## A Details of mm TS for Exponential Families

For a matrix (vector) $M$, we let $M_{i}$ denote its $i$-th row (element). Using this notation, we can write $\theta=\left(\theta_{s}\right)_{s \in \mathcal{S}}$ as a vector of parameters, one for each latent state; each $\theta_{s}$ parameterizes the reward under latent state $s$. We want to show that the sampling step in mmTS can be done tractably when the conditional reward distribution and model prior are in the exponential family.

First, we assume that we can write the conditional reward likelihood as,

$$
P(r \mid a, x, s ; \theta)=\exp \left[\phi(r, a, x)^{\top} \eta\left(\theta_{s}\right)-\psi(a, x)^{\top} g\left(\theta_{s}\right)\right]
$$

where $\phi(r, a, x)$ and $\psi(a, x)$ are sufficient statistics of the observed data, $\eta\left(\theta_{s}\right)$ are the natural parameters, and $g\left(\theta_{s}\right)$ satisfies $\psi(a, x)^{\top} g\left(\theta_{s}\right)=\log \sum_{r} \phi(r, a, x)^{\top} \eta\left(\theta_{s}\right)$, which is the log-partition function. Note that this assumption excludes rewards that are complex functions of the model parameters; however, it still handles all the non-contextual or linear bandit settings of prior work.
Then, we assume the prior over model parameters $\theta$ factors as $P_{1}(\theta)=\prod_{s \in \mathcal{S}} P_{1}\left(\theta_{s}\right)$. The prior over $\theta_{s}$ is assumed to be the conjugate prior of the likelihood and have the general form,

$$
P_{1}\left(\theta_{s}\right)=h\left(\phi_{s, 1}, m_{s, 1}\right) \exp \left[\phi_{s, 1}^{\top} \eta\left(\theta_{s}\right)-m_{s, 1}^{\top} g\left(\theta_{s}\right)\right]
$$

where $\phi_{s, 1}, m_{s, 1}$ are parameters controlling the prior, and $h\left(\phi_{s, 1}, m_{s, 1}\right)$ is the normalizing factor.
Recall that $H_{t}=\left(X_{\ell}, A_{\ell}, R_{\ell}\right)_{\ell=1}^{t-1}$ is the history up to round $t$. For round $t$, we can write the joint posterior as,

$$
\begin{align*}
P_{t}(s, \theta) & =P_{1}(s) P_{1}(\theta) P\left(H_{t} \mid s ; \theta\right) \\
& \propto P_{1}(s)\left[\prod_{s^{\prime} \neq s} P_{1}\left(\theta_{s^{\prime}}\right)\right] P_{1}\left(\theta_{s}\right) P\left(H_{t} \mid \theta_{s}\right) \propto P_{1}(s)\left[\prod_{s^{\prime} \neq s} P_{1}\left(\theta_{s^{\prime}}\right)\right] P_{t}\left(\theta_{s}\right) . \tag{7}
\end{align*}
$$

Given latent state $s$, we are only concerned with computing the posterior over the conditional model parameters $\theta_{s}$, denoted $P_{t}\left(\theta_{s}\right)$. This is given by,

$$
\begin{align*}
& P_{t}\left(\theta_{s}\right) \propto P_{1}\left(\theta_{s}\right) P\left(H_{t} \mid \theta_{s}\right)  \tag{8}\\
& \propto P_{1}\left(\theta_{s}\right) \prod_{\ell=1}^{t-1} \exp \left[\phi\left(R_{\ell}, A_{\ell}, X_{\ell}\right)^{\top} \eta\left(\theta_{s}\right)-\psi\left(A_{\ell}, X_{\ell}\right)^{\top} g\left(\theta_{s}\right)\right] \\
& \propto h\left(\phi_{s, 1}, m_{s, 1}\right) \exp \left[\phi_{s, 1}^{\top} \eta\left(\theta_{s}\right)-m_{s, 1}^{\top} g\left(\theta_{s}\right)+\left(\sum_{\ell=1}^{t-1} \phi\left(R_{\ell}, A_{\ell}, X_{\ell}\right)\right)^{\top} \eta\left(\theta_{s}\right)-\left(\sum_{\ell=1}^{t-1} \psi\left(A_{\ell}, X_{\ell}\right)\right)^{\top} g\left(\theta_{s}\right)\right] \\
& \propto h\left(\phi_{s, 1}, m_{s, 1}\right) \exp \left[\left(\phi_{s, 1}+\sum_{\ell=1}^{t-1} \phi\left(R_{\ell}, A_{\ell}, X_{\ell}\right)\right)^{\top} \eta\left(\theta_{s}\right)-\left(m_{s, 1}+\sum_{\ell=1}^{t-1} \psi\left(A_{\ell}, X_{\ell}\right)\right)^{\top} g\left(\theta_{s}\right)\right]
\end{align*}
$$

The general form for an exponential family likelihood is still retained. The prior-to-posterior conversion simply involves updating the prior parameters with sufficient statistics from the data. Specifically, updated parameters $\phi_{s, t} \leftarrow \phi_{s, 1}+\sum_{\ell=1}^{t-1} \phi\left(R_{\ell}, A_{\ell}, X_{\ell}\right)$ and $m_{s, t} \leftarrow m_{s, 1}+\sum_{\ell=1}^{t-1} \psi\left(A_{\ell}, X_{\ell}\right)$ form the conditional posterior $P_{t}\left(\theta_{s}\right)=h\left(\phi_{s, t}, m_{s, t}\right) \exp \left[\phi_{s, t}^{\top} \eta\left(\theta_{s}\right)-m_{s, t}^{\top} g\left(\theta_{s}\right)\right]$. Finally, the marginal posterior of $s$ is given by,

$$
\begin{align*}
P_{t}(s) \propto P_{1}(s) P\left(H_{t} \mid s\right) & \propto P_{1}(s) \int_{\theta_{s}} P_{1}\left(\theta_{s}\right) P\left(H_{t} \mid \theta_{s}\right) d \theta_{s} \\
& \propto P_{1}(s) \int_{\theta_{s}} h\left(\phi_{s, 1}, m_{s, 1}\right) \exp \left[\phi_{s, t}^{\top} \eta\left(\theta_{s}\right)-m_{s, t} g\left(\theta_{s}\right)\right] d \theta_{s}  \tag{9}\\
& \propto P_{1}(s) \frac{h\left(\phi_{s, 1}, m_{s, 1}\right)}{h\left(\phi_{s, t}, m_{s, t}\right)} .
\end{align*}
$$

The ratio on the right can be easily computed by tracking the posterior parameters. In practice, the marginal posterior can also be updated incrementally as

$$
\begin{align*}
P_{t}(s) & \propto P\left(s \mid H_{t-1}\right) P\left(R_{t-1} \mid A_{t-1}, X_{t-1}, s\right) \\
& \propto P\left(s \mid H_{t-1}\right) \int_{\theta_{s}} P\left(\theta_{s} \mid H_{t-1}\right) P\left(R_{t-1} \mid A_{t-1}, X_{t-1}, s ; \theta\right) d \theta_{s}  \tag{10}\\
& \propto P_{t-1}(s) \frac{h\left(\phi_{s, t-1}, m_{s, t-1}\right)}{h\left(\phi_{s, t}, m_{s, t}\right)}
\end{align*}
$$

where the second term is efffectively the posterior predictive of the observations of round $t-1$.
Thus, for all states $s$, and parameters $\theta$, the posterior probabilities $P_{t}(s)$ and $P_{t}\left(\theta_{s}\right)$ have analytic, closed-form solutions. Thus, sampling from the joint posterior can be done tractably by sampling state $s$ from its marginal posterior, then conditional parameters $\theta_{s}$ from their posterior.

## B Pseudocode of mmTS for Gaussians

Next, we provide specific variants of mmTS when both the model prior and conditional reward likelihood are Gaussian. This is a common assumption for Thompson sampling algorithms [3, 4, 2]. In this case, the joint posterior $P_{t}$ consists of Gaussians. We adopt the notation that $\mathcal{N}\left(r \mid \mu, \sigma^{2}\right) \propto$ $\exp \left[-(r-\mu)^{2} / 2 \sigma^{2}\right]$ is the Gaussian likelihood of $r$ given mean $\mu$ and variance $\sigma^{2}$.
We detail algorithms for two cases: Algorithm 4 is for a multi-armed bandit with independent arms (no context), and Algorithm 5 is for a linear bandit problem. In the first case, we have that $\theta_{s} \in \mathbb{R}^{K}$ are the mean reward vectors where $\mu(a, s ; \theta)=\theta_{s, a}$. In the other case, we assume that context is given by $x \in \mathbb{R}^{K \times d}$ where $x_{a} \in \mathbb{R}^{d}$ is the feature vector for arm $a$. Then, we have that $\theta_{s} \in \mathbb{R}^{d}$ are rank- $d$ vectors such that $\mu(a, x, s ; \theta)=x_{a}^{\top} \theta_{s}$. Both algorithms are efficient to implement, and perform exact sampling from the joint posterior.

```
Algorithm 4 Independent Gaussian mmTS (No Context)
    Input:
        Prior over model parameters \(P_{1}\left(\theta_{s}\right)=\mathcal{N}\left(\bar{\theta}_{s}, \sigma_{0}^{2} I\right), \forall s \in \mathcal{S}\)
        Prior over latent states \(P_{1}(s)\)
    for \(t \leftarrow 1,2, \ldots\) do
        \(\triangleright\) Step 1: sample latent state from marginal posterior.
        Sample \(B_{t} \sim P_{t}\)
        \(\triangleright\) Step 2: sample model parameters from conditional posteriors.
        For each \(s \in \mathcal{S}\), sample \(\widehat{\theta}_{s} \sim \mathcal{N}\left(\tilde{\theta}_{s}, \operatorname{diag}(\tilde{\sigma})\right)\), where
\[
\begin{aligned}
\tilde{\sigma}_{a} & \leftarrow\left(\sigma_{0}^{-2}+\sum_{\ell=1}^{t-1} \mathbb{1}\left\{A_{\ell}=a\right\} \sigma^{-2}\right)^{-1} \\
\tilde{\theta}_{s, a} & \leftarrow \tilde{\sigma}_{a}\left(\sigma_{0}^{-2} \bar{\theta}_{s, a}+\sigma^{-2} \sum_{\ell=1}^{t-1} \mathbb{1}\left\{A_{\ell}=a\right\} R_{\ell}\right)
\end{aligned}
\]
9: \(\quad\) Select \(A_{t} \leftarrow \arg \max _{a \in A} \widehat{\theta}_{B_{t}, a}\)
10: \(\triangleright\) Step 3: update marginal posterior.
11: Observe \(R_{t}\). Update
\[
P_{t}(s) \propto P_{t-1}(s) \mathcal{N}\left(R_{\ell} \mid \tilde{\theta}_{s, A_{\ell}}, \tilde{\sigma}_{A_{\ell}}+\sigma^{2}\right)
\]
```

```
Algorithm 5 Linear Gaussian mmTS
    Input:
        Prior over model parameters \(P_{1}\left(\theta_{s}\right)=\mathcal{N}\left(\bar{\theta}_{s}, \Sigma_{0}\right), \forall s \in \mathcal{S}\)
        Prior over latent states \(P_{1}(s)\)
    for \(t \leftarrow 1,2, \ldots\) do
        \(\triangleright\) Step 1: sample latent state from marginal posterior.
        Sample \(B_{t} \sim P_{t}\)
        \(\triangleright\) Step 2: sample model parameters from conditional posteriors.
        Define
\[
V_{t} \leftarrow I+\sum_{\ell=1}^{t-1} X_{\ell, A_{\ell}} X_{\ell, A_{\ell}}^{\top}, \quad F_{t} \leftarrow \sum_{\ell=1}^{t-1} X_{\ell, A_{\ell}} R_{\ell}
\]
For each \(s \in \mathcal{S}\), compute \(\widehat{\beta} \leftarrow V_{t}^{-1} F_{t}\), and \(\widehat{\Sigma} \leftarrow \sigma^{2} V_{t}^{-1}\)
For each \(s \in \mathcal{S}\), sample \(\widehat{\theta}_{s} \sim \mathcal{N}\left(\tilde{\theta}_{s}, \tilde{\Sigma}\right)\), where
\[
\tilde{\Sigma} \leftarrow\left(\Sigma_{0}^{-1}+(t-1) \widehat{\Sigma}^{-1}\right)^{-1}, \quad \tilde{\theta}_{s} \leftarrow \tilde{\Sigma}\left(\Sigma_{0}^{-1} \bar{\theta}_{s}+(t-1) \widehat{\Sigma}^{-1} \widehat{\beta}\right)
\]
11: \(\quad\) Select \(A_{t} \leftarrow \arg \max _{a \in A} X_{\ell, a}^{\top} \widehat{\theta}_{B_{t}}\)
\(\triangleright\) Step 3: update marginal posterior.
Observe \(R_{t}\). Update
\[
P_{t}(s) \propto P_{t-1}(s) \mathcal{N}\left(R_{\ell} \mid X_{\ell, A_{\ell}}^{\top} \tilde{\theta}_{s}, X_{\ell, A_{\ell}}^{\top} \tilde{\Sigma} X_{\ell, A_{\ell}}+\sigma^{2}\right)
\]
```


## C Proofs

Our proofs rely on the following concentration inequality, which is a straightforward extension of the Azuma-Hoeffding inequality to sub-Gaussian random variables.
Lemma 1. Let $\left(Y_{t}\right)_{t \in[n]}$ be a martingale difference sequence with respect to filtration $\left(\mathcal{F}_{t}\right)_{t \in[n]}$, that is $\mathbb{E}\left[Y_{t} \mid \mathcal{F}_{t-1}\right]=0$ for any $t \in[n]$. Let $Y_{t} \mid \mathcal{F}_{t-1}$ be $\sigma^{2}$-sub-Gaussian for any $t \in[n]$. Then for any $\varepsilon>0$,

$$
\mathbb{P}\left(\left|\sum_{t=1}^{n} Y_{t}\right| \geq \varepsilon\right) \leq 2 \exp \left[-\frac{\varepsilon^{2}}{2 n \sigma^{2}}\right]
$$

Proof. For any $\lambda>0$, which we tune later, we have

$$
\mathbb{P}\left(\sum_{t=1}^{n} Y_{t} \geq \varepsilon\right)=\mathbb{P}\left(\prod_{t=1}^{n} e^{\lambda Y_{t}} \geq e^{\lambda \varepsilon}\right) \leq e^{-\lambda \varepsilon} \mathbb{E}\left[\prod_{t=1}^{n} e^{\lambda Y_{t}}\right]
$$

The inequality is by Markov's inequality. From the conditional independence of $Y_{t}$ given $\mathcal{F}_{t-1}$, the right term becomes

$$
\mathbb{E}\left[\prod_{t=1}^{n} e^{\lambda Y_{t}}\right]=\mathbb{E}\left[\mathbb{E}\left[e^{\lambda Y_{n}} \mid \mathcal{F}_{n-1}\right] \prod_{t=1}^{n-1} e^{\lambda Y_{t}}\right] \leq e^{\frac{\lambda^{2} \sigma^{2}}{2}} \mathbb{E}\left[\prod_{t=1}^{n-1} e^{\lambda Y_{t}}\right] \leq e^{\frac{n \lambda^{2} \sigma^{2}}{2}}
$$

We use that $Y_{n} \mid \mathcal{F}_{n-1}$ is $\sigma^{2}$-sub-Gaussian in the first inequality, and recursively repeat for all rounds in the second. So we have

$$
\mathbb{P}\left(\sum_{t=1}^{n} Y_{t} \geq \varepsilon\right) \leq \min _{\lambda>0} e^{-\lambda \varepsilon+\frac{n \lambda^{2} \sigma^{2}}{2}}
$$

The minimum is achieved at $\lambda=\varepsilon /\left(n \sigma^{2}\right)$. Therefore

$$
\mathbb{P}\left(\sum_{t=1}^{n} Y_{t} \geq \varepsilon\right) \leq \exp \left[-\frac{\varepsilon^{2}}{2 n \sigma^{2}}\right]
$$

Now we apply the same proof to $\mathbb{P}\left(-\sum_{t=1}^{n} Y_{t} \geq \varepsilon\right)$, which yields a multiplicative factor of 2 in the upper bound. This concludes the proof.

## C. 1 Proof of Theorem 1

Recall that $s_{*} \in \mathcal{S}, \theta_{*} \in \Theta$ are the true latent state and model. Let $\mu(a, x)=\mu\left(a, x, s_{*} ; \theta_{*}\right)$ be the true mean rewards given observed context and action. Let

$$
\begin{equation*}
E_{t}=\left\{\forall s \in \mathcal{S}:\left|\sum_{\ell=1}^{t-1} \mathbb{1}\left\{B_{\ell}=s\right\}\left(\mu\left(A_{\ell}, X_{\ell}\right)-R_{\ell}\right)\right| \leq \sigma \sqrt{6 N_{t}(s) \log n}\right\} \tag{11}
\end{equation*}
$$

be the event that the total realized reward under each played latent state is close to its expectation. Let $E=\cap_{t=1}^{n} E_{t}$ be the event that this holds for all rounds, and $\bar{E}$ be its complement. We can rewrite the expected $n$-round regret by

$$
\begin{align*}
\mathcal{R}(n)= & \mathbb{E}[\mathbb{1}\{\bar{E}\} \mathcal{R}(n)]+\mathbb{E}[\mathbb{1}\{E\} \mathcal{R}(n)] \\
\leq & \mathbb{E}\left[\mathbb{1}\{\bar{E}\} \sum_{t=1}^{n} \mu\left(A_{t, *}, X_{t}\right)-\mu\left(A_{t}, X_{t}\right)\right] \\
& +\mathbb{E}\left[\mathbb{1}\{E\} \sum_{t=1}^{n}\left(\mu\left(A_{t, *}, X_{t}\right)-U_{t}\left(A_{t, *}\right)\right)\right]+\mathbb{E}\left[\mathbb{1}\{E\} \sum_{t=1}^{n}\left(U_{t}\left(A_{t}\right)-\mu\left(A_{t}, X_{t}\right)\right)\right], \tag{12}
\end{align*}
$$

where we use the regret decomposition in Eq. (5) in the inequality.
Our first lemma is that the probability of $\bar{E}$ occurring is low. Without context, the lemma would follow immediately from Hoeffding's inequality. Since we have context generated by some random process, we instead turn to martingales.

Lemma 2. Let $E_{t}$ be defined as in $E q$. (11) for all rounds $t, E=\cap_{t=1}^{n} E_{t}$, and $\bar{E}$ be its complement. Then $\mathbb{P}(\bar{E}) \leq 2|\mathcal{S}| n^{-1}$.

Proof. We see that the choice of action given observed context depends on past rounds. This is because the upper confidence bounds depend on which latent states are eliminated, which depend on the history of observed contexts.
For each latent state $s$ and round $t$, let $\mathcal{T}_{t, s}$ be the rounds where state $s$ was chosen until round $t$. For round $\ell \in \mathcal{T}_{t, s}$, let $Y_{\ell}(s)=\mu\left(A_{\ell}, X_{\ell}, s_{\ell} ; \theta_{*}\right)-R_{\ell}$. Observe that $Y_{\ell}(s) \mid X_{\ell}, H_{\ell}$ is $\sigma^{2}$-sub-Gaussian. This implies that $\left(Y_{\ell}(s)\right)_{\ell \in \mathcal{T}_{t, s}}$ is a martingale difference sequence with respect to context and history $\left(X_{\ell}, H_{\ell}\right)_{\ell \in \mathcal{T}_{t, s}}$, or $\mathbb{E}\left[Y_{\ell}(s) \mid X_{\ell}, H_{\ell}\right]=0$ for all rounds $\ell \in \mathcal{T}_{t, s}$.
For any round $t$, and state $s \in \mathcal{S}$, we have that $\mathcal{T}_{t, s}$ is a random quantity. First, we fix $\left|\mathcal{T}_{t, s}\right|=$ $N_{t}(s)=u$ where $u<t$ and yield the following due to Lemma 1 ,

$$
\mathbb{P}\left(\left|\sum_{\ell \in \mathcal{T}_{t, s}} Y_{\ell}(s)\right| \geq \sigma \sqrt{6 u \log n}\right) \leq 2 \exp [-3 \log n]=2 n^{-3}
$$

So, by the union bound, we have

$$
\mathbb{P}(\bar{E}) \leq \sum_{t=1}^{n} \sum_{s \in \mathcal{S}} \sum_{u=1}^{t-1} \mathbb{P}\left(\left|\sum_{\ell \in \mathcal{T}_{t, s}} Y_{\ell}(s)\right| \geq \sigma \sqrt{6 u \log n}\right) \leq 2|\mathcal{S}| n^{-1}
$$

The first term in Eq. (12) is small because the probability of $\bar{E}$ is small. Using Lemma 2, and that total regret is bounded by $n$, we have, $\mathbb{E}[\mathbb{1}\{\bar{E}\} \mathcal{R}(n)] \leq n \mathbb{P}(\bar{E}) \leq 2|\mathcal{S}|$.
For round $t$, the event $\mu\left(A_{t, *}, X_{t}\right) \geq U_{t}\left(A_{t, *}\right)$ occurs only if $s_{*} \notin C_{t}$ also occurs. By the design of $C_{t}$ in mUCB, this happens only if $G_{t}\left(s_{*}\right)>\sigma \sqrt{6 N_{t}(s) \log n}$. Event $E_{t}$ says that the opposite is true for all states, including true state $s_{*}$. So, the second term in Eq. (12) is at most 0 .
Now, consider the last term in Eq. (12). Let $T_{s}=\left\{t \leq n: B_{t}=s\right\}$ denote the set of rounds where latent state $s$ is selected. We have,

$$
\begin{aligned}
\mathbb{E}\left[\mathbb{1}\{E\} \sum_{t=1}^{n}\left(U_{t}\left(A_{t}\right)-\mu\left(A_{t}, X_{t}\right)\right)\right] & =\mathbb{E}\left[\mathbb{1}\{E\} \sum_{s \in S} \sum_{t \in T_{s}}\left(\mu\left(A_{t}, X_{t}, s\right)-\mu\left(A_{t}, X_{t}\right)\right)\right] \\
& =\mathbb{E}\left[\mathbb{1}\{E\} \sum_{s \in S} \sum_{t \in T_{s}}\left(\mu\left(A_{t}, X_{t}, s\right)-R_{t}+R_{t}-\mu\left(A_{t}, X_{t}\right)\right)\right] \\
& \leq \mathbb{E}\left[\mathbb{1}\{E\} \sum_{s \in S}\left(G_{n}(s)+\sum_{t \in T_{s}}\left(R_{t}-\mu\left(A_{t}, X_{t}\right)\right)\right)\right] \\
& \leq \sum_{s \in S}\left(1+2 \sigma \sqrt{6 N_{n}(s) \log n}\right) .
\end{aligned}
$$

For the first inequality, we use that the last round $t^{\prime}$ where state $s$ is selected, we have an upperbound on the prior gap $G_{t^{\prime}}(s) \leq \sigma \sqrt{6 N_{t^{\prime}}}(s) \log n$. Accounting for the last round yields an upper-bound of $\sigma \sqrt{6 N_{n}(s) \log n}+1$. For the last inequality, we use $E$ occurring to bound $\sum_{t \in T_{s}}\left(R_{t}-\mu\left(A_{t}, X_{t}\right)\right) \leq \sigma \sqrt{6 N_{n}(s) \log n}$.
This yields the desired bound on total regret,

$$
\begin{aligned}
\mathcal{R}(n) & \leq 3|\mathcal{S}|+2 \sigma \sqrt{6 \log n}\left(\sum_{s \in S} \sqrt{N_{n}(s)}\right) \\
& \leq 3|\mathcal{S}|+2 \sigma \sqrt{6|\mathcal{S}| \log n \sum_{s \in S} N_{n}(s)}=3|\mathcal{S}|+2 \sigma \sqrt{6|\mathcal{S}| n \log n}
\end{aligned}
$$

where the last inequality comes from the Cauchy-Schwarz inequality.

## C. 2 Proof of Corollary 1

The true latent state $S_{*} \in \mathcal{S}$ is random under Bayes regret. In this case, we still assume that we are given the true model $\theta_{*}$, so only $S_{*} \sim P_{1}$ for known $P_{1}$. We also have that the optimal action $A_{t, *}=\arg \max _{a \in \mathcal{A}} \mu\left(a, X_{t}, S_{*} ; \theta_{*}\right)$ is random not only due to context, but also $S_{*}$.

We define $U_{t}(a)=\arg \max _{s \in C_{t}} \mu\left(a, X_{t}, S_{*} ; \theta_{*}\right)$ as in mUCB. Note the additional randomness due to $S_{*}$. We can rewrite the Bayes regret as $\mathcal{B R}(n)=\mathbb{E}\left[\mathcal{R}\left(n ; S_{*}, \theta_{*}\right)\right]$, where the outer expectation is over $S_{*} \sim P_{1}$. The expression inside the expectation can be decomposed as

$$
\begin{aligned}
& \mathcal{R}\left(n, S_{*}, \theta_{*}\right)=\mathbb{E}\left[\mathbb{1}\{\bar{E}\} \sum_{t=1}^{n} \mu\left(A_{t, *}, X_{t}, S_{*}\right)-\mu\left(A_{t}, X_{t}, S_{*}\right)\right] \\
& \quad+\mathbb{E}\left[\mathbb{1}\{E\} \sum_{t=1}^{n}\left(\mu\left(A_{t, *}, X_{t}, S_{*}\right)-U_{t}\left(A_{t, *}\right)\right)\right]+\mathbb{E}\left[\mathbb{1}\{E\} \sum_{t=1}^{n}\left(U_{t}\left(A_{t}\right)-\mu\left(A_{t}, X_{t}, S_{*}\right)\right)\right]
\end{aligned}
$$

where $E, \bar{E}$ are defined as in Appendix C.1, and we use the decomposition in Eq. (6).
Each above expression can be bounded exactly as in Theorem 1 . The reason is that the original upper bounds hold for any $S_{*}$, and therefore also in expectation over $S_{*} \sim P_{1}$. This yields the desired Bayes regret bound.

## C. 3 Proof of Theorem 2

The only difference in the analysis is that we need to incorporate the additional error due to model misspecification.
Let $\mathcal{E}=\{\forall a \in \mathcal{A}, x \in \mathcal{X}, s \in \mathcal{S}:|\widehat{\mu}(a, x, s)-\mu(a, x, s)| \leq \varepsilon\}$ be the event that model $\widehat{\theta}$ has bounded misspecification and $\overline{\mathcal{E}}$ be its complement. Also let $E, \bar{E}$ be defined as in Appendix C.1.
If $\mathcal{E}$ does not hold, then the best possible upper-bound on regret is $n$; fortunately, we assume in the theorem that the probability of that occurring is bounded by $\delta$. So we can bound the expected $n$-round regret as

$$
\begin{align*}
\mathcal{R}(n)= & \mathbb{E}[\mathbb{1}\{\overline{\mathcal{E}}\} \mathcal{R}(n)]+\mathbb{E}[\mathbb{1}\{\bar{E}, \mathcal{E}\} \mathcal{R}(n)]+\mathbb{E}[\mathbb{1}\{E, \mathcal{E}\} \mathcal{R}(n)] \\
\leq & \delta n+\mathbb{E}\left[\mathbb{1}\{\bar{E}, \mathcal{E}\} \sum_{t=1}^{n} \mu\left(A_{t, *}, X_{t}\right)-\mu\left(A_{t}, X_{t}\right)\right]  \tag{13}\\
& +\mathbb{E}\left[\mathbb{1}\{E, \mathcal{E}\} \sum_{t=1}^{n}\left(\mu\left(A_{t, *}, X_{t}\right)-U_{t}\left(A_{t, *}\right)\right)\right]+\mathbb{E}\left[\mathbb{1}\{E, \mathcal{E}\} \sum_{t=1}^{n}\left(U_{t}\left(A_{t}\right)-\mu\left(A_{t}, X_{t}\right)\right)\right]
\end{align*}
$$

where we use the regret decomposition in Eq. (5).
The second term in Eq. (13) is small because the probability of $\bar{E}$ is small. Using Lemma 2, and that total regret is bounded by $n$, we have, $\mathbb{E}[\mathbb{1}\{\bar{E}, \mathcal{E}\} \mathcal{R}(n)] \leq n \mathbb{P}(\bar{E}) \leq 2|\mathcal{S}|$.

If $\mathcal{E}$ occurs, the event $\mu\left(A_{t, *}, X_{t}\right)-U_{t}\left(A_{t, *}\right)>\varepsilon$ for any round $t$ occurs only if $s_{*} \notin C_{t}$ also occurs. By the design of $C_{t}$ in mmUCB, this happens if $G_{t}\left(s_{*}\right) \geq \sigma \sqrt{6 N_{t}(s) \log n}$. Since

$$
G_{t}\left(s_{*}\right)=\sum_{\ell=1}^{t-1} \mathbb{1}\left\{B_{\ell}=s_{*}\right\}\left(\widehat{\mu}\left(A_{\ell}, X_{\ell}\right)-\varepsilon-R_{\ell}\right) \leq \sum_{\ell=1}^{t-1} \mathbb{1}\left\{B_{\ell}=s_{*}\right\}\left(\mu\left(A_{\ell}, X_{\ell}\right)-R_{\ell}\right),
$$

we see that event $E_{t}$ says that the opposite is true for all states, including true state $s_{*}$. Hence, the third term in Eq. (13) is bounded by $\varepsilon n$.

Now, consider the last term in Eq. (13). Let $T_{s}=\left\{t \leq n: B_{t}=s\right\}$ denote the set of rounds where latent state $s$ is selected. We have,

$$
\begin{aligned}
& \mathbb{E}[\mathbb{1}\left.\{E, \mathcal{E}\} \sum_{t=1}^{n}\left(U_{t}\left(A_{t}\right)-\mu\left(A_{t}, X_{t}\right)\right)\right] \\
&=\mathbb{E}\left[\mathbb{1}\{E, \mathcal{E}\} \sum_{s \in S}\left(\sum_{t \in T_{s}} \widehat{\mu}\left(A_{t}, X_{t}, s\right)-\mu\left(A_{t}, X_{t}\right)\right)\right] \\
&=\varepsilon n+\mathbb{E}\left[\mathbb{1}\{E, \mathcal{E}\} \sum_{s \in S} \sum_{t \in T_{s}}\left(\widehat{\mu}\left(A_{t}, X_{t}, s\right)-\varepsilon-R_{t}+R_{t}-\mu\left(A_{t}, X_{t}\right)\right)\right] \\
& \quad \leq \varepsilon n+\mathbb{E}\left[\mathbb{1}\{E, \mathcal{E}\} \sum_{s \in S}\left(G_{n}(s)+\sum_{t \in T_{s}}\left(R_{t}-\mu\left(A_{t}, X_{t}\right)\right)\right)\right] \\
& \quad \leq \varepsilon n+\sum_{s \in S}\left(1+2 \sigma \sqrt{6 N_{n}(s) \log n}\right) .
\end{aligned}
$$

For the first inequality, we use that the last round $t^{\prime}$ where state $s$ is selected, we have an upperbound on the prior gap $G_{t^{\prime}}(s) \leq \sigma \sqrt{6 N_{t^{\prime}}(s) \log n}$. Accounting for the last round yields and upper-bound of $\sigma \sqrt{6 N_{n}(s) \log n}+1$. For the last inequality, we use $E$ occurring to bound $\sum_{t \in T_{s}}\left(R_{t}-\mu\left(A_{t}, X_{t}\right)\right) \leq \sigma \sqrt{6 N_{n}(s) \log n}$.
This yields the desired bound on total regret,

$$
\begin{aligned}
\mathcal{R}(n) & \leq \delta n+3|\mathcal{S}|+2 \varepsilon n+2 \sigma \sqrt{6 \log n}\left(\sum_{s \in S} \sqrt{N_{n}(s)}\right) \\
& \leq \delta n+3|\mathcal{S}|+2 \varepsilon n+2 \sigma \sqrt{6|\mathcal{S}| \log n \sum_{s \in S} N_{n}(s)}=\delta n+3|\mathcal{S}|+2 \varepsilon n+2 \sigma \sqrt{6|\mathcal{S}| n \log n}
\end{aligned}
$$

where the last inequality comes from the Cauchy-Schwarz inequality.

## C. 4 Proof of Corollary 2

Both latent state $S_{*} \in \mathcal{S}$ and model $\theta_{*} \in \Theta$ are random, and drawn as $S_{*}, \theta_{*} \sim P_{1}$, where the prior $P_{1}$ is known. In this case, the true model $\theta_{*}$ is not known to us.
Using marginalized means $\bar{\mu}(a, x, s)$, and $\varepsilon, \delta>0$ as defined in the statement of the corollary, we write,

$$
G_{t}(s)=\sum_{\ell=1}^{t-1} \mathbb{1}\left\{B_{\ell}=s\right\}\left(\bar{\mu}\left(A_{\ell}, X_{\ell}, s\right)-\varepsilon-R_{\ell}\right)
$$

and $U_{t}(a)=\arg \max _{s \in C_{t}} \bar{\mu}\left(a, X_{t}, s\right)$. This is in contrast to $G_{t}(s)$ and $U_{t}(a)$ in mmUCB, which use $\widehat{\mu}(a, x, s)$ from a single model. Conceptually though, both $\widehat{\mu}(a, x, s)$ and $\bar{\mu}(a, x, s)$ are just $\varepsilon$-close point estimates of $\mu(a, x, s)$ due to the assumptions made about the true model $\theta_{*}$ in the theorem and corollary, respectively.
We can rewrite the Bayes regret as $\mathcal{B} \mathcal{R}(n)=\mathbb{E}\left[\mathcal{R}\left(n ; S_{*}, \theta_{*}\right)\right]$, where the outer expectation is over $S_{*}, \theta_{*} \sim P_{1}$. The expression inside the expectation can be written as,
$\mathcal{R}\left(n ; S_{*}, \theta_{*}\right) \leq \delta n+\mathbb{E}\left[\mathbb{1}\{\bar{E}, \mathcal{E}\} \sum_{t=1}^{n} \mu\left(A_{t, *}, X_{t}, S_{*} ; \theta_{*}\right)-\mu\left(A_{t}, X_{t}, S_{*} ; \theta_{*}\right)\right]$
$+\mathbb{E}\left[\mathbb{1}\{E, \mathcal{E}\} \sum_{t=1}^{n}\left(\mu\left(A_{t, *}, X_{t}, S_{*} ; \theta_{*}\right)-U_{t}\left(A_{t, *}\right)\right)\right]+\mathbb{E}\left[\mathbb{1}\{E, \mathcal{E}\} \sum_{t=1}^{n}\left(U_{t}\left(A_{t}\right)-\mu\left(A_{t}, X_{t}, S_{*} ; \theta_{*}\right)\right)\right]$,
where $\mathcal{E}, E, \bar{E}$ are defined as in Appendix C.3, and we use the decomposition in Eq. (6).
The expressions can be bounded exactly as in Theorem 2. The upper bound is worst-case and holds for any $S_{*}, \theta_{*}$, and thus also holds after taking an expectation over the prior $S_{*}, \theta_{*} \sim P_{1}$. This bounds the Bayes regret, as desired.

