Appendix of WAFR 2022 Paper on

Sample-efficient Safe Learning for Online Nonlinear Control with Control Barrier Functions

*Equation indexes from (1)-(19) follow the original indexes appearing in the paper submission and new equations start from (20) in this appendix.

A Remark 1

Remark 1. *In general for nonlinear function $h^s(\cdot)$ and nonlinear dynamical system, the constraint in Eq. 4 is nonlinear with respect to the control u. When both $\hat{f}$ and d are affine control functions in the form of $G_1(x) + G_2(x)u$, the constraint in Eq. 4 becomes linear with respect to u.*

Proof. Here we discuss how to derive the control constraints with nonlinear control barrier function $h^s(\cdot)$ from Eq. 6 that fulfills Eq. 4 (and hence fulfills Proposition 1). Recall the constraint Eq. 6 as follows (we have $u_h = \pi(x_h)$).

$$h^s\left(\hat{f}(x_h, u_h) + d(x_h, u_h)\right) - L\bar{\sigma}\sqrt{2n\ln\left(\frac{Hn}{\delta_s}\right)} - h^s(x_h) \geq -\eta h^s(x_h) \tag{20}$$

Given that both the known nominal discrete dynamics $\hat{f}$ and the unknown part d are affine in control as $\hat{f}(x_h, u_h) = \hat{F}(x_h) + \hat{G}(x_h)u_h$ and $d(x_h, u_h) = g_1(x_h) + g_2(x_h)u_h$, where $\hat{F}, \hat{G}, g_1, g_2$ are assumed locally Lipschitz continuous. Then with the continuously differentiable function $h^s(\cdot)$, we have $h^s\left(\hat{f}(x_h, u_h) + d(x_h, u_h)\right) - h^s(x_h) = L^{\Delta}_{\hat{F}+g_1} h^s(x_h) + L^{\Delta}_{\hat{G}+g_2} h^s(x_h)u_h$ and hence Eq.20 can be re-written as

$$L^{\Delta}_{\hat{F}+g_1} h^s(x_h) + L^{\Delta}_{\hat{G}+g_2} h^s(x_h)u_h \geq -\eta h^s(x_h) + L\bar{\sigma}\sqrt{2n\ln\left(\frac{Hn}{\delta_s}\right)} \tag{21}$$

where $L^{\Delta}_{\hat{F}+g_1} h^s(x_h)$ and $L^{\Delta}_{\hat{G}+g_2} h^s(x_h)$ are discrete-time Lie-derivatives of $h^s(x_h)$ obtained through Taylor's theorem along $\hat{F}(x_h) + g_1(x_h)$ and $\hat{G}(x_h) + g_2(x_h)$ respectively. To that end, the condition in Eq. 4 and Eq. 6 hold by enforcing the linear control constraint Eq. 21 on u_h. Thus, we conclude the proof. □

B Assumption 2

Assumption 2. *(Calibrated model) With $\overline{W}_0, V_0$ from the initial data $(x_i, u_i, x_i')_{i=1}^N$ and $\epsilon, \delta \in (0,1)$, we can build the initial confidence ball describing the uncertain region of the linear mapping $W^\star$ with probability at least $1-\delta$ as follows:*

$$\texttt{Ball}_0 = \left\{W : \left\|(W - \overline{W}_0)V_0^{1/2}\right\|_2 \leq \beta, \quad \|W\|_2 \leq \|W^\star\|_2\right\} \tag{10}$$

where β is a hyper-parameter describing an appropriate confidence radius. Then for all $\widetilde{W} \in \texttt{Ball}_0$, we have:

$$\forall x, u \in \mathcal{X} \times \mathcal{U} : \left\| \left(\widetilde{W} - W^\star \right) \phi(x, u) \right\|_2 \leq \mathcal{O}(\epsilon).$$

We introduce the following Lemma 1 to provide a practical example on how to derive such calibrated model in Eq. 10.

Lemma 1. *[Pre-train guarantee of calibrated model from pre-collected data] Fix a pair (ϵ, δ) with $\epsilon, \delta \in (0, 1)$. Denote $\Phi \in \mathbb{R}^{r \times N}$ where each column of Φ corresponds to the feature vector $\phi(x, u)$ for $(x, u) \in \mathcal{X} \times \mathcal{U}$. Assume span$(\Phi) = r$. Via John's theorem, denote $\mathcal{B} \subset \mathcal{X} \times \mathcal{U}$ as the core set of John's ellipsoid, and μ as the corresponding sampling distribution with support on $\mathcal{B}$ defined by $\mu = \arg\max_{\rho \in \Delta(\mathcal{X} \times \mathcal{U})} \ln\det\left(\mathbb{E}_{x,u \sim \rho} \phi(x, u)\phi(x, u)^\top\right)$ from John's ellipsoid. Then draw N triples $\mathcal{D} = \{x_i, u_i, x_i'\}_{i=1}^N$ as pre-collected offline dataset with $x_i, u_i \sim \mu, x_i' \sim P(\cdot|x_i, u_i)$, and compute the initialization $\overline{W}_0 = \sum_{i=1}^N (x_i' - \hat{f}(x_i, u_i))\phi(x_i, u_i)^\top V_0^{-1}$ with $V_0 = \sum_{i=1}^N \phi(x_i, u_i)\phi(x_i, u_i)^\top + \lambda I$. Then with probability at least $1 - \delta$, we have:*

$$\forall x, u \in \mathcal{X} \times \mathcal{U}, \quad \left\| (\overline{W}_0 - W^\star)\phi(x, u) \right\|_2 \leq O(\epsilon),$$

with polynomially number of samples, i.e., N scaling polynomially with respect to the relevant parameters:

$$N = \mathcal{O}\Big(\frac{r C_1^2 \lambda + r\bar{\sigma}^2 n + \ln(1/\delta) + r^2\bar{\sigma}^2}{\epsilon^2} + \frac{C_1^2 r^2 \ln(r/\delta)}{\epsilon^4} \Big)$$

After deriving $\overline{W}_0, V_0$ from the initial data $(x_i, u_i, x_i')_{i=1}^N$, we can build the initial confidence ball describing the uncertain region of $W^\star$ as follows:

$$\texttt{Ball}_0 = \left\{ W : \left\| (W - \overline{W}_0) V_0^{1/2} \right\|_2 \leq \beta, \quad \|W\|_2 \leq \|W^\star\|_2 \right\} \tag{10}$$

where β is the confidence radius as $\beta := \sqrt{\lambda} C_1 + \bar{\sigma}\sqrt{8n \ln(5) + 8r \ln(1 + N/\lambda) + 8\ln(1/\delta)}$. For all $\widetilde{W} \in \texttt{Ball}_0$, we also have

$$\forall x, u \in \mathcal{X} \times \mathcal{U} : \left\| \left(\widetilde{W} - W^\star \right) \phi(x, u) \right\|_2 \leq \mathcal{O}(\epsilon).$$

Proof. First note that we can compute the exact difference between the least square solution $\overline{W}_0$ and $W^\star$:

$$\overline{W}_0 - W^\star = -\lambda W^\star (V_0)^{-1} + \sum_{i=1}^N \epsilon_i \phi(x_i, u_i)^\top V_0^{-1}.$$

Continue, we have

$$\begin{aligned}
\left\| (\overline{W}_0 - W^\star) V_0^{1/2} \right\|_2 &\leq \left\| \lambda W^\star V_0^{-1/2} \right\|_2 + \left\| \sum_{i=1}^N \epsilon_i \phi(x_i, u_i)^\top V_0^{-1/2} \right\|_2 \\
&\leq \sqrt{\lambda} C_1 + \bar{\sigma}\sqrt{8n \ln(5) + 8 \ln\left(\det(1 + V_0/\lambda)\right) + 8\ln(1/\delta)} \\
&\leq \underbrace{\sqrt{\lambda} C_1 + \bar{\sigma}\sqrt{8n \ln(5) + 8r \ln(1 + N/\lambda) + 8\ln(1/\delta)}}_{:=\beta}
\end{aligned}$$

where C_1 denotes the standard assumption of bounded norm $\|W^\star\|_2 \leq C_1$. Denote $\Sigma = \mathbb{E}_{x,u\sim\mu}\phi(x,u)\phi(x,u)^\top$. Via matrix Bernstein's inequality, we get that with probability at least $1-\delta$, for any x with $\|x\|_2 \leq 1$,

$$\left| x^\top \left(\sum_{i=1}^{N} \phi(x_i,u_i)\phi(x_i,u_i)^\top/N - \Sigma \right) x \right| \leq \frac{2\ln(8r/\delta)}{3N} + \sqrt{\frac{2\ln(8r/\delta)}{N}} := \varepsilon.$$

Thus we will have that for any x with $\|x\|_2 \leq 1$:

$$x^\top(\overline{W}_0 - W^\star)V_0(\overline{W}_0 - W^\star)^\top x \geq x^\top(\overline{W}_0 - W^\star)(\Sigma N + \lambda)(\overline{W}_0 - W^\star)^\top x - 2\varepsilon N C_1,$$

which means that:

$$\begin{aligned}\left\|(\overline{W}_0 - W^\star)(\Sigma+\lambda/N)^{1/2}\right\|_2^2 \leq& \beta^2/N + 2C_1\varepsilon \\ \leq& \frac{\lambda C_1^2}{N} + \frac{\bar\sigma^2(n + r\ln(1+N/\lambda+\ln(1/\delta))}{N} + \frac{2C_1\sqrt{\ln(8r/\delta)}}{\sqrt{N}}\end{aligned}$$

For any x, u, we have:

$$\left|(\overline{W}_0 - W^\star)\phi(x,u)\right|^2 \leq \left\|(\overline{W}_0 - W^\star)(\Sigma+\lambda/N)^{1/2}\right\|_2^2 \left\|(\Sigma+\lambda/N)^{-1/2}\phi(x,u)\right\|_2^2$$

Note that for any x, we have:

$$x^\top\Sigma^{-1}x \geq x^\top(\Sigma+\lambda/N)^{-1}x.$$

Using the John's theorem, we get that:

$$\phi(x,u)^\top(\Sigma+\lambda/N)^{-1}\phi(x,u) \leq \phi(x,u)^\top\Sigma^{-1}\phi(x,u) \leq r$$

Hence, we have:

$$\begin{aligned}\left|(\overline{W}_0 - W^\star)\phi(x,u)\right| \leq& \sqrt{\left(\frac{\beta^2}{N} + 2C_1\varepsilon\right) r} \\ \leq& \sqrt{\frac{r\lambda C_1^2}{N}} + \sqrt{\frac{r\bar\sigma^2(n + r\ln(1+N/\lambda+\ln(1/\delta))}{N}} + \sqrt{\frac{2C_1 r\sqrt{\ln(8r/\delta)}}{\sqrt{N}}}\end{aligned}$$

Now setting $N = \mathcal{O}\Big(\frac{rC_1^2\lambda + r\bar\sigma^2 n + \ln(1/\delta) + r^2\bar\sigma^2}{\epsilon^2} + \frac{C_1^2 r^2\ln(r/\delta)}{\epsilon^4}\Big)$, we ensure that:

$$\left|(\overline{W}_0 - W^\star)\phi(x,u)\right| \leq O(\epsilon).$$

Then starting from triangle inequality, we get:

$$\begin{aligned}\left|(\widetilde{W} - W^\star)\phi(x,u)\right| \leq& \left|(\widetilde{W} - \overline{W}_0)\phi(x,u)\right| + \left|(\overline{W}_0 - W^\star)\phi(x,u)\right| \\ \leq& \left\|(\widetilde{W} - \overline{W}_0)(\Sigma+\lambda/N)^{1/2}\right\|_2 \left\|(\Sigma+\lambda/N)^{-1/2}\phi(x,u)\right\|_2 \\ &+ \left\|(\overline{W}_0 - W^\star)(\Sigma+\lambda/N)^{1/2}\right\|_2 \left\|(\Sigma+\lambda/N)^{-1/2}\phi(x,u)\right\|_2 \\ \leq& \left\|(\widetilde{W} - \overline{W}_0)(\Sigma+\lambda/N)^{1/2}\right\|_2 \sqrt{r} + \left\|(\overline{W}_0 - W^\star)(\Sigma+\lambda/N)^{1/2}\right\|_2 \sqrt{r}\end{aligned}$$

We also know that for any two W_1 and W_0 with $\|W_j\|_2 \leq C_1$ with $j \in \{1, 2\}$, we have:

$$x^\top (W_1 - W_2) V_0 (W_1 - W_2)^\top x \geq x^\top (W_1 - W_2)(\Sigma N + \lambda)(W_1 - W_2)^\top x - 2\varepsilon N C_1,$$

which means that:

$$\left\| (\overline{W}_0 - \widetilde{W})(\Sigma + \lambda/N)^{1/2} \right\|_2^2 \leq \beta^2/N + 2C_1\varepsilon,$$
$$\left\| (\overline{W}_0 - W^\star)(\Sigma + \lambda/N)^{1/2} \right\|_2 \leq \beta^2/N + 2C_1\varepsilon.$$

This implies that:

$$\left| (\widetilde{W} - W^\star)\phi(x, u) \right| \leq 2\sqrt{r}\sqrt{\beta^2/N + 2C_1\varepsilon}.$$

Now recall the setup of N, β, and ε, we conclude the proof. □

As the typical assumption similar to [6], Assumption 2 represents an initially calibrated model $\overline{W}_0$, whose initial confidence region $\texttt{Ball}_0$ in Eq. 10 could yield approximately good prediction for all $\widetilde{W} \in \texttt{Ball}_0$.

C Proof of Theorem 1

Theorem 1 (Policy for Approximate High-Probability Safety Guarantee with Learned Dynamics)**.** *Under Assumption 2, consider any $\widetilde{W} \in \texttt{Ball}_0$, and define any policy $\pi_s : \mathcal{X} \mapsto \mathcal{U}$ that satisfies the CBF constraint parameterized by $\widetilde{W}$, i.e.,*

$$\forall x \in \mathcal{X} : \pi_s(x) \in \mathcal{U}_x := \Big\{ u : h^s\left(\hat{f}(x, u) + \widetilde{W}\phi(x, u)\right) - L\bar{\sigma}\sqrt{2n \ln\left(\frac{Hn}{\delta_s}\right)} \geq (1 - \eta) h^s(x) \Big\} \quad (11)$$

Then with probability at least $1 - \delta_s$, starting at any safe initial state $h^s(x_0) \geq 0$, π_s generates a safe trajectory $\{x_0, u_0, \ldots, x_{H-1}, u_{H-1}\}$, such that for all time steps $h \in [H]$, $h^s(x_h) \geq -\mathcal{O}(\frac{L\epsilon}{\eta})$, where L is the Lipschitz constant of $h^s(\cdot)$ under bounded $x \in \mathcal{X}$.

Proof. Starting from Assumption 2, we know that for any $\widetilde{W} \in \texttt{Ball}_0$, we have:

$$\left\| \left(\widetilde{W} - W^\star\right) \phi(x, u) \right\|_2 \leq \mathcal{O}(\epsilon), \forall x, u \in \mathcal{X} \times \mathcal{U}.$$

From Eq. 11 the policy selects action u_h for all time steps $h \in [H]$ such that:

$$h^s(\hat{f}(x_h, u_h) + \widetilde{W}\phi(x_h, u_h)) - L\bar{\sigma}\sqrt{2n \ln\left(\frac{Hn}{\delta_s}\right)} \geq (1 - \eta) h^s(x_h)$$

This means that for $W^\star$, we have:

$$\begin{aligned} h^s(x_{h+1}) &= h^s(\hat{f}(x_h,u_h) + W^\star\phi(x_h,u_h) + \epsilon_h) \\ &\geq h^s(\hat{f}(x_h,u_h) + \widetilde{W}\phi(x_h,u_h)) \\ &\quad - L\left\|(\widetilde{W} - W^\star)\phi(x_h,u_h)\right\|_2 - L\|\epsilon_h\|_2 \\ &\geq (1-\eta)h^s(x_h) + L\bar{\sigma}\sqrt{2n\ln\left(\frac{Hn}{\delta_s}\right)} \\ &\quad - L\epsilon - L\|\epsilon_h\|_2 \\ &\geq (1-\eta)h^s(x_h) - L\epsilon \\ &\geq (1-\eta)^2 h^s(x_{h-1}) - L\left(\epsilon + (1-\eta)\epsilon\right) \\ &\geq (1-\eta)^{h+1}h^s(x_0) - \frac{L}{\eta}\epsilon \end{aligned} \tag{22}$$

Using the initial condition that $h^s(x_0) \geq 0$, we conclude the proof. □

Despite the unbounded stochasticity of the dynamics, Eq. 22 with $h^s(x_{h+1}) \geq (1-\eta)^{h+1}h^s(x_0) - \frac{L}{\eta}\epsilon$ ensures that for all time steps $h \in [H]$, $h^s(x_h)$ is always lower bounded with a high probability, implying the probabilistic safety guarantee for the entire trajectory generated under π_s in Eq. 11.

D Proof of Proposition 2

Proposition 2. *Given the uncertainty regions $W^\star \in \{W : \|(W - \overline{W}_t)\Sigma_t^{1/2}\|_2 \leq \beta_t\}$ (Proof of Lemma B.5 in [18]) and $\texttt{Ball}_0$ (Eq. 10) with the probability of $Pr(W^\star \in \{W : \|(W - \overline{W}_t)\Sigma_t^{1/2}\|_2 \leq \beta_t\}) \geq 1-\delta$ and $Pr(W^\star \in \texttt{Ball}_0) \geq 1-\delta$, then for all t we have*

$$Pr\left(W^\star \in \texttt{Ball}_t := \texttt{Ball}_0 \cap \left\{W : \|(W - \overline{W}_t)\Sigma_t^{1/2}\|_2 \leq \beta_t\right\}\right) \geq 1 - 2\delta \tag{16}$$

where $\Pr(\cdot)$ denotes the probability of an event.

Proof. By definition,

$$\Pr\left(W^\star \notin \{W : \|(W - \overline{W}_t)\Sigma_t^{1/2}\|_2 \leq \beta_t\}\right) \leq \delta$$
$$\Pr\left(W^\star \notin \texttt{Ball}_0\right) \leq \delta$$

Thus, we have

$$\begin{aligned} &\Pr\left(W^\star \in \texttt{Ball}_t := \texttt{Ball}_0 \cap \left\{W : \|(W - \overline{W}_t)\Sigma_t^{1/2}\|_2 \leq \beta_t\right\}\right) \\ =&1 - \Pr\left(W^\star \notin \{W : \|(W - \overline{W}_t)\Sigma_t^{1/2}\|_2 \leq \beta_t\} \text{ Or } W^\star \notin \texttt{Ball}_0\right) \\ \geq&1 - 2\delta \end{aligned}$$

which concludes the proof. □

E Proof of Eq. 18 ⇒ Eq. 17

Here we show that for all state $x \in \mathcal{X}$, any $u \in \mathcal{U}$ satisfying Eq. 18 ensures $u \in \pi_s(x)$ defined in Eq. 17, thus constructing the safe policy class $\Pi_{\widetilde{W}}$.

Recall the definition of $\Pi_{\widetilde{W}}$ as follows.

$$\Pi_{\widetilde{W}} = \Big\{\pi_s \in \Pi : \forall x \in \mathcal{X}, \pi_s(x) \in \Big\{u : h^s\left(\hat{f}(x,u) + \widetilde{W}\phi(x,u)\right) - L\bar{\sigma}\sqrt{2n\ln\left(\frac{Hn}{\delta_s}\right)} \geq (1-\eta)h^s(x)\Big\}\Big\} \quad (17)$$

Consider Eq. 18:

$$u \in \mathcal{U} : L_{\hat{F}}^{\Delta}h^s(x) + L_{\hat{G}}^{\Delta}h^s(x)u - L\bar{\sigma}\sqrt{2n\ln\left(\frac{Hn}{\delta_s}\right)} \geq -\eta h^s(x) + \underbrace{|\Delta h^s(x)\widetilde{W}\phi(x,u^\star)| + |\Delta h^s(x)\widetilde{W}L_{x,\phi}(u^+ - u^-)|}_{K(x,u^\star)} \quad (18)$$

where $L_{\hat{F}}^{\Delta}h^s(x)$ and $L_{\hat{G}}^{\Delta}h^s(x)$ are discrete-time Lie-derivatives of $h^s(x)$ obtained through Taylor's theorem along $\hat{F}(x)$ and $\hat{G}(x)$ respectively. $L_{x,\phi}$ is the local Lipschitz constant vector for the known feature mapping ϕ w.r.t. u at x. $\Delta h^s(x)$ is the discrete derivative of h^s and $u^\star, u^+, u^-$ are the nominal, max and min value of u respectively. Thus we have

$$\begin{aligned} K(x,u^\star) &= |\Delta h^s(x)\widetilde{W}\phi(x,u^\star)| + |\Delta h^s(x)\widetilde{W}L_{x,\phi}(u^+ - u^-)| \\ &\geq |\Delta h^s(x)\widetilde{W}\phi(x,u^\star)| + |\Delta h^s(x)\widetilde{W}L_{x,\phi}(u - u^\star)| \\ &\geq \left|\Delta h^s(x)\widetilde{W}\phi(x,u^\star) + \Delta h^s(x)\widetilde{W}L_{x,\phi}(u - u^\star)\right| \\ &\geq -\Delta h^s(x)\widetilde{W}\phi(x,u) \end{aligned} \quad (23)$$

Then with $K(x,u^\star) \geq -\Delta h^s(x)\widetilde{W}\phi(x,u)$, from Eq. 18 we have

$$\begin{aligned} u \in \mathcal{U} : & L_{\hat{F}}^{\Delta}h^s(x) + L_{\hat{G}}^{\Delta}h^s(x)u - L\bar{\sigma}\sqrt{2n\ln\left(\frac{Hn}{\delta_s}\right)} \\ &\geq -\eta h^s(x) + K(x,u^\star) \\ &\geq -\eta h^s(x) - \Delta h^s(x)\widetilde{W}\phi(x,u) \end{aligned} \quad (24)$$

and hence

$$
\begin{aligned}
u \in \mathcal{U} : & \overbrace{L_{\hat{F}}^{\Delta} h^s(x) + L_{\hat{G}}^{\Delta} h^s(x) u + \Delta h^s(x) \widetilde{W} \phi(x,u)}^{h^s\left(\hat{f}(x,u)+\widetilde{W}\phi(x,u)\right)-h^s(x)} \\
& - L\bar{\sigma}\sqrt{2n \ln\left(\frac{Hn}{\delta_s}\right)} \geq -\eta h^s(x) \\
\Rightarrow \quad & h^s\left(\hat{f}(x,u) + \widetilde{W}\phi(x,u)\right) - L\bar{\sigma}\sqrt{2n \ln\left(\frac{Hn}{\delta_s}\right)} \\
& \geq (1-\eta) h^s(x) \qquad (25)
\end{aligned}
$$

As Eq. 25 is equivalent to the constraint in Eq. 17, we conclude the proof. □

Note that if the ground truth dynamics d is only state-dependent as assumed in [6, 9, 37], then Eq. 18 is also linear in control where $L_{x,\phi} = \mathbf{0}$ and feature mapping becomes $\phi(x, u^\star) = \phi(x)$.

F Proof of Theorem 2

Below we first briefly summarize the theorem of LC^3 regret from [18] as follows.

Theorem 3. *(LC^3 Regret for finite dimensional, bounded features, See Theorem 1.1 in [18]) Consider the finite dimension of ϕ as d_ϕ and that ϕ is uniformly bounded with $\|\phi(x,u)\|_2 \leq B$. The LC^3 algorithm (Algorithm 1 in [18]) enjoys the following expected regret bound:*

$$
\begin{aligned}
& \mathbb{E}_{LC^3}[Regret_T] \\
\leq & \widetilde{\mathcal{O}}\left(\sqrt{d_\phi(d_\phi + d_{\mathcal{X}} + H)H^3 T} \cdot \log\left(1 + \frac{B^2 \|W^\star\|_2^2}{\sigma^2}\right)\right) \qquad (26)
\end{aligned}
$$

where $\widetilde{\mathcal{O}}(\cdot)$ notation drops logarithmic factors in T and H.

By revisiting this result, we provide our main statement as follows.

Theorem 2. *[Main Result] Set $\lambda = \bar{\sigma}^2 / C_1^2$. Our algorithm learns a sequence of policies $\pi^0, \ldots, \pi^{T-1}$ in T episodes, such that in expectation, we have:*

$$
\mathbb{E}\left[Regret_T\right] \leq \widetilde{\mathcal{O}}\left(H\sqrt{Hr(r+n+H)T}\right).
$$

Also with probability at least $1 - O(\delta_s)$, we have that for all $t \in [T], h^s \in [H]$, $h(x_h^t) \geq -\mathcal{O}(L\epsilon/\eta)$.

Proof. For safety consideration, we proved that the sequence of policies learned from our Algorithm 1 satisfying Eq. 18 (and thus Eq. 17) are all approximately safe, i.e. $h^s(x_h^t) \geq -\mathcal{O}(L\epsilon/\eta)$, with probability at least $1 - O(\delta_s)$ for all $t \in [T], h \in [H]$ (See Section E and Theorem 1).

For the regret analysis, our proof mainly follows Theorem 3 for LC^3 algorithm and its proofs in [18]. Readers are encouraged to refer to [18] for more details. One key assumption that allows for regret bound in Eq. 26 lies in the setting of optimism in the face of uncertainty that computes the optimal policy from unconstrained policy class $\mathit{\Pi}$

$$\pi^t := \arg\min_{\pi \in \mathit{\Pi}} \min_{W \in \texttt{Ball}_t} J^\pi(x_0^t; c, W) \tag{27}$$

Similarly, in our analysis, by considering the constrained policy class $\mathit{\Pi}_{\widehat{W}}$ defined in Eq. 17 and our optimism setup in Eq. 19 analogous to Eq. 27, our regret analysis naturally follows LC^3 regret in Eq. 26 and enjoys the regret bound with safety guarantee as follows

$$\mathbb{E}\left[\text{Regret}_T\right] \leq \widetilde{\mathcal{O}}\left(H\sqrt{Hr(r+n+H)T}\right)$$

where $\widetilde{\mathcal{O}}(\cdot)$ notation drops logarithmic factors. Thus we conclude the proof of Theorem 2. □