Appendix

A PRELIMINARIES

A.1 Transportation Inequalities

For any function $f : \mathcal{X} \to \mathbb{R}$, we define its span as $\mathcal{S}(f) := \max_{x \in \mathcal{X}} f(x) - \min_{x \in \mathcal{X}} f(x)$. For a probability distribution $P$ supported on the set $\mathcal{X}$, let $\mathbb{E}_P[f] := \mathbb{E}_P[f(X)]$ and $\mathbb{V}_P[f] := \mathbb{V}_P[f(X)] = \mathbb{E}_P[f(X)^2] - \mathbb{E}_P[f(X)]^2$ denote the mean and variance of the random variable $f(X)$, respectively. We now state the following transportation inequalities, which can be adapted from [Boucheron et al. 2013] Lemma 4.18.

**Lemma 1** (Transportation inequalities). Assume $f$ is such that $\mathcal{S}(f)$ and $\mathbb{V}_P[f]$ are finite. Then it holds

\[
\forall Q \ll P, \quad \mathbb{E}_Q[f] - \mathbb{E}_P[f] \leq \sqrt{2 \mathbb{V}_P[f] \text{KL}(Q, P)} + \frac{2 \mathcal{S}(f)}{3} \text{KL}(Q, P),
\]

\[
\forall Q \ll P, \quad \mathbb{E}_P[f] - \mathbb{E}_Q[f] \leq \sqrt{2 \mathbb{V}_P[f] \text{KL}(Q, P)}.
\]

A.2 Bregman Divergence

For a Legendre function $F : \mathbb{R}^d \to \mathbb{R}$, the Bregman divergence between $\theta', \theta \in \mathbb{R}^d$ associated with $F$ is defined as

\[
B_{F}(\theta', \theta) := F(\theta') - F(\theta) - (\theta' - \theta)^\top \nabla F(\theta).
\]

Now, for any fixed $\theta \in \mathbb{R}^d$, we introduce the function

\[
B_{F,\theta}(\lambda) := B_{F}(\theta + \lambda) = F(\theta + \lambda) - F(\theta) - \lambda^\top \nabla F(\theta).
\]

It then follows that $B_{F,\theta}$ is a convex function, and we define its dual as

\[
B_{\star F,\theta}(x) = \sup_{\lambda \in \mathbb{R}^d} \left( \lambda^\top x - B_{F,\theta}(\lambda) \right).
\]

We have for any $\theta, \theta' \in \mathbb{R}^d$:

\[
B_{F}(\theta', \theta) = B_{\star F,\theta}(\nabla F(\theta) - \nabla F(\theta')).
\]

To see this, we observe that

\[
B_{\star F,\theta'}(\nabla F(\theta) - \nabla F(\theta'))
= \sup_{\lambda \in \mathbb{R}^d} \lambda^\top (\nabla F(\theta) - \nabla F(\theta')) - [F(\theta' + \lambda) - F(\theta') - \lambda^\top \nabla F(\theta')]
= \sup_{\lambda \in \mathbb{R}^d} \lambda^\top \nabla F(\theta) - F(\theta' + \lambda) + F(\theta').
\]

Now an optimal $\lambda$ must satisfy $\nabla F(\theta) = \nabla F(\theta' + \lambda)$. One possible choice is $\lambda = \theta - \theta'$. Since, by definition, $F$ is strictly convex, the supremum will indeed be attained at $\lambda = \theta - \theta'$. Plugin-in this value, we obtain

\[
B_{\star F,\theta'}(\nabla F(\theta) - \nabla F(\theta')) = (\theta - \theta')^\top \nabla F(\theta) - F(\theta) + F(\theta') = B_{F}(\theta', \theta).
\]

(See that [1] holds for any convex function $F$. Only difference is that, in this case, $B_{F}(\cdot, \cdot)$ won’t correspond to the Bregman divergence.)

A.3 Exponential Family

In this section, we detail some useful results related to exponential families in our model.

**Derivatives** Let us first take a closer look at the derivative of the log-partition function $Z_{s,a}$. As usual with exponential families, these are intimately linked to moments of the random variable. We have on the one hand,

\[
(\nabla_i Z_{s,a})(\theta) = \int_S \psi(s')^\top A_i \varphi(s, a) \frac{h(s', s, a) \exp \left( \sum_{i=1}^d \theta_i \psi(s')^\top A_i \varphi(s, a) \right)}{\int_S h(s', s, a) \exp \left( \sum_{i=1}^d \theta_i \psi(s')^\top A_i \varphi(s, a) \right) \, ds'} \, ds'
\]

\[
= \mathbb{E}_{s,a}^{\theta} \left[ \psi(s')^\top A_i \varphi(s, a) \right].
\]
On the other hand, the entries of the Hessian of $Z$ are given by

$$
\begin{align*}
(\nabla^2_{i,j} Z_{s,a})(\theta) &= \int_s \psi(s')^T A_i \varphi(s,a) \psi(s')^T A_j \varphi(s,a) \frac{h(s',s,a) \exp\left(\sum_{i=1}^d \theta_i \psi(s')^T A_i \varphi(s,a)\right)}{\int_s h(s',s,a) \exp\left(\sum_{i=1}^d \theta_i \psi(s')^T A_i \varphi(s,a)\right) ds'} ds' \\
&- \int_s \psi(s')^T A_i \varphi(s,a) \frac{h(s',s,a) \exp\left(\sum_{i=1}^d \theta_i \psi(s')^T A_i \varphi(s,a)\right)}{\int_s h(s',s,a) \exp\left(\sum_{i=1}^d \theta_i \psi(s')^T A_i \varphi(s,a)\right) ds'} ds' \nabla^2_{i,j} Z_{s,a}(\theta)
\end{align*}
$$

$$
\begin{align*}
&= \mathbb{E}^\theta_{s,a} \left[ \psi(s')^T A_i \varphi(s,a) \psi(s')^T A_j \varphi(s,a) \right] \\
&- \mathbb{E}^\theta_{s,a} \left[ \psi(s')^T A_i \varphi(s,a) \right] \mathbb{E}^\theta_{s,a} \left[ \psi(s')^T A_j \varphi(s,a) \right] \\
&= \varphi(s,a)^T A_i^T \mathbb{C}^\theta_{s,a} \left[ \psi(s') \psi(s')^T \right] \mathbb{E}^\theta_{s,a} \left[ \psi(s')^T \right] A_j \varphi(s,a) \\
&= \varphi(s,a)^T A_i^T \mathbb{C}^\theta_{s,a} \left[ \psi(s') \psi(s')^T \right] \mathbb{E}^\theta_{s,a} \left[ \psi(s')^T \right] A_j \varphi(s,a)
\end{align*}
$$

where we introduce in the last line the $p \times p$ covariance matrix given by

$$
\mathbb{C}^\theta_{s,a} \left[ \psi(s') \right] = \mathbb{E}^\theta_{s,a} \left[ \psi(s') \psi(s')^T \right] - \mathbb{E}^\theta_{s,a} \left[ \psi(s') \right] \mathbb{E}^\theta_{s,a} \left[ \psi(s')^T \right]
$$

**KL Divergence** For any two $\theta, \theta'$ and for some pair $(s,a)$, we are interested in the following useful relations

$$
\begin{align*}
\log \left( \frac{P_{\theta}(s'|s,a)}{P_{\theta'}(s'|s,a)} \right) &= \sum_{i=1}^d (\theta_i - \theta'_i) \psi(s')^T A_i \varphi(s,a) - Z_{s,a}(\theta) + Z_{s,a}(\theta') \\
or \text{KL} \left( P_{\theta}(\cdot|s,a), P_{\theta'}(\cdot|s,a) \right) &= \sum_{i=1}^d (\theta_i - \theta'_i) \mathbb{E}^\theta_{s,a} \left[ \psi(s')^T \right] A_i \varphi(s,a) - Z_{s,a}(\theta) + Z_{s,a}(\theta') \\
&= \frac{1}{2} (\theta - \theta')^T (\nabla^2 Z_{s,a})(\tilde{\theta})(\theta - \theta')
\end{align*}
$$

where in the last line, we used, by a Taylor expansion, that $Z_{s,a}(\theta') = Z_{s,a}(\theta) + (\nabla Z_{s,a}(\theta))^T (\theta' - \theta) + \frac{1}{2}(\theta - \theta')^T (\nabla^2 Z_{s,a}(\tilde{\theta}))(\theta - \theta')$ for some $\tilde{\theta} \in [\theta, \theta']_\infty$. Here $[\theta, \theta']_\infty$ denotes the $d$-dimensional hypercube joining $\theta$ to $\theta'$.

**B METHOD OF MIXTURES FOR CONDITIONAL EXPONENTIAL FAMILIES: PROOF OF THEOREM [1]**

**Step 1: Martingale Construction** First note that by our hypothesis of strict convexity, the log-partition function $Z_{s,a}$ is a Legendre function[1]. Now for the conditional exponential family model, the KL divergence b/w $P_{\theta}(\cdot|s,a)$ and $P_{\theta'}(\cdot|s,a)$ can be expressed as a Bregman divergence associated to $Z_{s,a}$ with the parameters reversed, i.e.,

$$
\text{KL}_{s,a}(\theta, \theta') := \text{KL} \left( P_{\theta}(\cdot|s,a), P_{\theta'}(\cdot|s,a) \right) = B_{Z_{s,a}}(\theta', \theta) \ .
$$

(5)

Now, for any $\lambda \in \mathbb{R}^d$, we introduce the function $B_{Z_{s,a}, \theta^*}(\lambda) = B_{Z_{s,a}}(\theta^* + \lambda, \lambda)$ and define

$$
M^\lambda_n = \exp \left( \lambda^T S_n - \sum_{t=1}^n B_{Z_{s,t}, \theta^*}(\lambda) \right)
$$

Since we will only use [4] in the proof, the final result would hold even if $Z_{s,a}$ is only convex.
where \( \forall i \leq d \), we denote
\[
(S_n)_i = \sum_{t=1}^n \left( \psi(s'_t) - \mathbb{E}_{s',a_t}^\theta [\psi(s') \mid a_t] \right)^\top A_i \varphi(s_t, a_t).
\]
Note that \( M_n^\lambda > 0 \) and it is \( \mathcal{F}_n \)-measurable. Furthermore, we have for all \((s, a)\),
\[
\mathbb{E}_{s,a}^{\theta^*} \left[ \exp \left( \sum_{i=1}^d \lambda_i \left( \psi(s') - \mathbb{E}_{s,a}^{\theta^*} [\psi(s') \mid a_t] \right)^\top A_i \varphi(s, a) \right) \right]
\]
\[
= \exp \left( -\lambda^\top \nabla Z_{s,a}(\theta^*) \right) \int_S h(s', s, a) \exp \left( \sum_{i=1}^d (\lambda_i + \lambda) \psi(s')^\top A_i \varphi(s, a) - Z_{s,a}(\theta^*) \right) ds'
\]
\[
= \exp \left( Z_{s,a}(\theta^* + \lambda) - Z_{s,a}(\theta^*) - \lambda^\top \nabla Z_{s,a}(\theta^*) \right) = \exp \left( B_{Z_{s,a}}(\theta^*) \right).
\]
This implies \( \mathbb{E} \left[ \exp \left( \lambda^\top S_n \right) \mathcal{F}_{n-1} \right] = \exp \left( \lambda^\top S_n - B_{Z_{s,a}}(\theta^*) \right) \) and thus, in turn, \( \mathbb{E}[M_n^\lambda \mid \mathcal{F}_{n-1}] = M_{n-1}^\lambda \). Therefore \( \{M_n^\lambda \}_{n=0}^\infty \) is a non-negative martingale adapted to the filtration \( \{\mathcal{F}_n\}_{n=0}^\infty \) and actually satisfies \( \mathbb{E} \left[ M_n^\lambda \right] = 1 \). For any prior density \( q(\theta) \) for \( \theta \), we now define a mixture of martingales
\[
M_n = \int_{\mathbb{R}^d} M_n^\lambda q(\theta^* + \lambda) d\lambda.
\]
Then \( \{M_n\}_{n=0}^\infty \) is also a non-negative martingale adapted to \( \{\mathcal{F}_n\}_{n=0}^\infty \) and in fact, \( \mathbb{E} \left[ M_n \right] = 1 \).

**Step 2:** Method Of Mixtures and Martingale Control

Considering the prior density \( \mathcal{N} \left( 0, (\eta A)^{-1} \right) \), we obtain from (6) that
\[
M_n = c_0 \int_{\mathbb{R}^d} \exp \left( \lambda^\top S_n - \sum_{i=1}^n B_{Z_{s,i},a_i}(\theta^* - \eta \lambda^2) + \eta \lambda \theta - \eta \lambda^2 \right) d\lambda,
\]
where \( c_0 = \int_{\mathbb{R}^d} \exp \left( -\frac{1}{2} \|\theta^2\|_A^2 \right) d\theta \). We now introduce the function
\[
Z_n(\theta) = \sum_{i=1}^n Z_{s_i,a_i}(\theta).
\]
Note that \( Z_n(\theta) \) is a also Legendre function and its associated Bregman divergence satisfies
\[
B_{Z_n}(\theta, \theta') = \sum_{i=1}^n \left( Z_{s_i,a_i}(\theta' - Z_{s_i,a_i}(\theta) - (\theta' - \theta)^\top \nabla Z_{s_i,a_i}(\theta)) \right) = \sum_{i=1}^n B_{Z_{s_i,a_i}}(\theta', \theta).
\]
Furthermore, we have \( \sum_{i=1}^n B_{Z_{s_i,a_i}}(\theta^* - \theta) = B_{Z_n}(\theta^* - \theta) \).

From the penalized likelihood formula (2), recall that
\[
\forall i \leq d, \quad \sum_{t=1}^n \nabla_i Z_{s_i,a_i}(\theta_n) + \frac{\eta}{2} \nabla_i \|\theta_n\|_A^2 = \sum_{i=1}^n \psi(s'_t)^\top A_i \varphi(s_t, a_t).
\]
This yields
\[
S_n = \sum_{i=1}^n (\nabla Z_{s_i,a_i}(\theta_n) - \nabla Z_{s_i,a_i}(\theta^*)) + \eta \lambda \theta_n = \nabla Z_n(\theta_n) - \nabla Z_n(\theta^*) + \eta \lambda \theta_n.
\]
We now obtain from (7) and (8) that
\[
M_n = c_0 \cdot \exp \left( -\frac{\eta}{2} \|\theta^2\|_A^2 \right) \int_{\mathbb{R}^d} \exp \left( \lambda^\top x_n - B_{Z_n}(\lambda) + g_n(\lambda) \right) d\lambda,
\]
where we have introduced \( g_n(\lambda) = \frac{\eta}{2} \left( 2\lambda^\top A \lambda_n + \|\theta^2\|_A^2 - \|\theta^* + \lambda\|_A^2 \right) \) and \( x_n = \nabla Z_n(\theta_n) - \nabla Z_n(\theta^*) \).

Now, note that \( \sup_{\lambda \in \mathbb{R}^d} g_n(\lambda) \leq \frac{\eta}{2} \|\theta^2\|_A^2 \), where the supremum is attained at \( \lambda^* = \theta_n - \theta^* \). We then have
\[
g_n(\lambda) = g_n(\lambda) + \sup_{\lambda \in \mathbb{R}^d} g_n(\lambda) - g_n(\lambda^*)
\]
\[
= \frac{\eta}{2} \|\theta_n - \theta^*\|_A^2 + \eta \lambda^2 (\lambda^* + \lambda\lambda^*) + \frac{\eta}{2} \|\theta^* + \lambda\|_A^2 - \frac{\eta}{2} \|\theta^* + \lambda\|_A^2
\]
\[
= B_{Z_n}(\lambda^*, \theta_n) + (\lambda^* + \lambda^\top \nabla Z_0(\lambda^* + \lambda)) + Z_0(\theta^* + \lambda) - Z_0(\theta^* + \lambda),
\]
where we have introduced the Legendre function \( Z_0(\theta) = \frac{\eta}{2} \|\theta\|_A^2 \). We now have from (4) that
\[
\sup_{\lambda \in \mathbb{R}^d} \left( \lambda^\top x_n - B_{Z_n}(\lambda) \right)
\]
\[
= B_{Z_n}(\lambda^*, x_n) = B_{Z_n}(\lambda^* - \nabla Z_n(\theta_n) - \nabla Z_n(\theta^*)) = B_{Z_n}(\theta^*, \theta_n).
\]
Further, any optimal \( \lambda \) must satisfy
\[
\nabla Z_n(\theta^* + \lambda) = \nabla Z_n(\theta^*) = x_n \implies \nabla Z_n(\theta^* + \lambda) = \nabla Z_n(\theta_n).
\]
One possible solution is $\lambda = \lambda^*$. Now, since $Z_n$ is strictly convex, the supremum is indeed attained at $\lambda = \lambda^*$. We then have
\[
\lambda^\top x_n - B_{Z_n, \theta^*}(\lambda)
= \lambda^\top x_n - B_{Z_n, \theta^*}(\lambda) + B_{Z_n}(\theta^*, \theta_n) - (\lambda^\top x_n - B_{Z_n, \theta^*}(\lambda))
= B_{Z_n}(\theta^*, \theta_n) + (\lambda - \lambda^*)^\top \nabla Z_n(\theta^* + \lambda^*) + B_{Z_n, \theta^*}(\lambda^*) - B_{Z_n, \theta^*}(\lambda) - (\lambda - \lambda^*)^\top \nabla Z_n(\theta^*)
= B_{Z_n}(\theta^*, \theta_n) + (\lambda - \lambda^*)^\top \nabla Z_n(\theta^* + \lambda^*) + Z_n(\theta^* + \lambda^*) - Z_n(\theta^* + \lambda).
\]
Plugging \([10]\) and \([11]\) in \([9]\), we now obtain
\[
M_n = c_0 \cdot \exp \left( \sum_{j \in \{0, n\}} B_{Z_j}(\theta^*, \theta_j) - \frac{\eta}{2} \|\theta^*\|_A^2 \right)
\times \int_{\mathbb{R}^d} \exp \left( \sum_{j \in \{0, n\}} ((\lambda - \lambda^*)^\top \nabla Z_j(\theta^* + \lambda^*) + Z_j(\theta^* + \lambda^*) - Z_j(\theta^* + \lambda)) \right) d\lambda
= c_0 \cdot \exp \left( \sum_{j \in \{0, n\}} B_{Z_j}(\theta^*, \theta_n) - \frac{\eta}{2} \|\theta^*\|_A^2 \right) \cdot \exp \left( - \sum_{j \in \{0, n\}} ((\theta^* + \lambda^*)^\top \nabla Z_j(\theta^* + \lambda^*) - Z_j(\theta^* + \lambda)) \right)
\times \int_{\mathbb{R}^d} \exp \left( \sum_{j \in \{0, n\}} ((\theta^* + \lambda^*)^\top \nabla Z_j(\theta^* + \lambda^*) - Z_j(\theta^* + \lambda)) \right) d\lambda
= \frac{c_0}{c_n} \cdot \exp \left( \sum_{j \in \{0, n\}} B_{Z_j}(\theta^*, \theta_n) - \frac{\eta}{2} \|\theta^*\|_A^2 \right) \cdot \frac{\int_{\mathbb{R}^d} \exp \left( \sum_{j \in \{0, n\}} ((\theta^* + \lambda^*)^\top \nabla Z_j(\theta^* + \lambda^*) - Z_j(\theta^* + \lambda)) \right) d\lambda \cdot 1}{\int_{\mathbb{R}^d} \exp \left( \sum_{j \in \{0, n\}} ((\theta^* + \lambda^*)^\top \nabla Z_j(\theta^* + \lambda^*) - Z_j(\theta^*)) \right) d\theta^*}
= \frac{c_0}{c_n} \cdot \exp \left( \sum_{j \in \{0, n\}} B_{Z_j}(\theta^*, \theta_n) - \frac{\eta}{2} \|\theta^*\|_A^2 \right) \cdot \frac{1}{\int_{\mathbb{R}^d} \exp \left( - \sum_{j \in \{0, n\}} B_{Z_j}(\theta^*, \theta_n) - \frac{\eta}{2} \|\theta^* - \theta_n\|_A^2 \right) d\theta^*}
= c_n \cdot \exp \left( \frac{\exp \left( \sum_{j \in \{0, n\}} ((\theta^* + \lambda^*)^\top \nabla Z_j(\theta^* + \lambda^*) - Z_j(\theta^* + \lambda^*)) \right)}{\int_{\mathbb{R}^d} \exp \left( \sum_{j \in \{0, n\}} ((\theta^* + \lambda^*)^\top \nabla Z_j(\theta^* + \lambda^*) - Z_j(\theta^*)) \right) d\theta^*} \cdot 1 \right)
= \frac{1}{\int_{\mathbb{R}^d} \exp \left( - \sum_{j \in \{0, n\}} B_{Z_j}(\theta^*, \theta_n) - \frac{\eta}{2} \|\theta^* - \theta_n\|_A^2 \right) d\theta^*}.
\]
where we have introduced $c_n = \frac{\exp \left( \sum_{j \in \{0, n\}} ((\theta^* + \lambda^*)^\top \nabla Z_j(\theta^* + \lambda^*) - Z_j(\theta^* + \lambda^*)) \right)}{\int_{\mathbb{R}^d} \exp \left( \sum_{j \in \{0, n\}} ((\theta^* + \lambda^*)^\top \nabla Z_j(\theta^* + \lambda^*) - Z_j(\theta^*)) \right) d\theta^*}$.

Therefore, we have from \([5]\) that
\[
C_{A,n} := c_n = \frac{\int_{\mathbb{R}^d} \exp \left( - \frac{\eta}{2} \|\theta^*\|_A^2 \right) d\theta^*}{\int_{\mathbb{R}^d} \exp \left( - \sum_{i=1}^n \mathbf{KL}_{\theta_i, a_i}(\theta_n, \theta_i) - \frac{\eta}{2} \|\theta^* - \theta_n\|_A^2 \right) d\theta^*}.
\]
An application of Markov’s inequality now yields
\[
P \left[ \sum_{i=1}^n \mathbf{KL}_{\theta_i, a_i}(\theta_n, \theta_i) + \frac{\eta}{2} \|\theta^* - \theta_n\|_A^2 - \frac{\eta}{2} \|\theta^*\|_A^2 \geq \log \left( \frac{C_{A,n}}{\delta} \right) \right] = P \left[ M_n \geq \frac{1}{\delta} \right] \leq \delta \cdot E [M_n] = \delta.
\]

**Step 3: A Stopped Martingale and Its Control** Let $N$ be a stopping time with respect to the filtration $\{\mathcal{F}_t\}_{t=0}^\infty$. Now, by the martingale convergence theorem, $M_\infty = \lim_{n \to \infty} M_n$ is almost surely well-defined, and thus $M_N$ is well-defined as well irrespective of whether $N < \infty$ or not. Let $Q_n = M_{\min\{N,n\}}$ be a stopped version of $\{M_n\}_n$. Then an application of Fatou’s lemma yields
\[
E [M_N] = E \left[ \liminf_{n \to \infty} Q_n \right] \leq \liminf_{n \to \infty} E [Q_n] = \liminf_{n \to \infty} E \left[ M_{\min\{N,n\}} \right] \leq 1,
\]
since the stopped martingale $\{M_{\min\{N,n\}}\}_{n \geq 1}$ is also a martingale. Therefore, by the properties of $M_n$, \([12]\) also holds for any random stopping time $N < \infty$.

To complete the proof, we now employ a random stopping time construction as in Abbasi-Yadkori et al. (2011).
We define a random stopping time \( N \) by
\[
N = \min \left\{ n \geq 1 : \sum_{t=1}^{n} \text{KL}_{s_t,a_t}(\theta_n, \theta^*) + \frac{n}{2} \| \theta^* - \theta_n \|_{\text{A}}^2 \geq \log \left( \frac{C_{\text{A},n}}{\delta} \right) \right\},
\]
with \( \min \{ \emptyset \} := \infty \) by convention. We then have
\[
P \left[ \exists n \geq 1, \sum_{t=1}^{n} \text{KL}_{s_t,a_t}(\theta_n, \theta^*) + \frac{n}{2} \| \theta^* - \theta_n \|_{\text{A}}^2 \geq \log \left( \frac{C_{\text{A},n}}{\delta} \right) \right] = P \left[ N < \infty \right] \leq \delta,
\]
which concludes the proof of the first part.

**Proof of Second Part: Upper Bound on \( C_{\text{A},n} \)** First, we have for some \( \hat{\theta} \in [\theta_n, \theta^*] \), that
\[
\text{KL}_{s,a}(\theta_n, \theta^*) = \frac{1}{2} \sum_{i,j=1}^{d} (\theta^* - \theta_n)i_i \varphi(s, a) \top \text{A}_{i} \text{C}_{i,a}^{\beta} [\psi(s')] A_j \varphi(s, a)(\theta^* - \theta_n) \, .
\]
(13)

Now (13) implies that
\[
\sum_{t=1}^{n} \text{KL}_{s_t,a_t}(\theta_n, \theta^*) \leq \frac{\beta}{2} \sum_{t=1}^{n} \sum_{i,j=1}^{d} (\theta^* - \theta_n)i_i \varphi(s, a_t) \top \text{A}_{i} \text{A}_j \varphi(s, a_t)(\theta^* - \theta_n) = \frac{\beta}{2} \| \theta^* - \theta_n \|_{\text{A}}^2,
\]
where \( \beta := \sup_{\theta, s,a} \lambda_{\max} \left( \text{C}_{i,a}^{\beta} [\psi(s')] \right) \) and \( \forall i, j \leq d \), \( (G_{s,a})_{i,j} := \varphi(s, a) \top \text{A}_i \text{A}_j \varphi(s, a) \). Therefore, we obtain
\[
C_{\text{A},n} \leq \frac{\int_{\mathbb{R}^d} \exp \left( - \frac{\beta}{2} \| \theta^* \|_{\text{A}}^2 \right) d\theta^*}{\int_{\mathbb{R}^d} \exp \left( - \frac{\beta}{2} \| \theta - \theta_n \|_{\text{A}}^2 \right) d\theta \sum_{t=1}^{n} G_{s_t,a_t} + \eta A_{i}} \frac{(2\pi)^{d/2}}{\det(\eta A)^{1/2}},
\]
which completes the proof of the second part.

**C REGRENT BOUND OF Exp-UCRL: PROOF OF THEOREM 2**

**Step 1: Optimism** Let us consider the start of episode \( t \), i.e., when the total number of steps completed is \( n = (t-1)H \). Recall that \( \theta_n \equiv \theta_{(t-1)H} \) denotes the penalized MLE and \( \Theta_n \equiv \Theta_{(t-1)H} \) the confidence set around the MLE after \( n \) steps. Now, let \( \hat{\theta}_n \equiv \hat{\theta}_{(t-1)H} \) denotes the most optimistic realization from the confidence set \( \Theta_n \), i.e.,
\[
V_{\hat{\theta}_n,1}^\pi (s_1^t) = \max_{\pi \in \Pi} \max_{\theta \in \Theta_n} V_{\hat{\theta}_n,1}^\pi (s_1^t),
\]
where \( s_1^t \) denotes the starting state at episode \( t \). Therefore, as long as the true parameter \( \theta_n \) belongs to \( \Theta_n \), \( V_{\hat{\theta}_n,1}^\pi (s_1^t) \) gives an optimistic estimate of the value \( V_{\theta_n,1}^\pi (s_1^t) \) of the episode, i.e.,
\[
V_{\hat{\theta}_n,1}^\pi (s_1^t) \geq V_{\theta_n,1}^\pi (s_1^t) \, .
\]
(14)

An application of [1] implies that with probability at least \( 1 - \delta/2 \), \( \theta^* \in \Theta_n \) across all episodes. We then have from [1] that with probability at least \( 1 - \delta/2 \), the cumulative regret is controlled by
\[
\mathcal{R}(N) \leq \sum_{t=1}^{T} \left( V_{\hat{\theta}_n,1}^\pi (s_1^t) - V_{\theta_n,1}^\pi (s_1^t) \right) \, ,
\]
(15)

where \( N = TH \) denotes the total number of steps completed after \( T \) episodes.

**Step 2: Bellman Recursion, Transportation Inequalities and Martingale Control** For any parameter \( \theta \in \mathbb{R}^d \) and policy \( \pi \in \Pi \), the Bellman operator \( \mathcal{T}_{\theta,h}^\pi : \mathcal{S} \to \mathbb{R} \) is defined for all \( s \in \mathcal{S} \) and \( h \in [H] \) as
\[
\mathcal{T}_{\theta,h}^\pi (V)(s) = R(s, \pi(s, h)) + \mathbb{E}_{s, \pi(s, h)} [V] \, ,
\]
where \( V : \mathcal{S} \to \mathbb{R} \). By the Bellman equation, we have
\[
V_{\theta,h}^\pi (s) = \mathcal{T}_{\theta,h}^\pi (V_{\theta,h+1}^\pi)(s), \quad \forall h \in [H] \quad (\text{with } V_{\theta,H+1}^\pi (s) := 0).
\]
Following, e.g., Chowdhury and Gopalan (2019), a recursive application of Bellman equation now yields
\[
V_{\theta_n^{t+1}}(s_{t+1}^t) - V_{\theta_*}^{t+1}(s_{t+1}^t) = \sum_{h=1}^{H} \left( T_{\theta_n,h}^{t} \left( V_{\theta_n}^{t+1} \right) (s_{t+1}^h) - T_{\theta_*}^{t} \left( V_{\theta_*}^{t+1} \right) (s_{t+1}^h) + m_h^t \right),
\]
where \( m_h^t = \mathbb{E}_{s_h^{t,a_h^t}}^{\theta_h^{t} \cdot a_h^t} \left[ V_{\theta_n}^{t+1}(s_{t+1}^h) - V_{\theta_*}^{t+1}(s_{t+1}^h) \right] - \left( V_{\theta_n}^{t+1}(s_{t+1}^h) - V_{\theta_*}^{t+1}(s_{t+1}^h) \right) \). Note that \( \{m_h^t\}_{t,h} \) is a martingale sequence satisfying \( |m_h^t| \leq 2H \). Therefore, by the Azuma-Hoeffding inequality (Boucheron et al., 2013), with probability at least \( 1 - \delta/2 \), we obtain
\[
\sum_{t=1}^{T} \sum_{h=1}^{H} m_h^t \leq 2H \sqrt{2TH \ln(2/\delta)} = 2H \sqrt{2N \ln(2/\delta)}.
\]
Then, using a union bound argument along with (15), the cumulative regret can be upper bounded with probability at least \( 1 - \delta \) as
\[
\mathcal{R}(N) \leq \sum_{t=1}^{T} \sum_{h=1}^{H} \left( T_{\theta_n,h}^{t} \left( V_{\theta_n}^{t+1} \right) (s_{t+1}^h) - T_{\theta_*}^{t} \left( V_{\theta_*}^{t+1} \right) (s_{t+1}^h) + 2H \sqrt{2N \ln(2/\delta)} \right).
\]  
(16) 
We now proceed to bound the first term in (16). Since \( V_{\theta_n}^{t+1}(s_{t+1}^h) \leq H \), \( \forall s_h \), we have its span \( S \left( V_{\theta_n}^{t+1} \right) \leq H \) and variance \( \mathbb{V}_{s_h,a_h}^{\theta} \left[ V_{\theta_n}^{t+1} \right] \leq H^2, \ \forall \theta, \ \forall(s,a) \). Therefore, we obtain
\[
T_{\theta_n,h}^{t} \left( V_{\theta_n}^{t+1} \right) (s_{t+1}^h) - T_{\theta_*}^{t} \left( V_{\theta_*}^{t+1} \right) (s_{t+1}^h) = \mathbb{E}_{s_h^{t,a_h^t}}^{\theta_h^{t} \cdot a_h^t} \left[ V_{\theta_n}^{t+1} \right] - \mathbb{E}_{s_h^{t,a_h^t}}^{\theta_h^{t} \cdot a_h^t} \left[ V_{\theta_*}^{t+1} \right] + \mathbb{E}_{s_h^{t,a_h^t}}^{\theta_h^{t} \cdot a_h^t} \left[ V_{\theta_*}^{t+1} \right] - \mathbb{E}_{s_h^{t,a_h^t}}^{\theta_h^{t} \cdot a_h^t} \left[ V_{\theta_*}^{t+1} \right]
\leq H \sqrt{2KL_{s_h^{t,a_h^t}}(\theta_n, \theta_n) + H \sqrt{2KL_{s_h^{t,a_h^t}}(\theta_n, \theta_n) + 2H/3KL_{s_h^{t,a_h^t}}(\theta_n, \theta_n)}},
\]
where the last step follows from the transportation inequalities (Lemma 1). We then obtain from (16) that
\[
\mathcal{R}(N) \leq H \sum_{t=1}^{T} \sum_{h=1}^{H} \left( \sqrt{2KL_{s_h^{t,a_h^t}}(\theta_n, \theta_n) + 2H/3KL_{s_h^{t,a_h^t}}(\theta_n, \theta_n)} \right) + 2H \sqrt{2N \ln(2/\delta)}.
\]  
(17) 

Step 3: Sum of KL Divergences Along the Transition Trajectory First, we obtain from (13) that
\[
\forall(s,a) \in S \times A, \ \forall \theta, \theta' \in \mathbb{R}^d, \ \frac{\alpha}{2} \|\theta' - \theta\|^2 \leq KL_{s,a}(\theta, \theta') \leq \frac{\beta}{2} \|\theta' - \theta\|^2,
\]
where \( \alpha := \inf_{\theta, \theta'} \lambda_{\min} \left( \mathbb{C}_{s,a}[\psi(s')] \right), \ \beta := \sup_{\theta, \theta'} \lambda_{\max} \left( \mathbb{C}_{s,a}[\psi(s')] \right), \ \forall i, j \leq d, \ (G_{s,a})_{i,j} := \varphi(s,a)^T A_i^T A_j \varphi(s,a) \). We then have
\[
\forall(s,a), \ \forall \theta, \ \KL_{s,a}(\theta_n, \theta) \leq \frac{\beta}{2} \|\theta - \theta_n\|^2_{\mathbb{C}_{s,a}} \leq \beta \|\mathbb{G}^{-1/2}_{n, s,a} \mathbb{G}^{-1/2}_{n, s,a} \| \frac{1}{2} \|\theta - \theta_n\|^2_{\mathbb{G}_{n, s,a}},
\]
where \( \mathbb{G}_n \equiv \mathbb{G}_{n,(t-1)H} := G_n + \alpha^{-1} \eta A \) and \( G_n \equiv G_{n,(t-1)H} := \sum_{t=1}^{H} \sum_{i=1}^{H} KL_{s_i^{t,a_i^t}}(\theta_n, \theta) \). Furthermore, note that
\[
\frac{1}{2} \|\theta - \theta_n\|^2_{\mathbb{G}_{n, s,a}} = \frac{\alpha^{-1} \eta}{2} \|\theta - \theta_n\|^2_{\mathbb{A}} + \sum_{t=1}^{H} \frac{1}{2} \|\theta - \theta_n\|^2_{\mathbb{G}_{n, s,a}} \leq \alpha^{-1} \left( \frac{\eta}{2} \|\theta - \theta_n\|^2_{\mathbb{A}} + \sum_{t=1}^{H} KL_{s_i^{t,a_i^t}}(\theta_n, \theta) \right).
\]
Therefore, for any \( \theta \in \Theta_n \), we obtain
\[
\forall(s,a), \ \KL_{s,a}(\theta_n, \theta) \leq \frac{\beta}{\alpha} \cdot \beta_n(\delta) \frac{\mathbb{G}_{n,(t-1)H}^{-1/2}}{\mathbb{G}_{n,(t-1)H}^{-1/2}} = \frac{\beta}{\alpha} \cdot \beta_n(\delta) \|\mathbb{G}_{n,(t-1)H}^{-1} \mathbb{G}_{n,(t-1)H} \|, \ \forall(s,a),
\]  
(18) 
Now, since \( G_n \) is positive semi-definite, we have \( \mathbb{G}_n \geq \alpha^{-1} \eta \mathbb{A} \), and thus, in turn
\[
\|\mathbb{G}_{n,(t-1)H}^{-1/2} \mathbb{G}_{s,a} \| \leq \frac{\alpha}{\eta} \|\mathbb{A}^{-1} \mathbb{G}_{s,a} \| \leq \frac{\alpha B_{s,a}}{\eta}, \ \forall(s,a),
\]
where $B_{\varphi, A} := \sup_{s,a} \| A^{-1} G_{s,a} \|$. This further yields
\begin{equation}
\left\| I + \mathcal{G}^{-1}_n \sum_{h=1}^H G_{s_h, a_h} \right\| \leq 1 + \sum_{h=1}^H \left\| \mathcal{G}^{-1}_n G_{s_h, a_h} \right\| \leq 1 + \frac{\alpha B_{\varphi, A} H}{\eta} .
\end{equation}

Now, we define $\mathcal{G}^{-1}_{n,H} := \mathcal{G}_n + \sum_{h=1}^H G_{s_h, a_h}$. Hence, $\mathcal{G}^{-1}_{n,H} G_{s,a} = \left( I + \mathcal{G}^{-1}_n \sum_{h=1}^H G_{s_h, a_h} \right)^{-1} \mathcal{G}^{-1}_n G_{s,a}$. We therefore deduce from (19) that
\begin{equation}
\forall (s,a), \quad \left\| \mathcal{G}^{-1}_n G_{s,a} \right\| = \left\| \left( I + \mathcal{G}^{-1}_n \sum_{h=1}^H G_{s_h, a_h} \right) \mathcal{G}^{-1}_{n,H} G_{s,a} \right\| \leq \left( 1 + \frac{\alpha B_{\varphi, A} H}{\eta} \right) \left\| \mathcal{G}^{-1}_{n,H} G_{s,a} \right\| .
\end{equation}

Now see that
\begin{equation}
\sum_{t=1}^T \sum_{h=1}^H \left\| \mathcal{G}^{-1}_{n,H} G_{s_h, a_h} \right\| \leq \sum_{t=1}^T \sum_{h=1}^H \text{tr} \left( \mathcal{G}^{-1}_{n,H} G_{s_h, a_h} \right) = \sum_{t=1}^T \text{tr} \left( \mathcal{G}^{-1}_{n,H} \left( \mathcal{G}_{n,H} - \mathcal{G}_n \right) \right) \leq \sum_{t=1}^T \log \frac{\det(\mathcal{G}_{n,H})}{\det(\mathcal{G}_n)} ,
\end{equation}
where we have used that for two positive definite matrices $A$ and $B$ such that $A-B$ is positive semi-definite, $\text{tr}(A^{-1}(A-B)) \leq \log \frac{\det(A)}{\det(B)}$. We can now control the R.H.S. of the above equation, as
\begin{equation}
\sum_{t=1}^T \log \frac{\det(\mathcal{G}_{n,H})}{\det(\mathcal{G}_n)} = \sum_{t=1}^T \log \frac{\det(\mathcal{G}_{n+H})}{\det(\mathcal{G}_n)} = \log \frac{\det(\mathcal{G}_{n+H})}{\det(\mathcal{G}_n)} = \log \frac{\det(\mathcal{G}_n)}{\det(\mathcal{G}_{n+H})} = \log \det \left( I + \beta^{-1} A^{-1} G_N \right) .
\end{equation}

Therefore, we have from (20) and that
\begin{equation}
\sum_{t=1}^T \sum_{h=1}^H \left\| \mathcal{G}^{-1}_n G_{s_h, a_h} \right\| \leq \left( 1 + \frac{\beta B_{\varphi, A} H}{\eta} \right) \log \det \left( I + \beta^{-1} A^{-1} G_N \right) ,
\end{equation}
where we have used that $\alpha \leq \beta$.

It now remains to bound the log determinant term in the above equation. By the trace-determinant inequality, we have
\begin{equation}
\det \left( I + \beta^{-1} A^{-1} G_N \right) \leq \left( \frac{\text{tr} \left( I + \beta^{-1} A^{-1} G_N \right)}{d} \right)^d \leq \left( 1 + \frac{\beta^{-1}}{d} \text{tr} \left( A^{-1} G_N \right) \right)^d .
\end{equation}

Now see that $\text{tr} \left( A^{-1} G_N \right) \leq n \sup_{s,a} \text{tr} \left( A^{-1} G_{s,a} \right) \leq dB_{\varphi, A} n$. Therefore, we have
\begin{equation}
\log \det \left( I + \beta^{-1} A^{-1} G_N \right) \leq d \log \left( 1 + \beta^{-1} B_{\varphi, A} n \right) .
\end{equation}
This further implies that the confidence radius
\begin{equation}
\beta_n(d) \leq \frac{n}{2} B_{\varphi}^2 + \log \left( 2 \det \left( I + \beta^{-1} A^{-1} G_N \right) / \delta \right) \leq \frac{n}{2} B_{\varphi}^2 + d \log \left( 1 + \beta^{-1} B_{\varphi, A} n \right) + \log(2/\delta) ,
\end{equation}
which is an increasing function in the total number of steps $n$, hence, in the number of episodes $t$. We then have from (15) and (21) that
\begin{equation}
\forall \theta \in \Theta_n, \quad \sum_{t=1}^T \sum_{h=1}^H \text{KL} \left( s_h, a_h \right)(\theta, \theta) \leq \frac{\beta}{\alpha} \left( 1 + \frac{\beta B_{\varphi, A} H}{\eta} \right) \beta_n(\delta) \gamma_N ,
\end{equation}
where we define $\gamma_N := d \log \left( 1 + \beta^{-1} B_{\varphi, A} N \right)$ and $\beta_n(\delta) := \frac{n}{2} B_{\varphi}^2 + \gamma_N + \log(2/\delta)$.

**Final Step:** First, an application of Cauchy-Schwartz’s inequality yields
\begin{equation}
\forall \theta \in \Theta_n, \quad \sum_{t=1}^T \sum_{h=1}^H \sqrt{\text{KL} \left( s_h, a_h \right)(\theta, \theta)} \leq \sqrt{N} \sum_{t=1}^T \sum_{h=1}^H \text{KL} \left( s_h, a_h \right)(\theta, \theta) \leq \sqrt{\frac{\beta}{\alpha} \left( 1 + \frac{\beta B_{\varphi, A} H}{\eta} \right) \beta_n(\delta)N \gamma_N} .
\end{equation}
At this point, we note that by design, $\hat{\theta}_n \in \Theta_n$ and by Theorem 1, $\theta^* \in \Theta_n$ with probability at least $1 - \delta/2$. We now obtain from (17), (23) and (24) that the cumulative regret
\begin{equation}
\mathcal{R}(N) \leq 2H \sqrt{\frac{\beta}{\alpha} \left( 1 + \frac{\beta B_{\varphi, A} H}{\eta} \right) 2\beta_n(\delta)N \gamma_N} + 2H \sqrt{2N \ln(2/\delta)} + \frac{2H}{3} \frac{\beta}{\alpha} \left( 1 + \frac{\beta B_{\varphi, A} H}{\eta} \right) \beta_n(\delta) \gamma_N ,
\end{equation}
which completes the proof.
D REGRET BOUND OF Exp-PSRL: PROOF OF THEOREM 3

Let us consider the start of episode $t$, i.e., when the total number of steps completed is $n = (t-1)H$. Recall that we sample $\hat{\theta}_n \equiv \hat{\theta}_{(t-1)H} \sim \mu_{\mathcal{H}_n}$, where $\mu_{\mathcal{H}_n} = \mathbb{P}(\theta^* \in \mathcal{H}_n)$ denotes the posterior distribution of $\theta^*$, given the history of transitions $\mathcal{H}_n = \mathcal{H}_{(t-1)H} = \{(s_{h-1}^t, a_{h-1}^t, s_{h+1}^t)_{t \leq H} \}$. A key property of posterior sampling is that for any $\sigma(\mathcal{H}_n)$-measurable function $f$, we have $\mathbb{E}[f(\hat{\theta}_n)] = \mathbb{E}[f(\theta^*)]$ [Osband et al., 2013]. This implies that the optimal policy $\pi^*$ and selected policy $\pi_t^*$ are identically distributed conditioned on the history $\mathcal{H}_n$. Therefore, we have $\mathbb{E} \left[ V_{\pi_t^*,1}^t(s_1^t) \right] = \mathbb{E} \left[ V_{\pi_t^*,1}(s_1^t) \right]$, and thus, in turn, the Bayes regret

$$\mathbb{E}[\mathcal{R}(N)] = \mathbb{E} \left[ \sum_{t=1}^T \left( V_{\pi_t^*,1}^t(s_1^t) - V_{\pi_t^*,1}(s_1^t) \right) \right].$$

A recursive application of the Bellman equation now yields a result similar to (10):

$$\mathbb{E}[\mathcal{R}(N)] = \mathbb{E} \left[ \sum_{t=1}^T \sum_{h=1}^H \left( \sqrt{2} KL_{s_{h+1}^t,a_{h+1}^t}(\theta_n, \hat{\theta}_n) + \sqrt{2} KL_{s_{h+1}^t,a_{h+1}^t}(\theta^*, \theta^*) + \frac{2}{3} KL_{s_{h+1}^t,a_{h+1}^t}(\theta_n, \theta^*) \right) \right],$$

where $m_h^t = \mathbb{E}_{s_{h+1}^t,a_{h+1}^t} \left[ V_{\pi_t^*,1}(s_{h+1}^t) - V_{\theta^*,1}(s_{h+1}^t) \right]$ is a martingale difference sequence satisfying $\mathbb{E}[m_h^t] = 0$. Then an application of the transportation inequalities (Lemma 1) yields a result similar to (17):

$$\mathbb{E}[\mathcal{R}(N)] \leq H \mathbb{E} \left[ \sum_{t=1}^T \sum_{h=1}^H \left( \sqrt{2} KL_{s_{h+1}^t,a_{h+1}^t}(\theta_n, \hat{\theta}_n) + \sqrt{2} KL_{s_{h+1}^t,a_{h+1}^t}(\theta^*, \theta^*) + \frac{2}{3} KL_{s_{h+1}^t,a_{h+1}^t}(\theta_n, \theta^*) \right) \right],$$

where $\theta_n \equiv \hat{\theta}_{(t-1)H}$ denotes the penalized MLE (as computed by Exp-UCRL) after $n = (t-1)H$ steps.

We now define for any $\delta \in (0,1)$, the events $\mathcal{E}^* = \{ \forall t \geq 1, \theta^* \in \Theta_t \}$ and $\tilde{\mathcal{E}} = \{ \forall t \geq 1, \theta_n \in \Theta_t \}$, where $\Theta_t \equiv \Theta_{(t-1)H}$ is confidence set (as constructed by Exp-UCRL) after $n = (t-1)H$ steps. Under the event $\mathcal{E}^* \cap \tilde{\mathcal{E}}$, we have from (23) and (24) that

$$\sum_{t=1}^T \sum_{h=1}^H KL_{s_{h+1}^t,a_{h+1}^t}(\theta_n, \theta^*) \leq \frac{\beta}{\alpha} \left( 1 + \frac{\beta B_{\mathcal{Z},AH}}{\eta} \right) \beta_N(\delta) \gamma_N,$$

$$\sum_{t=1}^T \sum_{h=1}^H KL_{s_{h+1}^t,a_{h+1}^t}(\theta_n, \theta^*) \leq \frac{\beta}{\alpha} \left( 1 + \frac{\beta B_{\mathcal{Z},AH}}{\eta} \right) \beta_N(\delta) \gamma_N \quad \text{and}$$

$$\sum_{t=1}^T \sum_{h=1}^H KL_{s_{h+1}^t,a_{h+1}^t}(\theta_n, \hat{\theta}_n) \leq \frac{\beta}{\alpha} \left( 1 + \frac{\beta B_{\mathcal{Z},AH}}{\eta} \right) \beta_N(\delta) \gamma_N.$$

Therefore, we obtain from (25), the following:

$$\mathbb{E}[\mathcal{R}(N) \mathbb{I}_{\mathcal{E}^* \cap \tilde{\mathcal{E}}}] \leq 2H \sqrt{\frac{\beta}{\alpha} \left( 1 + \frac{\beta B_{\mathcal{Z},AH}}{\eta} \right) 2 \beta_N(\delta) \gamma_N + \frac{2}{3} \frac{\beta}{\alpha} \left( 1 + \frac{\beta B_{\mathcal{Z},AH}}{\eta} \right) \beta_N(\delta) \gamma_N}.$$

Since we can always bound $\mathcal{R}(N) \leq N$, we have

$$\mathbb{E}[\mathcal{R}(N)] = \mathbb{E} \left[ \mathcal{R}(N) \mathbb{I}_{\mathcal{E}^* \cap \tilde{\mathcal{E}}} + \mathcal{R}(N) \mathbb{I}_{(\mathcal{E}^* \cap \tilde{\mathcal{E}})^c} \right] \leq \mathbb{E}[\mathcal{R}(N) \mathbb{I}_{\mathcal{E}^* \cap \tilde{\mathcal{E}}}] + N(1 - \mathbb{P}(\mathcal{E}^* \cap \tilde{\mathcal{E}})).$$

Now from the property of posterior sampling, $\mathbb{P}(\tilde{\mathcal{E}}) = \mathbb{P}(\mathcal{E}^*)$ and from Theorem 1, $\mathbb{P}(\mathcal{E}^*) \geq 1 - \delta/2$. Therefore, by a union bound, $\mathbb{P}(\mathcal{E}^* \cap \tilde{\mathcal{E}}) \geq 1 - \delta$. This implies for any $\delta \in (0,1]$ that the Bayes regret

$$\mathbb{E}[\mathcal{R}(N)] \leq 2H \sqrt{\frac{\beta}{\alpha} \left( 1 + \frac{\beta B_{\mathcal{Z},AH}}{\eta} \right) 2 \beta_N(\delta) \gamma_N + \frac{2}{3} \frac{\beta}{\alpha} \left( 1 + \frac{\beta B_{\mathcal{Z},AH}}{\eta} \right) \beta_N(\delta) \gamma_N + N\delta}.$$

The proof now can be completed by setting $\delta = \frac{1}{N}$.

E ON THE CHOICE OF PENALTY FUNCTION

In this paper, we have considered the penalty function $\operatorname{pen}(\theta) = \frac{1}{2} \| \theta \|^2_r$, where $\forall i, j \leq d, \ A_{i,j} = \text{tr}(A_i A_j^T)$. We however note that all our results (Theorem 1, 2, 3) hold for any choice of the (regularizing) matrix $A$. For any
such choice of $A$, we only need to ensure that there exist a known constant $B_A$ such that $\|\theta^*\|_A \leq B_A$. In fact for our particular choice, as we have seen in Section $4$, we obtain $A = I$ for factored and tabular MDPs and $A = m_1 I$ for the linearly controlled dynamical systems. (The scaling with $m_1$ arises because of our parameterization and can be suppressed for the special case of $\Sigma_{s,a} = cI$, $c > 0$, $\forall (s, a)$ by using a reparameterization.) We leave it to future work to study the effect of other possible regularizing matrices and penalty functions.