A Proofs for Offline Policy Optimization

Recall that we have a fixed latent sequence $z_{1:T}$ such that for round t, latent state z_t parameterizes the underlying distribution of reward $r_t \in [0, 1]$. Also recall that we have IPS estimator $\hat{V}$ given in (1), where the clipping parameter M can be ignored by only considering policies in $\mathcal{H}$. In this section, we denote by $\tilde{V}$ the IPS estimator in (1) with the true latent states $z_{1:T}$. By Lemma 1, we know that $\tilde{V}$ is unbiased.

Our first result bounds the discrepancy between the two IPS estimators $\tilde{V}(\Pi)$ and $\hat{V}(\Pi)$:

Lemma 4. *For any $\Pi \in \mathcal{H}^{\mathcal{Z}}$ and $\delta \in (0, 1]$, $\left|\hat{V}(\Pi) - \tilde{V}(\Pi)\right| \leq M\varepsilon(T, \delta)$ holds with probability at least $1 - \delta$.*

Proof. The claim is proved as

$$\left|\hat{V}(\Pi) - \tilde{V}(\Pi)\right| = \left|\sum_{t=1}^{T} \frac{\pi_{\hat{z}_t}(a_t \mid x_t)}{p_t} r_t - \frac{\pi_{z_t}(a_t \mid x_t)}{p_t} r_t\right| \leq M \sum_{t=1}^{T} \mathbb{1}[\hat{z}_t \neq z_t] \leq M\varepsilon(T, \delta)\,.$$

The first inequality is by assuming that $\mathcal{H}$ in $\mathcal{H}^{\mathcal{Z}}$ satisfy (2). The second inequality is by Assumption 1 in Section 4 and holds with probability at least $1 - \delta$. □

Next, we bound the estimation error of $\tilde{V}(\Pi)$ from $V(\Pi)$. This error is due to the randomness in $\mathcal{D}$.

Lemma 5. *For any $\Pi \in \mathcal{H}^{\mathcal{Z}}$, logged data $\mathcal{D}$, and $\delta \in (0, 1]$, $\left|\tilde{V}(\Pi) - V(\Pi)\right| \leq M\sqrt{2T \log(2/\delta)}$ holds with probability at least $1 - \delta$.*

Proof. We define a martingale sequence $(U_t)_{t \in [T] \cup \{0\}}$ over rounds t and then use Azuma's inequality. The sequence is defined as $U_0 = 0$ and

$$U_t = U_{t-1} + \frac{\pi_{z_t}(a_t \mid x_t)}{p_t} r_t - V_t(\pi_{z_t})$$

for $t > 0$. It is easy to verify that this is a martingale. In particular, since z_t is fixed,

$$\mathbb{E}_{x_t, a_t, r_t \sim P_{z_t}, \pi_0}\left[\frac{\pi_{z_t}(a_t \mid x_t)}{p_t} r_t - V_t(\pi_{z_t}) \,\middle|\, U_0, \ldots, U_{t-1}\right] = \mathbb{E}_{x_t, a_t, r_t \sim P_{z_t}, \pi_{z_t}}\left[r_t\right] - V_t(\pi_{z_t}) = 0\,,$$

and $\mathbb{E}\left[U_t \mid U_0, \ldots, U_{t-1}\right] = U_{t-1}$ for any round t. Also, since $\Pi \in \mathcal{H}^{\mathcal{Z}}$, we have

$$\left|\frac{\pi_{z_t}(a_t \mid x_t)}{p_t} r_t - V_t(\pi_{z_t})\right| \leq M\,.$$

Finally, by Azuma's inequality, we get

$$\mathbb{P}\left(|\tilde{V}(\Pi) - V(\Pi)| \geq M\sqrt{2T \log(2/\delta)}\right) = \mathbb{P}\left(|U_T - U_0| \geq M\sqrt{2T \log(2/\delta)}\right) \leq 2 \exp\left[-\frac{4M^2 T \log(2/\delta)}{2M^2 T}\right] \leq \delta\,.$$

This concludes the proof. □

Using Lemmas 4 and 5 above, we can derive the results stated in the main paper.

Lemma 2. *For any policy $\Pi \in \mathcal{H}^{\mathcal{Z}}$, its IPS estimate $\hat{V}(\Pi)$ in (1), and true value $V(\Pi)$, we have that*

$$|V(\Pi) - \hat{V}(\Pi)| \leq M\varepsilon(T, \delta_1/2) + M\sqrt{2T \log(4/\delta_2)}$$

holds with probability at least $1 - \delta_1 - \delta_2$.

Proof. We have

$$\left|\hat{V}(\Pi) - V(\Pi)\right| \leq \left|\hat{V}(\Pi) - \tilde{V}(\Pi)\right| + \left|\tilde{V}(\Pi) - V(\Pi)\right|$$

from the triangle inequality. The result follows from Lemma 4 and Lemma 5. □

Theorem 2. *Let*

$$\hat{\Pi} = \arg\max_{\Pi \in \mathcal{H}^{\mathcal{Z}}} \hat{V}(\Pi), \quad \Pi^* = \arg\max_{\Pi \in \mathcal{H}^{\mathcal{Z}}} V(\Pi)$$

be the optimal latent policies w.r.t. the off-policy estimated value and the true value respectively. Then for any $\delta_1, \delta_2 \in (0, 1]$, we have that

$$V(\hat{\Pi}) \geq V(\Pi^*) - 2M\varepsilon(T, \delta_1/2) - 2M\sqrt{2T \log(4/\delta_2)}$$

holds with probability at least $1 - \delta_1 - \delta_2$.

Proof. We have

$$V(\Pi^*) - V(\hat{\Pi}) = \left[V(\Pi^*) - \hat{V}(\hat{\Pi})\right] + \left[\hat{V}(\hat{\Pi}) - V(\hat{\Pi})\right] \leq \left[V(\Pi^*) - \hat{V}(\Pi^*)\right] + \left[\hat{V}(\hat{\Pi}) - V(\hat{\Pi})\right],$$

where the inequality is from $\hat{\Pi}$ maximizing $\hat{V}$. By Lemma 2, we have for any $\Pi \in \mathcal{H}^{\mathcal{Z}}$ that

$$|\hat{V}(\Pi) - V(\Pi)| \leq M\varepsilon(T, \delta_1/2) + 2M\sqrt{T \log(4/\delta_2)}$$

holds with probability at least $1 - \delta_1/2 - \delta_2/2$. We apply the lemma to both $\hat{\Pi}$ and Π^*, and get the desired result. □

B Proofs for Change-Point Detector

Recall that S is the number of stationary segments, and $\tau_0 = 1 < \tau_1 < \ldots < \tau_{S-1} < T = \tau_S$ are the change-points. Also recall that we have change-point detector given by Algorithm 1 that on a high-level, computes differences in total reward across sliding windows of length w and detects a change-point if a difference exceeds threshold c. For any $i \in [S-1]$, let $W_i = [\tau_i - w, \tau_i + w]$ be w-close rounds to change-point τ_i. We also define $W = \bigcup_i W_i$ as all rounds w-close to any change-point.

First, we bound the probability of false positives, or that we declare any round $t \notin W$ as a change-point:

Lemma 6. *For any round $t \notin W$, the probability of a false detection is bounded from above as*

$$\mathbb{P}\left(\left|\mu_t^- - \mu_t^+\right| \geq c\right) \leq 4\exp\left[-\frac{wc^2}{2}\right].$$

Proof. Since $t \notin \bigcup_i W_i$, we have $\mathbb{E}\left[\mu_t^-\right] = \mathbb{E}\left[\mu_t^+\right]$. By Hoeffding's inequality, we get

$$\mathbb{P}\left(\left|\mu_t^- - \mu_t^+\right| \geq c\right) \leq \mathbb{P}\left(\left|\mu_t^- - \mathbb{E}\left[\mu_t^-\right]\right| \geq c/2\right) + \mathbb{P}\left(\left|\mu_t^+ - \mathbb{E}\left[\mu_t^+\right]\right| \geq c/2\right) \leq \exp\left[-\frac{wc^2}{2}\right].$$

This concludes the proof. □

Next we bound the probability of failing to detect a change-point in W:

Lemma 7. *For any positive $c \leq \Delta/2$ and W_i, a change-point is not detected in W_i with probability at most*

$$\mathbb{P}\left(\forall t \in W_i : \left|\mu_t^- - \mu_t^+\right| \leq c\right) \leq 4\exp\left[-\frac{wc^2}{2}\right].$$

Proof. Fix $s = \tau_i$. From $s \in W_i$, we have

$$\mathbb{P}\left(\forall t \in W_i : \left|\mu_t^- - \mu_t^+\right| \leq c\right) = 1 - \mathbb{P}\left(\exists t \in W_i : \left|\mu_t^- - \mu_t^+\right| > c\right) \leq 1 - \mathbb{P}\left(\left|\mu_s^- - \mu_s^+\right| > c\right)$$
$$= \mathbb{P}\left(\left|\mu_s^- - \mu_s^+\right| \leq c\right).$$

Note that $|\mu_s^- - \mu_s^+| \leq c$ implies that either μ_s^- or μ_s^+ is not close to its mean. More specifically, since $\mathbb{E}\left[\mu_s^-\right] = V_{s-1}(\pi_0)$, $\mathbb{E}\left[\mu_s^+\right] = V_s(\pi_0)$, and $|V_s(\pi_0) - V_{s-1}(\pi_0)| \geq \Delta$, we have

$$\mathbb{P}\left(\left|\mu_s^- - \mu_s^+\right| \leq c\right) \leq \mathbb{P}\left(\left|\mu_s^- - \mathbb{E}\left[\mu_s^-\right]\right| \geq \frac{\Delta - c}{2}\right) + \mathbb{P}\left(\left|\mu_s^+ - \mathbb{E}\left[\mu_s^+\right]\right| \geq \frac{\Delta - c}{2}\right).$$

From $2c \leq \Delta$ and by Hoeffding's inequality, the first term is bounded as

$$\mathbb{P}\left(\left|\mu_s^- - \mathbb{E}\left[\mu_s^-\right]\right| \geq \frac{\Delta - c}{2}\right) \leq \mathbb{P}\left(\left|\mu_s^- - \mathbb{E}\left[\mu_s^-\right]\right| \geq c/2\right) \leq 2\exp\left[-\frac{wc^2}{2}\right].$$

The second term is bounded analogously. Finally, we chain all inequalities and get our claim. □

Finally, we prove Theorem 1 by applying Lemma 6 to all rounds $t \notin W$, Lemma 7 to all change-points, and then chaining them by the union bound.

Theorem 1. *Let $\tau_i - \tau_{i-1} > 4w$ for all $i \in [L]$. For any $\delta \in (0, 1]$, and c and w in Algorithm 1 such that*

$$\Delta/2 \geq c \geq \sqrt{2\log(8T/\delta)/w}\,,$$

then Algorithm 1 estimates $\hat{z}_{1:T}$ so $\sum_{t=1}^T \mathbb{1}[\hat{z}_t \neq z_t] \leq Sw$ holds with probability at least $1 - \delta$.

Proof. Define $\delta \in (0, 1]$. We see that given w, setting c as described satisfies,

$$4T\exp\left[\frac{-wc^2}{2}\right], \quad 4k\exp\left[\frac{-wc^2}{2}\right] \leq \frac{\delta}{2}.$$

We know that $\varepsilon(T, \delta) = kw$ when all the estimated changepoints are in W (at most w rounds from a true change-point), and every $W_i \in W$ contains exactly one estimated change-point. This cannot happen if (1) a change-point is falsely detected outside W, and (2), no change-point is detected in some $W_i \in W$.

We can bound from above the probability of any error occurring with the union bound. Proposition 3 applied to every round upper-bounds the probability of (1) by $4T\exp\left(-wc^2/2\right)$. Meanwhile, Proposition 4 applied to every change-point upper-bounds the probability of (2) by $4k\exp\left(-wc^2/2\right)$. From Algorithm 1, we remove a $4w$-window around each detected changepoint, and under the assumption that $\tau_i - \tau_{i-1} > 4w$ for all $i \in [k]$, we guarantee that exactly one changepoint is detected in each W_i for true changepoint τ_i. Combining yields the total probability of an error,

$$4T\exp\left[\frac{-wc^2}{2}\right] + 4k\exp\left[\frac{-wc^2}{2}\right] \leq \delta,$$

which is the desired result. □

C Proofs for Online Deployment

Recall that we have a mixture-of-experts algorithm $\mathcal{E}$ and experts/sub-policies $\hat{\Pi} = (\hat{\pi})_{z \in \mathcal{Z}}$, such that for each round t, actions are sampled according to $a_t \sim \mathcal{E}_t(x_t, \hat{\pi})$. Let $\mathcal{E}$ be Exp4.S as described in Algorithm 6; this is similar to one proposed in Luo et al. (2018), but for stochastic experts.

Algorithm 6: Exp4.S

Input: vector of expert sub-policies $\hat{\Pi} = (\hat{\pi}_z)_{z \in \mathcal{Z}}$ with $|\mathcal{Z}| = L$, and hyperparameters $\beta, \eta > 0, \gamma \in (0, 1]$

Initialize $w_1 = (1/L, \dots, 1/L) \in [0, 1]^L$.

for $t \leftarrow 1, 2, \dots, T$ **do**

Observe x_t, and expert feedback $\hat{\pi}_z(\cdot \mid x_t)$, $\forall z \in \mathcal{Z}$.

Choose $a_t \sim \mathcal{E}_t$, where for each $a \in \mathcal{A}$,

$$\mathcal{E}_t(a) = (1 - \gamma) \sum_{z \in \mathcal{Z}} w_t(z) \hat{\pi}_z(a \mid x_t) + \frac{\gamma}{L}.$$

Observe r_t. Estimate the action costs under full feedback $\hat{c}_t(a) = \mathbb{1}[a_t = a] \frac{1 - r_t}{\mathcal{E}_t(a)}$, $\forall a \in \mathcal{A}$.

Propagate the cost to the experts $\tilde{c}_t(z) = \hat{c}_t(a_t) \hat{\pi}_z(a_t \mid x_t)$, $\forall z \in \mathcal{Z}$.

Update the distribution weights, $\tilde{w}_{t+1}(z) \propto w_t(z) \exp\left(-\eta \tilde{c}_t(z)\right)$, $\forall z \in \mathcal{Z}$.

Mix with uniform weights, $w_{t+1}(z) = (1 - \beta) w_t(z) + \beta$, $\forall z \in \mathcal{Z}$.

end

Our first result is the following regret guarantee over any stationary segment. A version of this proof for deterministic experts is in Theorem 2 of Luo et al. (2018).

Lemma 8. *Let $\mathcal{E}$ be Exp4.S as in Algorithm 6. Also, let $\gamma = 0, \eta = \sqrt{\log(L)/(\ell K)}$, and $\beta = 1/L$. Then, for any stationary segment $[\tau_{s-1}, \tau_s - 1]$ of length at most ℓ, any history up to τ_{s-1}, and any latent state $z \in \mathcal{Z}$, the regret is bounded as*

$$\sum_{t=\tau_{s-1}}^{\tau_s - 1} \mathbb{E}_{z_t, \hat{\pi}_z}\left[r_t\right] - \mathbb{E}_{z_t, \mathcal{E}_t}\left[r_t\right] \le \sqrt{2 \ell K \log(L)}\,.$$

Proof. First, we have the following upper-bound,

$$\begin{aligned} \log\left[\sum_{z' \in \mathcal{Z}} w_t(z') \exp(-\eta \tilde{c}_t(z'))\right] &\le \log\left[\sum_{z' \in \mathcal{Z}} w_t(z') \left(1 - \eta \tilde{c}_t(z') + \eta^2 \tilde{c}_t(z')^2\right)\right] \\ &\le -\eta \sum_{z' \in \mathcal{Z}} w_t(z') \tilde{c}_t(z') + \eta^2 \sum_{z' \in \mathcal{Z}} w_t(z') \tilde{c}_t(z')^2\,, \end{aligned}$$

where we use that $\exp(-x) \le 1 - x + x^2$, and $\log(1 + x) \le x$ for all $x \ge 0$. Meanwhile, for any $z \in \mathcal{Z}$, we can also bound the same quantity from below,

$$\begin{aligned} \log\left[\sum_{z' \in \mathcal{Z}} w_t(z') \exp(-\eta \tilde{c}_t(z'))\right] = \log\left[\frac{w_t(z) \exp(-\eta \tilde{c}_t(z))}{\tilde{w}_{t+1}(z)}\right] &= \log\left[\frac{w_t(z)(1 - \beta)}{w_{t+1}(z) - \beta}\right] - \eta \tilde{c}_t(z) \\ &\ge \log\left[\frac{w_t(z)}{w_{t+1}(z)}\right] - 2\beta - \eta \tilde{c}_t(z)\,, \end{aligned}$$

where for the last inequality, we use that $\log(1 - \beta) \ge -\beta/(1 - \beta) \ge -2\beta$. Combining the two inequalities, summing over all $t \in [\tau_{s-1}, \tau_s - 1]$, and telescoping yields,

$$\begin{aligned} \sum_{t=\tau_{s-1}}^{\tau_s - 1} \sum_{z' \in \mathcal{Z}} w_t(z') \tilde{c}_t(z') - \tilde{c}_t(z) &\le \frac{1}{\eta} \log\left[\frac{w_{\tau_s}(z)}{w_{\tau_{s-1}}(z)}\right] + \frac{2\beta\ell}{\eta} + \eta \sum_{t=\tau_{s-1}}^{\tau_s - 1} \sum_{z' \in \mathcal{Z}} w_t(z') \tilde{c}_t(z')^2 \\ &\le \frac{\log(1/\beta) + 2\beta\ell}{\eta} + \eta \sum_{t=\tau_{s-1}}^{\tau_s - 1} \sum_{z' \in \mathcal{Z}} w_t(z') \tilde{c}_t(z')^2\,, \end{aligned}$$

where we use that $w_t(z) \in [\beta, 1]$ for all rounds t.

When $\gamma = 0$ we know that $\hat{c}_t(a_t)$ is unbiased, or $\mathbb{E}_{z_t, \mathcal{E}_t}\left[\hat{c}_t(a_t)\right] = 1 - \mathbb{E}_{z_t, \mathcal{E}_t}\left[r_t\right]$. We also have that for any $z' \in \mathcal{Z}$,

$$\mathbb{E}_{z_t, \mathcal{E}_t}\left[\tilde{c}_t(z')\right] = \mathbb{E}_{z_t, \mathcal{E}_t}\left[\sum_{a \in \mathcal{A}} \hat{\pi}_{z'}(a \mid x_t) \hat{c}_t(a)\right] = 1 - \mathbb{E}_{z_t, \hat{\pi}_z}\left[r_t\right]\,.$$

Taking the expectation of both sides leads to,

$$\sum_{t=\tau_{s-1}}^{\tau_s-1} \mathbb{E}_{z_t,\hat{\pi}_z}\left[r_t\right] - \mathbb{E}_{z_t,\mathcal{E}_t}\left[r_t\right] \leq \frac{\log(1/\beta)+2\beta\ell}{\eta} + \eta \sum_{t=\tau_{s-1}}^{\tau_s-1} \sum_{z'\in\mathcal{Z}} \mathbb{E}_{z_t,\mathcal{E}_t}\left[w_t(z')\tilde{c}_t(z')^2\right] .$$

Next, we have that for any $z' \in \mathcal{Z}$,

$$\mathbb{E}_{z_t,\mathcal{E}_t}\left[\tilde{c}_t(z')^2\right] = \mathbb{E}_{z_t,\mathcal{E}_t}\left[\left(\frac{\hat{\pi}_{z'}(a_t \mid x_t)(1-r_t)}{\mathcal{E}_t(a_t)}\right)^2\right] \leq \sum_{a\in\mathcal{A}} \frac{\hat{\pi}_{z'}(a \mid x_t)}{\mathcal{E}_t(a)} ,$$

where we use that $a_t \sim \mathcal{E}_t$ and $r_t \in [0,1]$. Substituting this result yields,

$$\sum_{z'\in\mathcal{Z}} \mathbb{E}_{z_t,\mathcal{E}_t}\left[w_t(z')\tilde{c}_t(z')^2\right] \leq \sum_{a\in\mathcal{A}} \mathbb{E}_{z_t,\mathcal{E}_t}\left[\frac{1}{\mathcal{E}_t(a)} \sum_{z'\in\mathcal{Z}} w_t(z')\pi_{z'}(a_t \mid x_t)\right] \leq K ,$$

where we again use that $a_t \sim \mathcal{E}_t$. Substituting into the regret bound and using the values for η, β yields

$$\sum_{t=\tau_{s-1}}^{\tau_s-1} \mathbb{E}_{z_t,\hat{\pi}_z}\left[r_t\right] - \mathbb{E}_{z_t,\mathcal{E}_t}\left[r_t\right] \leq \frac{\log(1/\beta)+2\beta\ell}{\eta} + \eta K\ell \leq \sqrt{2\ell K\log(L)} ,$$

as desired. □

In practice, we do not know the lengths of stationary segments, and may not be able to find a tight upper-bound ℓ on the lengths of stationary segments. However, in our analysis, we can further partition stationary segments so that they do not exceed length ℓ at the cost of increasing the number of change-points. This is formalized in the following corollary.

Lemma 9. *Let $\mathcal{E}$ be Exp4.S as in Algorithm 6. Also, let $\gamma = 0, \eta = \sqrt{\log(L)/(\ell K)}$, and $\beta = 1/L$. Then, the total regret is bounded by*

$$\sum_{s=1}^{S} \max_{z\in\mathcal{Z}} \sum_{t=\tau_{s-1}}^{\tau_s-1} \mathbb{E}_{z_t,\hat{\pi}_z}\left[r_t\right] - \sum_{t=1}^{T} \mathbb{E}_{z_t,\mathcal{E}_t}\left[r_t\right] \leq \left(T/\sqrt{\ell} + S\sqrt{\ell}\right)\sqrt{2K\log(L)} .$$

Proof. Recall that S is the number of stationary segments within the T rounds, as defined in Section 3. Our goal is to divide the T rounds into stationary intervals of length at most ℓ, so that we can apply Lemma 8 on each interval. We do this as follows. First, we construct T/ℓ intervals of length at most T. Then, we additionally divide intervals that contain changepoints, so that each interval contains only a single latent state. This leads to at most $T/\ell + S$ stationary intervals. Finally, using Lemma 8 on each interval and summing the regrets the desired result. Note that though we consider $T/\ell + S$ intervals, we only need to consider the best latent sub-policy for each of S stationary segments, as intervals belonging to the same stationary segment have the same optimal sub-policy. □

Lemma 3. *The regret $\mathcal{R}(T;\mathcal{E},\hat{\Pi})$ is bounded from above as*

$$\begin{aligned}\mathcal{R}(T;\mathcal{E},\hat{\Pi}) \leq & \left[\sum_{t=1}^{T} \mathbb{E}_{z_t,\pi^*_{z_t}}\left[r_t\right] - \sum_{t=1}^{T} \mathbb{E}_{z_t,\hat{\pi}_{z_t}}\left[r_t\right]\right] \\ & + \left[\sum_{s=1}^{S} \max_{z\in\mathcal{Z}} \sum_{t=\tau_{s-1}}^{\tau_s-1} \mathbb{E}_{z_t,\hat{\pi}_z}\left[r_t\right] - \sum_{t=1}^{T} \mathbb{E}_{z_t,\mathcal{E}_t}\left[r_t\right]\right] .\end{aligned} \tag{4}$$

Proof. The regret can be decomposed as follows:

$$\begin{aligned}\mathcal{R}(T;\mathcal{E},\hat{\Pi}) &= \sum_{t=1}^{T} \mathbb{E}_{z_t,\pi^*_{z_t}}\left[r_t\right] - \sum_{t=1}^{T} \mathbb{E}_{z_t,\mathcal{E}_t}\left[r_t\right] \\ &= \left[\sum_{t=1}^{T} \mathbb{E}_{z_t,\pi^*_{z_t}}\left[r_t\right] - \sum_{t=1}^{T} \mathbb{E}_{z_t,\hat{\pi}_{z_t}}\left[r_t\right]\right] + \left[\sum_{t=1}^{T} \mathbb{E}_{z_t,\hat{\pi}_{z_t}}\left[r_t\right] - \sum_{t=1}^{T} \mathbb{E}_{z_t,\mathcal{E}_t}\left[r_t\right]\right] ,\end{aligned}$$

where we introduce $\hat{\Pi}$ that acts according to the true latent state. Then, recalling there are S stationary segments, the above expression can be further expressed as

$$\begin{aligned}
&\left[\sum_{t=1}^{T} \mathbb{E}_{z_t,\pi^*_{z_t}}\left[r_t\right] - \sum_{t=1}^{T} \mathbb{E}_{z_t,\hat{\pi}_{z_t}}\left[r_t\right]\right] + \left[\sum_{s=1}^{S} \sum_{t=\tau_{s-1}}^{\tau_s-1} \mathbb{E}_{z_t,\hat{\pi}_{z_t}}\left[r_t\right] - \sum_{t=1}^{T} \mathbb{E}_{z_t,\mathcal{E}_t}\left[r_t\right]\right] \\
&\leq \left[\sum_{t=1}^{T} \mathbb{E}_{z_t,\pi^*_{z_t}}\left[r_t\right] - \sum_{t=1}^{T} \mathbb{E}_{z_t,\hat{\pi}_{z_t}}\left[r_t\right]\right] + \left[\sum_{s=1}^{S} \max_{z\in\mathcal{Z}} \sum_{t=\tau_{s-1}}^{\tau_s-1} \mathbb{E}_{z_t,\hat{\pi}_{z}}\left[r_t\right] - \sum_{t=1}^{T} \mathbb{E}_{z_t,\mathcal{E}_t}\left[r_t\right]\right],
\end{aligned}$$

where we utilize the fact that each stationary segment has one optimal sub-policy. □

Theorem 3. *Let $\hat{\Pi}$ be defined as in Theorem 2 and $\mathcal{E}$ be Exp4.S Algorithm 6. Let $z_{1:T}$ be the same latent states as in offline data $\mathcal{D}$ and S be the number of stationary segments. Then for any $\delta_1, \delta_2 \in (0, 1]$, we have that*

$$\begin{aligned}
&\mathcal{R}(T; \mathcal{E}, \hat{\Pi}) \leq \\
&\quad 2M\varepsilon(T, \delta_1/2) + 2M\sqrt{2T\log(4/\delta_2)} + 2\sqrt{STK\log L}
\end{aligned}$$

holds with probability at least $1 - \delta_1 - \delta_2$.

Proof. We have the following regret decomposition due to Lemma 3,

$$\mathcal{R}(T; \mathcal{E}, \hat{\Pi}) \leq \left[\sum_{t=1}^{T} \mathbb{E}_{z_t,\pi^*_{z_t}}\left[r_t\right] - \sum_{t=1}^{T} \mathbb{E}_{z_t,\hat{\pi}_{z_t}}\left[r_t\right]\right] + \left[\sum_{s=1}^{S} \max_{z\in\mathcal{Z}} \sum_{t=\tau_{s-1}}^{\tau_s-1} \mathbb{E}_{z_t,\hat{\pi}_{z}}\left[r_t\right] - \sum_{t=1}^{T} \mathbb{E}_{z_t,\mathcal{E}_t}\left[r_t\right]\right].$$

The first term can be bounded using our offline analysis, which shows near-optimality of $\hat{\Pi}$ when the latent state is known. In the case where $z_{1:T}$ is the same both offline and online, we see that for each round t, $\mathbb{E}_{z_t,\pi^*_{z_t}}\left[r_t\right] - \mathbb{E}_{z_t,\hat{\pi}_{z_t}}\left[r_t\right] = V_t(\pi^*_{z_t}) - V_t(\hat{\pi}_{z_t})$. Hence, the first term is exactly $V(\Pi^*) - V(\hat{\Pi})$ and is bounded by Theorem 2 w.p. at least $1 - \delta_1 - \delta_2$. The second term is the switching regret of Exp4.S, and is bounded by choosing $\ell = T/S$ in Lemma 9. Combining the two bounds yields the desired result. □