemma follows from Lemma 8 in Strehl & Littman (2005).

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

\[\frac{1}{2} \chi_{t} \sigma_r \sqrt{\Gamma_{t-1}} \geq \alpha_t^{1/2} \sigma_r^r(s),\]

with probability at least \( 1 - \Delta^r \).

\[\text{Proof.}\] The lemma follows from Lemma 4 in Chowdhury & Gopalan (2017).

\[\text{D. ES}^2 \text{ algorithm}\]

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

\[s_t \in \mathcal{Y}_t \implies s_{t+1} \in \mathcal{Y}_t.\]
Lemma 8. Assume that \( \mathcal{Y}_t \subseteq \mathcal{X}_t^- \) holds, and let \( J^*_Y(s_t, b'_t, b''_t) \) be the value function calculated by SNO-MDP with the ES\(^2\) algorithm. Then, for all \( s_t \in \mathcal{X}_t^- \), \( J^*_Y(s_t, b'_t, b''_t) \) satisfies the following equation:

\[
J^*_Y(s_t, b'_t, b''_t) \geq V^*(s_t).
\]

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

\[
\begin{align*}
J^*_Y(s_t, b'_t, b''_t) - V^*(s_t) & = \max_{s_{t+1} \in \mathcal{Y}_t} \left[ U_i(s_{t+1}) + \gamma J^*_Y(s_{t+1}, b'_t, b''_t) \right] - \max_{s_{t+1} \in \mathcal{Y}_t} \left[ I(s_{t+1}) \cdot \left\{ r(s_{t+1}) + \gamma V^*_M(s_{t+1}) \right\} \right] \\
& = \max_{s_{t+1} \in \mathcal{Y}_t} \left[ U_i(s_{t+1}) + \gamma J^*_Y(s_{t+1}, b'_t, b''_t) \right] - \max_{s_{t+1} \in \mathcal{Y}_t} \left[ I(s_{t+1}) \cdot \left\{ r(s_{t+1}) + \gamma V^*_M(s_{t+1}) \right\} \right] \\
& = \max_{s_{t+1} \in \mathcal{Y}_t} \left[ U_i(s_{t+1}) + \gamma J^*_Y(s_{t+1}, b'_t, b''_t) \right] - \max_{s_{t+1} \in \mathcal{Y}_t} \left[ I(s_{t+1}) \cdot \left\{ r(s_{t+1}) + \gamma V^*_M(s_{t+1}) \right\} \right] \\
& \geq \min_{s_{t+1} \in \mathcal{Y}_t} \left[ U_i(s_{t+1}) + \gamma J^*_Y(s_{t+1}, b'_t, b''_t) \right] - r(s_{t+1}) + \gamma V^*_M(s_{t+1}) \\
& \geq \min_{s_{t+1} \in \mathcal{Y}_t} \left[ U_i(s_{t+1}) + \gamma J^*_Y(s_{t+1}, b'_t, b''_t) \right] - \{ r(s_{t+1}) + \gamma V^*_M(s_{t+1}) \} \\
& = \min_{s_{t+1} \in \mathcal{Y}_t} \left[ U_i(s_{t+1}) - r(s_{t+1}) + \gamma J^*_Y(s_{t+1}, b'_t, b''_t) - \gamma V^*_M(s_{t+1}) \right].
\end{align*}
\]

The second and third lines follow from the definitions of \( I \) and \( V^*_M \). 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 \( 1 - \Delta^v \):

\[
\min_{s_t \in \mathcal{X}_t^-} \left[ J^*_Y(s_t, b'_t, b''_t) - V^*(s_t) \right] \geq \gamma \cdot \min_{s_{t+1} \in \mathcal{Y}_t} \left[ J^*_Y(s_{t+1}, b'_t, b''_t) - V^*_M(s_{t+1}) \right] \\
\geq \gamma^2 \cdot \min_{s_{t+1} \in \mathcal{Y}_t} \left[ J^*_Y(s_{t+1}, b'_t, b''_t) - V^*_M(s_{t+1}) \right].
\]

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

\[
J^*_Y(s_t, b'_t, b''_t) \geq V^*(s_t).
\]

**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 \( g(s) \geq \bar{h} \) for all \( s \in S_0 \). Fix any \( \epsilon_g > 0 \) and \( \Delta^g \in (0, 1) \). Suppose that we conduct the stage of “exploration of safety” with the noise \( n^g_t \) being \( \sigma^* \)-sub-Gaussian, and that \( \beta_t = B^g + \sigma^g \sqrt{2(\Gamma_{t-1} + 1 + \log(1/\Delta^g))} \) until \( \max_{s \in G_t} w_t(s) < \epsilon_g \) is achieved. Finally, let \( t^* \) be the smallest integer satisfying

\[
\frac{t^*}{\beta_t \Gamma_t^g} \geq C_2 \frac{\tilde{P}_0(S_0)}{\epsilon_g^2} \cdot D(M),
\]

with \( C_2 = 8/\log(1 + \sigma_g^{-2}) \). Then, the following statements jointly hold with probability at least \( 1 - \Delta^g \):
• ∀t ≥ 1, g(s_t) ≥ h,

• ∃t ≥ t^*, \bar{R}_t(s_0) ⊆ X_{t^*} ⊆ \bar{R}_0(s_0).

\textbf{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. \hfill \square

\textbf{Theorem 2.} Assume that the reward function r satisfies \(\|r\|_r^2 ≤ B^r\), and that the noise is \(σ_r\)-sub-Gaussian. Let \(π_t\) denote the policy followed by SNO-MDP at time t, and let \(s_t\) and \(b_t\) be the corresponding state and beliefs, respectively. Let \(t^*\) be the smallest integer satisfying \(\frac{t^*}{βr + 1} ≥ \frac{C_1(\bar{R}_0(s_0))}{ε_σ} D(M)\), and fix any \(Δ^r ∈ (0, 1)\). Finally, set \(α_t = B^r + σ_r \sqrt{2(Γ_{t-1}^r + 1 + \log(1/Δ^r))}\) and

\[\epsilon^r_v = V_{\max} \cdot (Δ^g + Σ^r_t / R_{\max}),\]

with \(Σ^r_t = 1/2 \sqrt{\frac{C_1(α_t, Γ_{t-1}^r)}{r}}.\) Then, with high probability,

\[V^{π_t}(s_t, b_t^r, b_t^0) ≥ V^*(s_t) - \epsilon^r_v\]

— i.e., the algorithm is \(ε^r_v\)-close to the optimal policy — for all but \(t^*\) time steps, while guaranteeing safety with probability at least \(1 - Δ^g\).

\textbf{Proof.} Define \(\tilde{r}\) as the reward function (including the exploration bonus) that is used by SNO-MDP. Let \(r\) be a reward function equal to \(r\) on \(Ω\) and equal to \(\tilde{r}\) elsewhere. Furthermore, let \(\tilde{π}\) be the policy followed by SNO-MDP at time \(t\), that is, the policy calculated on the basis of the current beliefs, (i.e., \(b_t^r\) and \(b_t^0\)) and the reward \(\tilde{r}\). Finally, let \(A_Ω\) be the event in which \(\tilde{π}\) escapes from \(Ω\). Then,

\[V^{π_t}(r, s_t, b_t^r, b_t^0) ≥ V^{\tilde{π}}(\tilde{r}, s_t, b_t^r, b_t^0) - V_{\max} P(A_Ω)\]

by Lemma 5. In addition, note that, for all \(t \geq t^*\), because \(\tilde{r}\) and \(\tilde{r}\) differ by at most \(α_t \cdot σ_{t^*}^r\) at each state,

\[|V^{\tilde{π}}(\tilde{r}, s_t, b_t^r, b_t^0) - V^{\tilde{π}}(\tilde{r}, s_t, b_t^r, b_t^0)| ≤ \frac{1}{1 - γ} \cdot α_t \cdot σ_{t^*}^r(s)\]

\[≤ V_{\max}/R_{\max} \cdot Σ^r_t.\]  \hfill (6)

For the above inequality, we used Lemma 6. Here, consider the case of \(Ω = X_{t^*}\). Once the safe region is fully explored, \(P(A_Ω) ≤ Δ^g\) holds after \(t^*\) time steps. Then, the following chain of equations and inequalities holds:

\[V^{π_t}(R, s, b) ≥ V^{\tilde{π}}(\tilde{R}, s, b) - V_{\max} \cdot P(A_Ω)\]

\[= V^{\tilde{π}}(\tilde{R}, s, b) - V_{\max} \cdot P(A_{X^c-})\]

\[≥ V^{\tilde{π}}(\tilde{R}, s, b) - V_{\max} \cdot Δ^g\]

\[≥ V^{\tilde{π}}(\tilde{R}, s, b) - V_{\max} \cdot (Δ^g + Σ^r_t / R_{\max})\]

\[= J^*_X(\tilde{R}, s, b) - V_{\max} \cdot (Δ^g + Σ^r_t / R_{\max})\]

\[≥ V^*(\tilde{R}, s, b) - V_{\max} \cdot (Δ^g + Σ^r_t / R_{\max}).\]

In this derivation, the second line follows from the assumption of \(Ω = X^c\), the third line follows from \(P(A_{X^-}) ≤ Δ^g\), the fourth line follows from (6), the fifth line follows from the fact that \(\tilde{π}\) is precisely the optimal policy for \(\tilde{R}\) and \(b\), and the final line follows from Lemma 4. \hfill \square

\textbf{Theorem 3.} Assume that the reward function r satisfies \(\|r\|_r^2 ≤ B^r\), and that the noise is \(σ_r\)-sub-Gaussian. Let \(π_t\) denote the policy followed by SNO-MDP with the ES² algorithm at time t, and let \(s_t\) and \(b_t^r\), \(b_t^0\) be the corresponding state and beliefs, respectively. Let \(\hat{t}\) be the smallest integer for which (4) holds, and fix any \(Δ^r ∈ (0, 1)\). Finally, set \(α_t = B^r + σ_r \sqrt{2(Γ_{t-1}^r + 1 + \log(1/Δ^r))}\) and

\[\hat{ε}_v = V_{\max} \cdot (Δ^g + Σ^r_t / R_{\max}),\]
Safe Reinforcement Learning in Constrained Markov Decision Processes

with $\Sigma_\tilde{t} = \frac{1}{2} \sqrt{\frac{\Sigma_\gamma \alpha \tilde{t}}{\tilde{t}}}$. Then, with high probability,

$$V^{\pi_t}(s_t, b_t^r, b_t^g) \geq V^*(s_t) - \tilde{\epsilon}_V$$

— 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., $b_t^r$ and $b_t^g$) and the reward $\tilde{r}$. Finally, let $A_{\mathcal{Y}}$ be the event in which $\tilde{\pi}$ escapes from $\mathcal{Y}$. Then,

$$V^{\pi_t}(r, s_t, b_t^r, b_t^g) \geq V^{\tilde{\pi}}(\tilde{r}, s_t, b_t^r, b_t^g) - V_{\max} \cdot P(A_{\mathcal{Y}})$$

by Lemma 5. In addition, note that, for all $t \geq \tilde{t}$, because $\hat{r}$ and $\tilde{r}$ differ by at most $\alpha_1^{1/2} \sigma_t^r$ at each state,

$$|V^{\tilde{\pi}}(\hat{r}, s_t, b_t^r, b_t^g) - V^{\tilde{\pi}}(\tilde{r}, s_t, b_t^r, b_t^g)| \leq \frac{1}{1 - \gamma} \cdot \alpha_1^{1/2} \sigma_t^r(s) \leq V_{\max}/R_{\max} \cdot \Sigma_t^r.$$  (7)

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

$$V^{\pi_t}(R, s, b) = V^{\tilde{\pi}}(\tilde{R}, s, b) - V_{\max} \cdot P(A_{\mathcal{Y}})$$

$$\geq V^{\tilde{\pi}}(\tilde{R}, s, b) - V_{\max} \cdot \Delta^g$$

$$\geq V^{\tilde{\pi}}(\tilde{R}, s, b) - V_{\max} \cdot (\Delta^g + \Sigma_t^r/R_{\max})$$

$$= J^{\gamma}\ast(\tilde{R}, s, b) - V_{\max} \cdot (\Delta^g + \Sigma_t^r/R_{\max})$$

$$\geq V^*(R, s) - V_{\max} \cdot (\Delta^g + \Sigma_t^r/R_{\max}).$$

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 $R$ and $b$, and the final line follows from Lemma 8.  \[\Box\]