Appendix

A PRELIMINARIES

A.1 Transportation Inequalities

For any function $f : \mathcal{X} \to \mathbb{R}$, we define its span as $\mathbb{S}(f) := \max_{x\in\mathcal{X}} f(x) - \min_{x\in\mathcal{X}} f(x)$. For a probability distribution P supported on the set $\mathcal{X}$, let $\mathbb{E}_P[f] := \mathbb{E}_P[f(X)]$ and $\mathbb{V}_P[f] := \mathbb{V}_P[f(X)] = \mathbb{E}_P\left[f(X)^2\right] - \mathbb{E}_P[f(X)]^2$ 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 $\mathbb{S}(f)$ and $\mathbb{V}_P[f]$ are finite. Then it holds*

$$\forall Q \ll P, \quad \mathbb{E}_Q[f] - \mathbb{E}_P[f] \le \sqrt{2\,\mathbb{V}_P[f]\,\mathrm{KL}(Q,P)} + \frac{2\,\mathbb{S}(f)}{3}\,\mathrm{KL}(Q,P)\ ,$$
$$\forall Q \ll P, \quad \mathbb{E}_P[f] - \mathbb{E}_Q[f] \le \sqrt{2\,\mathbb{V}_P[f]\,\mathrm{KL}(Q,P)}\ .$$

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

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

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 $\theta, \theta' \in \mathbb{R}^d$:

$$B_F(\theta', \theta) = B^\star_{F,\theta'}\left(\nabla F(\theta) - \nabla F(\theta')\right)\ . \tag{4}$$

To see this, we observe that

$$\begin{aligned} & B^\star_{F,\theta'}(\nabla F(\theta) - \nabla F(\theta')) \\ =\ & \sup_{\lambda\in\mathbb{R}^d} \lambda^\top\left(\nabla F(\theta) - \nabla F(\theta')\right) - \left[F(\theta' + \lambda) - F(\theta') - \lambda^\top \nabla F(\theta')\right] \\ =\ & \sup_{\lambda\in\mathbb{R}^d} \lambda^\top \nabla F(\theta) - F(\theta' + \lambda) + F(\theta')\,. \end{aligned}$$

Now an optimal λ 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

$$B^\star_{F,\theta'}(\nabla F(\theta) - \nabla F(\theta')) = (\theta - \theta')^\top \nabla F(\theta) - F(\theta) + F(\theta') = B_F(\theta', \theta)\,.$$

(Note that (4) 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,

$$\begin{aligned} (\nabla_i Z_{s,a})(\theta) &= \int_{\mathcal{S}} \psi(s')^\top A_i \varphi(s,a) \frac{h(s', s, a) \exp\left(\sum_{i=1}^d \theta_i \psi(s')^\top A_i \varphi(s,a)\right)}{\int_{\mathcal{S}} h(s', s, a) \exp\left(\sum_{i=1}^d \theta_t \psi(s')^\top A_i \varphi(s,a)\right) ds'} ds' \\ &= \mathbb{E}^\theta_{s,a}\Big[\psi(s')\Big]^\top A_i \varphi(s,a)\,. \end{aligned}$$

On the other hand, the entries of the Hessian of Z are given by

$$
\begin{aligned}
(\nabla^2_{i,j} Z_{s,a})(\theta) &= \int_{\mathcal{S}} \psi(s')^\top A_i \varphi(s,a)\psi(s')^\top A_j\varphi(s,a) \frac{h(s',s,a)\exp\Big(\sum_{i=1}^d \theta_i \psi(s')^\top A_i \varphi(s,a)\Big)}{\int_{\mathcal{S}} h(s',s,a)\exp\Big(\sum_{i=1}^d \theta_t \psi(s')^\top A_i\varphi(s,a)\Big) ds'} ds' \\
&\quad - \int_{\mathcal{S}} \psi(s')^\top A_i \varphi(s,a) \frac{h(s',s,a)\exp\Big(\sum_{i=1}^d \theta_i \psi(s')^\top A_i \varphi(s,a)\Big)}{\int_{\mathcal{S}} h(s',s,a)\exp\Big(\sum_{i=1}^d \theta_t \psi(s')^\top A_i\varphi(s,a)\Big) ds'} ds' (\nabla_j Z_{s,a})(\theta) \\
&= \mathbb{E}^\theta_{s,a}\Big[\psi(s')^\top A_i\varphi(s,a)\psi(s')^\top A_j\varphi(s,a)\Big] \\
&\quad - \mathbb{E}^\theta_{s,a}\Big[\psi(s')^\top A_i\varphi(s,a)\Big]\mathbb{E}^\theta_{s,a}\Big[\psi(s')^\top A_j\varphi(s,a)\Big] \\
&= \varphi(s,a)^\top A_i^\top \Big(\mathbb{E}^\theta_{s,a}\big[\psi(s')\psi(s')^\top\big] - \mathbb{E}^\theta_{s,a}\big[\psi(s')\big]\mathbb{E}^\theta_{s,a}\big[\psi(s')^\top\big]\Big) A_j \varphi(s,a) \\
&= \varphi(s,a)^\top A_i^\top \mathbb{C}^\theta_{s,a}\big[\psi(s')\big] A_j\varphi(s,a)\,,
\end{aligned}
$$

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

$$
\mathbb{C}^\theta_{s,a}\big[\psi(s')\big] = \mathbb{E}^\theta_{s,a}\big[\psi(s')\psi(s')^\top\big] - \mathbb{E}^\theta_{s,a}\big[\psi(s')\big]\mathbb{E}^\theta_{s,a}\big[\psi(s')^\top\big]\,,
$$

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

$$
\begin{aligned}
\log\left(\frac{P_\theta(s'|s,a)}{P_{\theta'}(s'|s,a)}\right) &= \sum_{i=1}^d (\theta_i - \theta'_i)\psi(s')^\top A_i \varphi(s,a) - Z_{s,a}(\theta) + Z_{s,a}(\theta')\,, \\
\text{or KL}\Big(P_\theta(\cdot|s,a), P_{\theta'}(\cdot|s,a)\Big) &= \sum_{i=1}^d (\theta_i - \theta'_i)\mathbb{E}^\theta_{s,a}[\psi(s')]^\top A_i\varphi(s,a) - Z_{s,a}(\theta) + Z_{s,a}(\theta') \\
&= \frac{1}{2}(\theta-\theta')^\top(\nabla^2 Z_{s,a})(\tilde\theta)(\theta-\theta')\,,
\end{aligned}
$$

where in the last line, we used, by a Taylor expansion, that $Z_{s,a}(\theta') = Z_{s,a}(\theta) + (\nabla Z_{s,a}(\theta))^\top(\theta'-\theta) + \frac{1}{2}(\theta - \theta')^\top(\nabla^2 Z_{s,a}(\tilde\theta))(\theta-\theta')$ for some $\tilde\theta \in [\theta,\theta']_\infty$. Here $[\theta,\theta']_\infty$ denotes the d-dimensional hypercube joining θ to θ'.

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.[7] 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.,

$$
\text{KL}_{s,a}(\theta,\theta') := \text{KL}\left(P_\theta(\cdot|s,a), P_{\theta'}(\cdot|s,a)\right) = B_{Z_{s,a}}(\theta',\theta)\,. \tag{5}
$$

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

$$
M_n^\lambda = \exp\left(\lambda^\top S_n - \sum_{t=1}^n B_{Z_{s_t,a_t},\theta^\star}(\lambda)\right)\,,
$$

[7]Since we will only use (4) in the proof, the final result would hold even if $Z_{s,a}$ is only convex.

where $\forall i \le d$, we denote $(S_n)_i = \sum_{t=1}^n \left(\psi(s'_t) - \mathbb{E}^{\theta^\star}_{s_t,a_t}[\psi(s')]\right)^\top A_i \varphi(s_t,a_t)$. Note that $M_n^\lambda > 0$ and it is $\mathcal{F}_n$-measurable. Furthermore, we have for all (s,a),

$$\begin{aligned}
&\mathbb{E}^{\theta^\star}_{s,a}\left[\exp\left(\sum_{i=1}^d \lambda_i \left(\psi(s') - \mathbb{E}^{\theta^\star}_{s,a}[\psi(s')]\right)^\top A_i\varphi(s,a)\right)\right] \\
&= \exp\left(-\lambda^\top \nabla Z_{s,a}(\theta^\star)\right) \int_{\mathcal{S}} h(s',s,a) \exp\left(\sum_{i=1}^d (\theta_i^\star + \lambda_i)\psi(s')^\top A_i \varphi(s,a) - Z_{s,a}(\theta^\star)\right) ds' \\
&= \exp\left(Z_{s,a}(\theta^\star+\lambda) - Z_{s,a}(\theta^\star) - \lambda^\top \nabla Z_{s,a}(\theta^\star)\right) = \exp\left(B_{Z_{s,a}}(\theta^\star)\right) .
\end{aligned}$$

This implies $\mathbb{E}\left[\exp(\lambda^\top S_n)|\mathcal{F}_{n-1}\right] = \exp\left(\lambda^\top S_{n-1} + B_{Z_{s_n,a_n},\theta^\star}(\lambda)\right)$ and thus, in turn, $\mathbb{E}[M_n^\lambda|\mathcal{F}_{n-1}] = M_{n-1}^\lambda$. 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 θ, we now define a mixture of martingales

$$M_n = \int_{\mathbb{R}^d} M_n^\lambda q(\theta^\star + \lambda) d\lambda \ . \tag{6}$$

Then $\{M_n\}_{n=0}^\infty$ is also a non-negative martingale adapted to $\{\mathcal{F}_n\}_{n=0}^\infty$ and in fact, $\mathbb{E}[M_n] = 1$.

Step 2: Method Of Mixtures and Martingale Control Considering the prior density $\mathcal{N}\left(0, (\eta\mathbb{A})^{-1}\right)$, we obtain from (6) that

$$M_n = c_0 \int_{\mathbb{R}^d} \exp\left(\lambda^\top S_n - \sum_{t=1}^n B_{Z_{s_t,a_t},\theta^\star}(\lambda) - \frac{\eta}{2}\|\theta^\star + \lambda\|^2_{\mathbb{A}}\right) d\lambda \ , \tag{7}$$

where $c_0 = \frac{1}{\int_{\mathbb{R}^d} \exp\left(-\frac{\eta}{2}\|\theta'\|^2_{\mathbb{A}}\right) d\theta'}$. We now introduce the function $Z_n(\theta) = \sum_{t=1}^n Z_{s_t,a_t}(\theta)$. Note that Z_n is a also Legendre function and its associated Bregman divergence satisfies

$$B_{Z_n}(\theta',\theta) = \sum_{t=1}^n \left(Z_{s_t,a_t}(\theta') - Z_{s_t,a_t}(\theta) - (\theta'-\theta)^\top \nabla Z_{s_t,a_t}(\theta)\right) = \sum_{t=1}^n B_{Z_{s_t,a_t}}(\theta',\theta)$$

Furthermore, we have $\sum_{t=1}^n B_{Z_{s_t,a_t},\theta^\star}(\lambda) = B_{Z_n,\theta^\star}(\lambda)$.

From the penalized likelihood formula (2), recall that

$$\forall i \le d, \quad \sum_{t=1}^n \nabla_i Z_{s_t,a_t}(\theta_n) + \frac{\eta}{2}\nabla_i \|\theta_n\|^2_{\mathbb{A}} = \sum_{t=1}^n \psi(s'_t)^\top A_i \varphi(s_t,a_t) \ .$$

This yields

$$S_n = \sum_{t=1}^n \left(\nabla Z_{s_t,a_t}(\theta_n) - \nabla Z_{s_t,a_t}(\theta^\star)\right) + \eta\mathbb{A}\theta_n = \nabla Z_n(\theta_n) - \nabla Z_n(\theta^\star) + \eta\mathbb{A}\theta_n \ . \tag{8}$$

We now obtain from (7) and (8) that

$$M_n = c_0 \cdot \exp\left(-\frac{\eta}{2}\|\theta^\star\|^2_{\mathbb{A}}\right) \int_{\mathbb{R}^d} \exp\left(\lambda^\top x_n - B_{Z_n,\theta^\star}(\lambda) + g_n(\lambda)\right) d\lambda \ , \tag{9}$$

where we have introduced $g_n(\lambda) = \frac{\eta}{2}\left(2\lambda^\top\mathbb{A}\theta_n + \|\theta^\star\|^2_{\mathbb{A}} - \|\theta^\star+\lambda\|^2_{\mathbb{A}}\right)$ and $x_n = \nabla Z_n(\theta_n) - \nabla Z_n(\theta^\star)$.

Now, note that $\sup_{\lambda\in\mathbb{R}^d} g_n(\lambda) = \frac{\eta}{2}\|\theta^\star - \theta_n\|^2_{\mathbb{A}}$, where the supremum is attained at $\lambda^\star = \theta_n - \theta^\star$. We then have

$$\begin{aligned}
g_n(\lambda) &= g_n(\lambda) + \sup_{\lambda\in\mathbb{R}^d} g_n(\lambda) - g_n(\lambda^\star) \\
&= \frac{\eta}{2}\|\theta_n - \theta^\star\|^2_{\mathbb{A}} + \eta(\lambda-\lambda^\star)^\top\mathbb{A}(\theta^\star+\lambda^\star) + \frac{\eta}{2}\|\theta^\star+\lambda^\star\|^2_{\mathbb{A}} - \frac{\eta}{2}\|\theta^\star+\lambda\|^2_{\mathbb{A}} \\
&= B_{Z_0}(\theta^\star,\theta_n) + (\lambda-\lambda^\star)^\top \nabla Z_0(\theta^\star+\lambda^\star) + Z_0(\theta^\star+\lambda^\star) - Z_0(\theta^\star+\lambda) \ ,
\end{aligned} \tag{10}$$

where we have introduced the Legendre function $Z_0(\theta) = \frac{\eta}{2}\|\theta\|^2_{\mathbb{A}}$. We now have from (4) that

$$\begin{aligned}
&\sup_{\lambda\in\mathbb{R}^d}\left(\lambda^\top x_n - B_{Z_n,\theta^\star}(\lambda)\right) \\
&= B^\star_{Z_n,\theta^\star}(x_n) = B^\star_{Z_n,\theta^\star}(\nabla Z_n(\theta_n) - \nabla Z_n(\theta^\star)) = B_{Z_n}(\theta^\star,\theta_n) \ .
\end{aligned}$$

Further, any optimal λ must satisfy

$$\nabla Z_n(\theta^\star+\lambda) - \nabla Z_n(\theta^\star) = x_n \implies \nabla Z_n(\theta^\star+\lambda) = \nabla Z_n(\theta_n) \ .$$

One possible solution is $\lambda = \lambda^\star$. Now, since Z_n is strictly convex, the supremum is indeed attained at $\lambda = \lambda^\star$. We then have

$$\begin{aligned}
&\lambda^\top x_n - B_{Z_n,\theta^\star}(\lambda)\\
&= \lambda^\top x_n - B_{Z_n,\theta^\star}(\lambda) + B_{Z_n}(\theta^\star,\theta_n) - \left(\lambda^{\star\top} x_n - B_{Z_n,\theta^\star}(\lambda^\star)\right)\\
&= B_{Z_n}(\theta^\star,\theta_n) + (\lambda-\lambda^\star)^\top \nabla Z_n(\theta^\star+\lambda^\star) + B_{Z_n,\theta^\star}(\lambda^\star) - B_{Z_n,\theta^\star}(\lambda) - (\lambda-\lambda^\star)^\top \nabla Z_n(\theta^\star)\\
&= B_{Z_n}(\theta^\star,\theta_n) + (\lambda-\lambda^\star)^\top \nabla Z_n(\theta^\star+\lambda^\star) + Z_n(\theta^\star+\lambda^\star) - Z_n(\theta^\star+\lambda)\ . \qquad (11)
\end{aligned}$$

Plugging (10) and (11) in (9), we now obtain

$$\begin{aligned}
M_n &= c_0 \cdot \exp\Big(\sum_{j\in\{0,n\}} B_{Z_j}(\theta^\star,\theta_j) - \frac{\eta}{2}\|\theta^\star\|^2_{\mathbb{A}}\Big)\\
&\quad\times \int_{\mathbb{R}^d} \exp\Big(\sum_{j\in\{0,n\}} \big((\lambda-\lambda^\star)^\top \nabla Z_j(\theta^\star+\lambda^\star) + Z_j(\theta^\star+\lambda^\star) - Z_j(\theta^\star+\lambda)\big)\Big) d\lambda\\
&= c_0 \cdot \exp\Big(\sum_{j\in\{0,n\}} B_{Z_j}(\theta^\star,\theta_n) - \frac{\eta}{2}\|\theta^\star\|^2_{\mathbb{A}}\Big)\cdot \exp\Big(-\sum_{j\in\{0,n\}} \big((\theta^\star+\lambda^\star)^\top \nabla Z_j(\theta^\star+\lambda^\star) - Z_j(\theta^\star+\lambda^\star)\big)\Big)\\
&\quad\times \int_{\mathbb{R}^d} \exp\Big(\sum_{j\in\{0,n\}} \big((\theta^\star+\lambda)^\top \nabla Z_j(\theta^\star+\lambda^\star) - Z_j(\theta^\star+\lambda)\big)\Big) d\lambda\\
&= \frac{c_0}{c_n}\cdot \exp\Big(\sum_{j\in\{0,n\}} B_{Z_j}(\theta^\star,\theta_n) - \frac{\eta}{2}\|\theta^\star\|^2_{\mathbb{A}}\Big)\cdot \frac{\int_{\mathbb{R}^d}\exp\Big(\sum_{j\in\{0,n\}}\big((\theta^\star+\lambda)^\top\nabla Z_j(\theta^\star+\lambda^\star) - Z_j(\theta^\star+\lambda)\big)\Big)d\lambda}{\int_{\mathbb{R}^d}\exp\Big(\sum_{j\in\{0,n\}}\big((\theta')^\top\nabla Z_j(\theta^\star+\lambda^\star) - Z_j(\theta')\big)\Big)d\theta'}\\
&= \frac{c_0}{c_n}\cdot \exp\Big(B_{Z_n}(\theta^\star,\theta_n) + B_{Z_0}(\theta^\star,\theta_n) - \frac{\eta}{2}\|\theta^\star\|^2_{\mathbb{A}}\Big)\cdot 1\\
&= \frac{c_0}{c_n}\cdot \exp\Big(\sum_{t=1}^n B_{Z_{s_t,a_t}}(\theta^\star,\theta_n) + \frac{\eta}{2}\|\theta^\star-\theta_n\|^2_{\mathbb{A}} - \frac{\eta}{2}\|\theta^\star\|^2_{\mathbb{A}}\Big)\ ,
\end{aligned}$$

where we have introduced $c_n = \frac{\exp\left(\sum_{j\in\{0,n\}}\left((\theta^\star+\lambda^\star)^\top\nabla Z_j(\theta^\star+\lambda^\star) - Z_j(\theta^\star+\lambda^\star)\right)\right)}{\int_{\mathbb{R}^d}\exp\left(\sum_{j\in\{0,n\}}((\theta')^\top\nabla Z_j(\theta^\star+\lambda^\star)-Z_j(\theta'))\right)d\theta'}$. Since $\lambda^\star = \theta_n - \theta^\star$, we have

$$c_n = \frac{1}{\int_{\mathbb{R}^d}\exp\Big(-\sum_{j\in\{0,n\}} B_{Z_j}(\theta',\theta^\star+\lambda^\star)\Big)d\theta'} = \frac{1}{\int_{\mathbb{R}^d}\exp\Big(-\sum_{t=1}^n B_{Z_{s_t,a_t}}(\theta',\theta_n) - \frac{\eta}{2}\|\theta'-\theta_n\|^2_{\mathbb{A}}\Big)d\theta'}\ .$$

Therefore, we have from (5) that

$$C_{\mathbb{A},n} := \frac{c_n}{c_0} = \frac{\int_{\mathbb{R}^d}\exp\Big(-\frac{\eta}{2}\|\theta'\|^2_{\mathbb{A}}\Big)d\theta'}{\int_{\mathbb{R}^d}\exp\Big(-\sum_{t=1}^n \mathrm{KL}_{s_t,a_t}(\theta_n,\theta') - \frac{\eta}{2}\|\theta'-\theta_n\|^2_{\mathbb{A}}\Big)d\theta'}$$

An application of Markov's inequality now yields

$$\mathbb{P}\left[\sum_{t=1}^n \mathrm{KL}_{s_t,a_t}(\theta_n,\theta^\star) + \frac{\eta}{2}\|\theta^\star-\theta_n\|^2_{\mathbb{A}} - \frac{\eta}{2}\|\theta^\star\|^2_{\mathbb{A}} \ge \log\left(\frac{C_{\mathbb{A},n}}{\delta}\right)\right] = \mathbb{P}\left[M_n \ge \frac{1}{\delta}\right] \le \delta\cdot\mathbb{E}[M_n] = \delta\ . \qquad (12)$$

Step 3: A Stopped Martingale and Its Control Let N be a stopping time with respect to the filtration $\{\mathcal{F}_n\}_{n=0}^\infty$. 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

$$\mathbb{E}[M_N] = \mathbb{E}\left[\liminf_{n\to\infty} Q_n\right] \le \liminf_{n\to\infty}\mathbb{E}[Q_n] = \liminf_{n\to\infty}\mathbb{E}\left[M_{\min\{N,n\}}\right] \le 1\ ,$$

since the stopped martingale $\left\{M_{\min\{N,n\}}\right\}_{n\ge1}$ 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

$$N = \min\left\{ n \ge 1 : \sum_{t=1}^{n} \mathrm{KL}_{s_t,a_t}(\theta_n,\theta^\star) + \frac{\eta}{2}\left\|\theta^\star-\theta_n\right\|_{\mathbb{A}}^2 - \frac{\eta}{2}\left\|\theta^\star\right\|_{\mathbb{A}}^2 \ge \log\left(\frac{C_{\mathbb{A},n}}{\delta}\right)\right\} ,$$

with $\min\{\emptyset\} := \infty$ by convention. We then have

$$\mathbb{P}\left[\exists\, n \ge 1,\ \sum_{t=1}^{n} \mathrm{KL}_{s_t,a_t}(\theta_n,\theta^\star) + \frac{\eta}{2}\left\|\theta^\star-\theta_n\right\|_{\mathbb{A}}^2 - \frac{\eta}{2}\left\|\theta^\star\right\|_{\mathbb{A}}^2 \ge \log\left(\frac{C_{\mathbb{A},n}}{\delta}\right)\right] = \mathbb{P}\left[N<\infty\right] \le \delta ,$$

which concludes the proof of the first part.

Proof of Second Part: Upper Bound on $C_{\mathbb{A},n}$ First, we have for some $\tilde\theta \in [\theta_n,\theta']_\infty$ that

$$\mathrm{KL}_{s,a}(\theta_n,\theta') \quad = \quad \frac{1}{2}\sum_{i,j=1}^{d}(\theta'-\theta_n)_i \varphi(s,a)^\top A_i^\top \mathbb{C}_{s,a}^{\tilde\theta}\left[\psi(s')\right] A_j\varphi(s,a)(\theta'-\theta_n)_j \ . \tag{13}$$

Now (13) implies that

$$\sum_{t=1}^{n}\mathrm{KL}_{s_t,a_t}(\theta_n,\theta') \le \frac{\beta}{2}\sum_{t=1}^{n}\sum_{i,j=1}^{d}(\theta'-\theta_n)_i\varphi(s_t,a_t)^\top A_i^\top A_j\varphi(s_t,a_t)(\theta'-\theta_n)_j = \frac{\beta}{2}\left\|\theta'-\theta_n\right\|^2_{\sum_{t=1}^n G_{s_t,a_t}} ,$$

where $\beta := \sup_{\theta,s,a}\lambda_{\max}\left(\mathbb{C}_{s,a}^{\theta}[\psi(s')]\right)$ and $\forall i,j\le d,\ \ (G_{s,a})_{i,j} := \varphi(s,a)^\top A_i^\top A_j\varphi(s,a)$. Therefore, we obtain

$$\begin{aligned} C_{\mathbb{A},n} &\le \frac{\int_{\mathbb{R}^d}\exp\Big(-\frac{\eta}{2}\left\|\theta'\right\|_{\mathbb{A}}^2\Big)d\theta'}{\int_{\mathbb{R}^d}\exp\Big(-\frac{1}{2}\left\|\theta'-\theta_n\right\|^2_{(\beta\sum_{t=1}^n G_{s_t,a_t}+\eta\mathbb{A})}\Big)d\theta'} \\ &= \frac{(2\pi)^{d/2}}{\det(\eta\mathbb{A})^{1/2}}\cdot\frac{\det(\beta\sum_{t=1}^n G_{s_t,a_t}+\eta\mathbb{A})^{1/2}}{(2\pi)^{d/2}} = \det\Big(I+\beta\eta^{-1}\mathbb{A}^{-1}\sum_{t=1}^{n}G_{s_t,a_t}\Big) ,\end{aligned}$$

which completes the proof of the second part.

C REGRET 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 Θ_n, i.e.,

$$V^{\pi_t}_{\hat\theta_n,1}(s_1^t) = \max_{\pi\in\Pi}\max_{\theta\in\Theta_n} V^{\pi}_{\theta,1}(s_1^t) ,$$

where s_1^t denotes the starting state at episode t. Therefore, as long as the true parameter $\theta^\star$ belongs to Θ_n, $V^{\pi_t}_{\hat\theta_n,1}(s_1^t)$ gives an optimistic estimate of the value $V^{\pi^\star}_{\theta^\star,1}(s_1^t)$ of the episode, i.e.,

$$V^{\pi_t}_{\hat\theta_n,1}(s_1^t) \ge V^{\pi^\star}_{\theta^\star,1}(s_1^t) \ . \tag{14}$$

An application of 1 implies that with probability at least $1-\delta/2$, $\theta^\star\in\Theta_n$ across all episodes. We then have from (14) that with probability at least $1-\delta/2$, the cumulative regret is controlled by

$$\mathcal{R}(N) \le \sum_{t=1}^{T}\Big(V^{\pi_t}_{\hat\theta_n,1}(s_1^t) - V^{\pi_t}_{\theta^\star,1}(s_1^t)\Big) , \tag{15}$$

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 $\mathcal{T}^{\pi}_{\theta,h} : (\mathcal{S}\to\mathbb{R})\to(\mathcal{S}\to\mathbb{R})$ is defined for all $s\in\mathcal{S}$ and $h\in[H]$ as

$$\mathcal{T}^{\pi}_{\theta,h}\left(V\right)(s) = R\left(s,\pi(s,h)\right) + \mathbb{E}^{\theta}_{s,\pi(s,h)}\left[V\right] ,$$

where $V:\mathcal{S}\to\mathbb{R}$. By the Bellman equation, we have

$$V^{\pi}_{\theta,h}(s) = \mathcal{T}^{\pi}_{\theta,h}\left(V^{\pi}_{\theta,h+1}\right)(s), \quad \forall h\in[H] \quad (\text{with } V^{\pi}_{\theta,H+1}(s) := 0).$$

Following, e.g., Chowdhury and Gopalan (2019), a recursive application of Bellman equation now yields

$$V^{\pi_t}_{\hat{\theta}_n,1}(s_1^t) - V^{\pi_t}_{\theta^\star,1}(s_1^t) = \sum_{h=1}^{H}\left(\mathcal{T}^{\pi_t}_{\hat{\theta}_n,h}\left(V^{\pi_t}_{\hat{\theta}_n,h+1}\right)(s_h^t) - \mathcal{T}^{\pi_t}_{\theta^\star,h}\left(V^{\pi_t}_{\hat{\theta}_n,h+1}\right)(s_h^t) + m_h^t\right) \;,$$

where $m_h^t = \mathbb{E}^{\theta^\star}_{s_h^t,a_h^t}\left[V^{\pi_t}_{\hat{\theta}_n,h+1}(s_{h+1}^t) - V^{\pi_t}_{\theta^\star,h+1}(s_{h+1}^t)\right] - \left(V^{\pi_t}_{\hat{\theta}_n,h+1}(s_{h+1}^t) - V^{\pi_t}_{\theta^\star,h+1}(s_{h+1}^t)\right)$. Note that $\{m_h^t\}_{t,h}$ is a martingale sequence satisfying $|m_h^t| \le 2H$. Therefore, by the Azuma-Hoeffding inequality (Boucheron et al., 2013), with probability at least $1-\delta/2$, we obtain

$$\sum_{t=1}^{T}\sum_{h=1}^{H} m_h^t \le 2H\sqrt{2TH\ln(2/\delta)} = 2H\sqrt{2N\ln(2/\delta)} \;.$$

Then, using a union bound argument along with (15), the cumulative regret can be upper bounded with probability at least $1-\delta$ as

$$\mathcal{R}(N) \le \sum_{t=1}^{T}\sum_{h=1}^{H}\left(\mathcal{T}^{\pi_t}_{\hat{\theta}_n,h}\left(V^{\pi_t}_{\hat{\theta}_n,h+1}\right)(s_h^t) - \mathcal{T}^{\pi_t}_{\theta^\star,h}\left(V^{\pi_t}_{\hat{\theta}_n,h+1}\right)(s_h^t)\right) + 2H\sqrt{2N\ln(2/\delta)} \;. \tag{16}$$

We now proceed to bound the first term in (16). Since $V^{\pi_t}_{\hat{\theta}_n,h+1}(s) \le H,\ \forall s$, we have its span $\mathbb{S}\left(V^{\pi_t}_{\hat{\theta}_n,h+1}\right) \le H$ and variance $\mathbb{V}^{\theta}_{s_h^t,a_h^t}\left[V^{\pi_t}_{\hat{\theta}_n,h+1}\right] \le H^2,\ \forall\theta,\ \forall(s,a)$. Therefore, we obtain

$$\begin{aligned}
&\mathcal{T}^{\pi_t}_{\hat{\theta}_n,h}\left(V^{\pi_t}_{\hat{\theta}_n,h+1}\right)(s_h^t) - \mathcal{T}^{\pi_t}_{\theta^\star,h}\left(V^{\pi_t}_{\hat{\theta}_n,h+1}\right)(s_h^t)\\
&= \mathbb{E}^{\hat{\theta}_n}_{s_h^t,a_h^t}\left[V^{\pi_t}_{\hat{\theta}_n,h+1}\right] - \mathbb{E}^{\theta^\star}_{s_h^t,a_h^t}\left[V^{\pi_t}_{\hat{\theta}_n,h+1}\right]\\
&= \mathbb{E}^{\hat{\theta}_n}_{s_h^t,a_h^t}\left[V^{\pi_t}_{\hat{\theta}_n,h+1}\right] - \mathbb{E}^{\theta_n}_{s_h^t,a_h^t}\left[V^{\pi_t}_{\hat{\theta}_n,h+1}\right] + \mathbb{E}^{\theta_n}_{s_h^t,a_h^t}\left[V^{\pi_t}_{\hat{\theta}_n,h+1}\right] - \mathbb{E}^{\theta^\star}_{s_h^t,a_h^t}\left[V^{\pi_t}_{\hat{\theta}_n,h+1}\right]\\
&\le H\sqrt{2\ \mathrm{KL}_{s_h^t,a_h^t}\left(\theta_n,\hat{\theta}_n\right)} + H\sqrt{2\ \mathrm{KL}_{s_h^t,a_h^t}\left(\theta_n,\theta^\star\right)} + \frac{2H}{3}\mathrm{KL}_{s_h^t,a_h^t}\left(\theta_n,\theta^\star\right) \;,
\end{aligned}$$

where the last step follows from the transportation inequalities (Lemma 1). We then obtain from 16 that

$$\mathcal{R}(N) \le H\sum_{t=1}^{T}\sum_{h=1}^{H}\left(\sqrt{2\ \mathrm{KL}_{s_h^t,a_h^t}(\theta_n,\hat{\theta}_n)} + \sqrt{2\ \mathrm{KL}_{s_h^t,a_h^t}(\theta_n,\theta^\star)} + \frac{2}{3}\mathrm{KL}_{s_h^t,a_h^t}(\theta_n,\theta^\star)\right) + 2H\sqrt{2N\ln(2/\delta)} \;. \tag{17}$$

Step 3: Sum of KL Divergences Along the Transition Trajectory First, we obtain from (13) that

$$\forall(s,a)\in\mathcal{S}\times\mathcal{A},\quad \forall\theta,\theta'\in\mathbb{R}^d,\quad \frac{\alpha}{2}\left\|\theta'-\theta\right\|^2_{G_{s,a}} \le \mathrm{KL}_{s,a}(\theta,\theta') \le \frac{\beta}{2}\left\|\theta'-\theta\right\|^2_{G_{s,a}} \;,$$

where $\alpha := \inf_{\theta,s,a}\lambda_{\min}\left(\mathbb{C}^\theta_{s,a}[\psi(s')]\right)$, $\beta := \sup_{\theta,s,a}\lambda_{\max}\left(\mathbb{C}^\theta_{s,a}[\psi(s')]\right)$, and $\forall i,j\le d,\ (G_{s,a})_{i,j} := \varphi(s,a)^\top A_i^\top A_j\varphi(s,a)$. We then have

$$\forall(s,a),\quad \forall\theta,\quad \mathrm{KL}_{s,a}(\theta_n,\theta) \le \frac{\beta}{2}\left\|\theta-\theta_n\right\|^2_{G_{s,a}} \le \beta\left\|\overline{G}_n^{-1/2}G_{s,a}\overline{G}_n^{-1/2}\right\|\frac{1}{2}\left\|\theta-\theta_n\right\|^2_{\overline{G}_n} \;,$$

where $\overline{G}_n \equiv \overline{G}_{(t-1)H} := G_n + \alpha^{-1}\eta\mathbb{A}$ and $G_n \equiv G_{(t-1)H} := \sum_{\tau=1}^{t-1}\sum_{h=1}^{H}G_{s_h^\tau,a_h^\tau}$. Furthermore, note that

$$\frac{1}{2}\left\|\theta-\theta_n\right\|^2_{\overline{G}_n} = \frac{\alpha^{-1}\eta}{2}\left\|\theta-\theta_n\right\|^2_{\mathbb{A}} + \sum_{\tau=1}^{t-1}\sum_{h=1}^{H}\frac{1}{2}\left\|\theta-\theta_n\right\|^2_{G_{s_h^\tau,a_h^\tau}} \le \alpha^{-1}\left(\frac{\eta}{2}\left\|\theta-\theta_n\right\|^2_{\mathbb{A}} + \sum_{\tau=1}^{t-1}\sum_{h=1}^{H}\mathrm{KL}_{s_h^\tau,a_h^\tau}(\theta_n,\theta)\right) \;.$$

Therefore, for any $\theta\in\Theta_n$, we obtain

$$\forall(s,a),\quad \mathrm{KL}_{s,a}(\theta_n,\theta) \le \frac{\beta}{\alpha}\cdot\beta_n(\delta)\left\|\overline{G}_n^{-1/2}G_{s,a}\overline{G}_n^{-1/2}\right\| = \frac{\beta}{\alpha}\cdot\beta_n(\delta)\left\|\overline{G}_n^{-1}G_{s,a}\right\| \;, \tag{18}$$

where $\beta_n(\delta) \equiv \beta_{(t-1)H}(\delta) = \frac{\eta}{2}B^2_{\mathbb{A}} + \log\left(2C_{\mathbb{A},(t-1)H}/\delta\right)$.

Now, since G_n is positive semi-definite, we have $\overline{G}_n \succeq \alpha^{-1}\eta\mathbb{A}$, and thus, in turn

$$\left\|\overline{G}_n^{-1}G_{s,a}\right\| \le \frac{\alpha}{\eta}\left\|\mathbb{A}^{-1}G_{s,a}\right\| \le \frac{\alpha B_{\varphi,\mathbb{A}}}{\eta} \;,\ \forall(s,a) \;,$$

where $B_{\varphi,\mathbb{A}} := \sup_{s,a} \left\|\mathbb{A}^{-1}G_{s,a}\right\|$. This further yields

$$\left\| I + \overline{G}_n^{-1}\sum_{h=1}^{H} G_{s_h^t,a_h^t}\right\| \le 1 + \sum_{h=1}^{H}\left\|\overline{G}_n^{-1}G_{s_h^t,a_h^t}\right\| \le 1 + \frac{\alpha B_{\varphi,\mathbb{A}}H}{\eta}\ . \tag{19}$$

Now, we define $\overline{G}_{n+H} := \overline{G}_n + \sum_{h=1}^{H} G_{s_h^t,a_h^t}$. Hence, $\overline{G}_{n+H}^{-1}G_{s,a} = \left(I + \overline{G}_n^{-1}\sum_{h=1}^{H} G_{s_h^t,a_h^t}\right)^{-1}\overline{G}_n^{-1}G_{s,a}$. We therefore deduce from (19) that

$$\forall (s,a),\quad \left\|\overline{G}_n^{-1}G_{s,a}\right\| = \left\|\left(I + \overline{G}_n^{-1}\sum_{h=1}^{H} G_{s_h^t,a_h^t}\right)\overline{G}_{n+H}^{-1}G_{s,a}\right\| \le \left(1 + \frac{\alpha B_{\varphi,\mathbb{A}}H}{\eta}\right)\left\|\overline{G}_{n+H}^{-1}G_{s,a}\right\|\ . \tag{20}$$

Now see that

$$\sum_{t=1}^{T}\sum_{h=1}^{H}\left\|\overline{G}_{n+H}^{-1}G_{s_h^t,a_h^t}\right\| \le \sum_{t=1}^{T}\sum_{h=1}^{H}\operatorname{tr}\left(\overline{G}_{n+H}^{-1}G_{s_h^t,a_h^t}\right) = \sum_{t=1}^{T}\operatorname{tr}\left(\overline{G}_{n+H}^{-1}(\overline{G}_{n+H}-\overline{G}_n)\right) \le \sum_{t=1}^{T}\log\frac{\det(\overline{G}_{n+H})}{\det(\overline{G}_n)}\ ,$$

where we have used that for two positive definite matrices A and B such that $A-B$ is positive semi-definite, $\operatorname{tr}(A^{-1}(A-B)) \le \log\frac{\det(A)}{\det(B)}$. We can now control the R.H.S. of the above equation, as

$$\sum_{t=1}^{T}\log\frac{\det(\overline{G}_{n+H})}{\det(\overline{G}_n)} = \sum_{t=1}^{T}\log\frac{\det(\overline{G}_{tH})}{\det(\overline{G}_{(t-1)H})} = \log\frac{\det(\overline{G}_{TH})}{\det(\overline{G}_0)} = \log\frac{\det(\overline{G}_N)}{\det(\alpha^{-1}\eta\mathbb{A})} = \log\det\left(I + \alpha\eta^{-1}\mathbb{A}^{-1}G_N\right)\ .$$

Therefore, we have from (20) and that

$$\sum_{t=1}^{T}\sum_{h=1}^{H}\left\|\overline{G}_n^{-1}G_{s_h^t,a_h^t}\right\| \le \left(1 + \frac{\beta B_{\varphi,\mathbb{A}}H}{\eta}\right)\log\det\left(I + \beta\eta^{-1}\mathbb{A}^{-1}G_N\right)\ , \tag{21}$$

where we have used that $\alpha \le \beta$.

It now remains to bound the log determinant term in the above equation. By the trace-determinant inequality, we have

$$\det\left(I + \beta\eta^{-1}\mathbb{A}^{-1}G_n\right) \le \left(\frac{\operatorname{tr}\left(I + \beta\eta^{-1}\mathbb{A}^{-1}G_n\right)}{d}\right)^d \le \left(1 + \frac{\beta\eta^{-1}}{d}\operatorname{tr}\left(\mathbb{A}^{-1}G_n\right)\right)^d\ .$$

Now see that $\operatorname{tr}\left(\mathbb{A}^{-1}G_n\right) \le n\ \sup_{s,a}\operatorname{tr}\left(\mathbb{A}^{-1}G_{s,a}\right) \le dB_{\varphi,\mathbb{A}}\ n$. Therefore, we have

$$\log\det\left(I + \beta\eta^{-1}\mathbb{A}^{-1}G_n\right) \le d\log\left(1 + \beta\eta^{-1}B_{\varphi,\mathbb{A}}\ n\right)\ . \tag{22}$$

This further implies that the confidence radius

$$\beta_n(\delta) \le \frac{\eta}{2}B_{\mathbb{A}}^2 + \log\left(2\det\left(I + \beta\eta^{-1}\mathbb{A}^{-1}G_n\right)/\delta\right) \le \frac{\eta}{2}B_{\mathbb{A}}^2 + d\log\left(1 + \beta\eta^{-1}B_{\varphi,\mathbb{A}}\ n\right) + \log(2/\delta)\ ,$$

which is an increasing function in the total number of steps n, hence, in the number of episodes t. We then have from (18) and (21) that

$$\forall\theta\in\Theta_n,\quad \sum_{t=1}^{T}\sum_{h=1}^{H}\mathrm{KL}_{s_h^t,a_h^t}(\theta_n,\theta) \le \frac{\beta}{\alpha}\left(1 + \frac{\beta B_{\varphi,\mathbb{A}}H}{\eta}\right)\beta_N(\delta)\gamma_N\ , \tag{23}$$

where we define $\gamma_N := d\log\left(1 + \beta\eta^{-1}B_{\varphi,\mathbb{A}}\ N\right)$ and $\beta_N(\delta) := \frac{\eta}{2}B_{\mathbb{A}}^2 + \gamma_N + \log(2/\delta)$.

Final Step: First, an application of Cauchy-Schwartz's inequality yields

$$\forall\theta\in\Theta_n,\quad \sum_{t=1}^{T}\sum_{h=1}^{H}\sqrt{\mathrm{KL}_{s_h^t,a_h^t}(\theta_n,\theta)} \le \sqrt{N\sum_{t=1}^{T}\sum_{h=1}^{H}\mathrm{KL}_{s_h^t,a_h^t}(\theta_n,\theta)} \le \sqrt{\frac{\beta}{\alpha}\left(1 + \frac{\beta B_{\varphi,\mathbb{A}}H}{\eta}\right)\beta_N(\delta)N\gamma_N}\ . \tag{24}$$

At this point, we note that by design, $\hat{\theta}_n \in \Theta_n$ and by Theorem 1, $\theta^\star \in \Theta_n$ with probability at least $1-\delta/2$. We now obtain from (17), (23) and (24) that the cumulative regret

$$\mathcal{R}(N) \le 2H\sqrt{\frac{\beta}{\alpha}\left(1 + \frac{\beta B_{\varphi,\mathbb{A}}H}{\eta}\right)2\beta_N(\delta)N\gamma_N} + 2H\sqrt{2N\ln(2/\delta)} + \frac{2H}{3}\frac{\beta}{\alpha}\left(1 + \frac{\beta B_{\varphi,\mathbb{A}}H}{\eta}\right)\beta_N(\delta)\gamma_N\ ,$$

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 $\tilde{\theta}_n \equiv \tilde{\theta}_{(t-1)H} \sim \mu_n$, where $\mu_n \equiv \mu_{(t-1)H} = \mathbb{P}(\theta^\star \in \cdot|\mathcal{H}_n)$ denotes the posterior distribution of $\theta^\star$, given the history of transitions $\mathcal{H}_n \equiv \mathcal{H}_{(t-1)H} = \{(s_h^\tau, a_h^\tau, s_{h+1}^\tau)_{\tau<t,h\leq H}\}$. A key property of posterior sampling is that for any $\sigma(\mathcal{H}_n)$-measurable function f, we have $\mathbb{E}[f(\tilde{\theta}_n)] = \mathbb{E}[f(\theta^\star)]$ (Osband et al., 2013). This implies that the optimal policy $\pi^\star$ and selected policy π^t are identically distributed conditioned on the history $\mathcal{H}_n$. Therefore, we have $\mathbb{E}\left[V^{\pi^t}_{\tilde{\theta}_n,1}(s_1^t)\right] = \mathbb{E}\left[V^{\pi^\star}_{\theta^\star,1}(s_1^t)\right]$, and thus, in turn, the Bayes regret

$$\mathbb{E}[\mathcal{R}(N)] = \mathbb{E}\left[\sum_{t=1}^T \left(V^{\pi_t}_{\tilde{\theta}_n,1}(s_1^t) - V^{\pi_t}_{\theta^\star,1}(s_1^t)\right)\right] .$$

A recursive application of the Bellman equation now yields a result similar to (16):

$$\mathbb{E}[\mathcal{R}(N)] = \mathbb{E}\left[\sum_{t=1}^T\sum_{h=1}^H \left(\mathcal{T}^{\pi_t}_{\tilde{\theta}_n,h}\left(V^{\pi_t}_{\tilde{\theta}_n,h+1}\right)(s_h^t) - \mathcal{T}^{\pi_t}_{\theta^\star,h}\left(V^{\pi_t}_{\tilde{\theta}_n,h+1}\right)(s_h^t)\right) + \sum_{t=1}^T\sum_{h=1}^H m_h^t\right] ,$$

where $m_h^t = \mathbb{E}^{\theta^\star}_{s_h^t,a_h^t}\left[V^{\pi_t}_{\tilde{\theta}_n,h+1}(s_{h+1}^t) - V^{\pi_t}_{\theta^\star,h+1}(s_{h+1}^t)\right] - \left(V^{\pi_t}_{\tilde{\theta}_n,h+1}(s_{h+1}^t) - V^{\pi_t}_{\theta^\star,h+1}(s_{h+1}^t)\right)$ 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):

$$\mathbb{E}\left[\mathcal{R}(N)\right] \leq H\,\mathbb{E}\left[\sum_{t=1}^T\sum_{h=1}^H \left(\sqrt{2\ \mathrm{KL}_{s_h^t,a_h^t}(\theta_n,\tilde{\theta}_n)} + \sqrt{2\ \mathrm{KL}_{s_h^t,a_h^t}(\theta_n,\theta^\star)} + \frac{2}{3}\mathrm{KL}_{s_h^t,a_h^t}(\theta_n,\theta^\star)\right)\right] , \qquad (25)$$

where $\theta_n \equiv \theta_{(t-1)H}$ 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 $\mathcal{E}^\star = \{\forall t \geq 1, \theta^\star \in \Theta_n\}$ and $\tilde{\mathcal{E}} = \{\forall t \geq 1, \tilde{\theta}_n \in \Theta_n\}$, where $\Theta_n \equiv \Theta_{(t-1)H}$ is confidence set (as constructed by Exp-UCRL) after $n=(t-1)H$ steps. Under the event $\mathcal{E}^\star \cap \tilde{\mathcal{E}}$, we have from (23) and (24) that

$$\sum_{t=1}^T\sum_{h=1}^H \mathrm{KL}_{s_h^t,a_h^t}(\theta_n,\theta^\star) \leq \frac{\beta}{\alpha}\left(1+\frac{\beta B_{\varphi,\mathbb{A}}H}{\eta}\right)\beta_N(\delta)\gamma_N ,$$

$$\sum_{t=1}^T\sum_{h=1}^H \sqrt{\mathrm{KL}_{s_h^t,a_h^t}(\theta_n,\theta^\star)} \leq \sqrt{\frac{\beta}{\alpha}\left(1+\frac{\beta B_{\varphi,\mathbb{A}}H}{\eta}\right)\beta_N(\delta)N\gamma_N} \quad \text{and}$$

$$\sum_{t=1}^T\sum_{h=1}^H \sqrt{\mathrm{KL}_{s_h^t,a_h^t}(\theta_n,\tilde{\theta}_n)} \leq \sqrt{\frac{\beta}{\alpha}\left(1+\frac{\beta B_{\varphi,\mathbb{A}}H}{\eta}\right)\beta_N(\delta)N\gamma_N} .$$

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

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

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

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

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

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

The proof now can be completed by setting $\delta = 1/N$.

E ON THE CHOICE OF PENALTY FUNCTION

In this paper, we have considered the penalty function $\mathrm{pen}(\theta) = \frac{1}{2}\left\|\theta\right\|^2_{\mathbb{A}}$, where $\forall i,j \leq d,\ \mathbb{A}_{i,j} = \mathrm{tr}(A_iA_j^\top)$. We however note that all our results (Theorem 1, 2, 3) hold for any choice of the (regularizing) matrix $\mathbb{A}$. For any such choice of $\mathbb{A}$, we only need to ensure that there exist a known constant $B_{\mathbb{A}}$ such that $\|\theta^\star\|_{\mathbb{A}} \leq B_{\mathbb{A}}$. In fact for our particular choice, as we have seen in Section 4, we obtain $\mathbb{A} = I$ for factored and tabular MDPs and $\mathbb{A} = m_1 I$ 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.