A Some Useful Lemmas

Lemma 8. Let \( V \in S^+(m) \) be positive definite, \( (M_t)_{t=1,2,...} \subset S^+(m) \) be positive semidefinite matrices and define \( V_t = V + \sum_{k=1}^{t-1} M_s, t = 1, 2, \ldots \) If \( \text{trace}(M_t) \leq L^2 \) for all \( t \), then

\[
\sum_{t=1}^{T} \min(1, \| V_t^{-\frac{1}{2}} M_t V_t^{-\frac{1}{2}} \|_2) \leq 2 \left\{ \log \det(V_{T+1}) - \log \det V \right\} \leq 2 \left\{ m \log \left( \frac{\text{trace}(V) + TL^2}{m} \right) - \log \det V \right\} .
\]

Proof. On the one hand, we have

\[
det(V_T) = det(V_{T-1} + M_{T-1}) = det(V_{T-1}(I + V_{T-1}^{-\frac{1}{2}} M_{T-1} V_{T-1}^{-\frac{1}{2}}))
\]

\[
= det(V_{T-1}) det(I + V_{T-1}^{-\frac{1}{2}} M_{T-1} V_{T-1}^{-\frac{1}{2}})
\]

\[
\vdots
\]

\[
= det(V) \prod_{t=1}^{T-1} det(I + V_t^{-\frac{1}{2}} M_t V_t^{-\frac{1}{2}}).
\]

One the other hand, thanks to \( x \leq 2 \log(1 + x) \), which holds for all \( x \in [0, 1] \),

\[
\sum_{t=1}^{T} \min(1, \| V_t^{-\frac{1}{2}} M_t V_t^{-\frac{1}{2}} \|_2) \leq 2 \sum_{t=1}^{T} \log(1 + \| V_t^{-\frac{1}{2}} M_t V_t^{-\frac{1}{2}} \|_2)
\]

\[
\leq 2 \sum_{t=1}^{T} \log(det(I + V_t^{-\frac{1}{2}} M_t V_t^{-\frac{1}{2}}))
\]

\[
= 2(\log(det V_{T+1}) - \log(det V)) ,
\]

where the second inequality follows since \( V_t^{-\frac{1}{2}} M_t V_t^{-\frac{1}{2}} \) is positive semidefinite, hence all eigenvalues of \( I + V_t^{-\frac{1}{2}} M_t V_t^{-\frac{1}{2}} \) are above one and the largest eigenvalue of \( I + V_t^{-\frac{1}{2}} M_t V_t^{-\frac{1}{2}} \) is \( 1 + \| V_t^{-\frac{1}{2}} M_t V_t^{-\frac{1}{2}} \|_2 \), proving the first inequality. For the second inequality, note that for any positive definite matrix \( S \in S^+(m) \), \( \log det S \leq m \log(\text{trace}(S)/m) \). Applying this to \( V_T \) and using the condition that \( \text{trace}(M_t) \leq L^2 \), we get \( \log det V_T \leq m \log((\text{trace}(V) + TL^2)/m) \). Plugging this into the previous upper bound, we get the second part of the statement. \( \square \)

Lemma 9 (Lemma 11 of Abbasi-Yadkori and Szepesvári (2011)). Let \( A \in \mathbb{R}^{m \times m} \) and \( B \in \mathbb{R}^{m \times m} \) be positive semidefinite matrices such that \( A \succ B \). Then, we have

\[
\sup_{X \neq 0} \frac{\| X^\top A X \|_2}{\| X^\top B X \|_2} \leq \frac{\det(A)}{\det(B)} .
\]

B Proofs

Proof of Proposition 1. Note that if ACOE (1) holds for \( h \), then for any constant \( C \), it also holds that

\[
J(\Theta) + (h(x, \Theta) + C) = \min_{a \in A} \left\{ \ell(x, a) + \int (h(y, \Theta) + C) p(dy \mid x, a, \Theta) \right\} .
\]

As by our assumption, the value function is bounded from below, we can choose \( C \) such that the \( h'(\cdot, \Theta) = h(\cdot, \Theta) + C \) is nonnegative valued. In fact, if \( h \) assumes a minimizer \( x_0 \), by this reasoning, without loss of generality, we can assume that \( h(x_0) = 0 \) and so for any \( x \in \mathcal{X} \), \( 0 \leq h(x) = h(x) - h(x_0) \leq B \| x - x_0 \| \leq BX \). The argument trivially extends to the general case when \( h \) may fail to have a minimizer over \( \mathcal{X} \). \( \square \)
Proof of Theorem 2. The proof follows that of the main result of Abbasi-Yadkori and Szepesvári (2011). First, we decompose the regret into a number of terms, which are then bound one by one. Define \( \overline{x}_{t+1}^a = f(x_t, a, \tilde{\Theta}_t, z_{t+1}) \), where \( f \) is the map of Assumption A1 and let \( h_t(x) = h(x, \tilde{\Theta}_t) \) be the solution of the ACOE underlying \( p(\cdot | x, a, \tilde{\Theta}_t) \). By Assumption A3 (i), \( h_t \) exists and \( h_t(x) \in [0, H] \) for any \( x \in \mathcal{X} \). By Assumption A1, for any \( g \in L^1(p(\cdot | x, a, \tilde{\Theta}_t)) \), \( \int g(dy)p(dy|x, a, \tilde{\Theta}_t) = \mathbb{E} \left[ g(\overline{x}_{t+1}^a) | \mathcal{F}_t, \tilde{\Theta}_t \right] \). Hence, from (1) and (2),

\[
J(\tilde{\Theta}_t) + h_t(x_t) = \min_{a \in A} \left\{ \ell(x_t, a) + \mathbb{E} \left[ h_t(\overline{x}_{t+1}^a) | \mathcal{F}_t, \tilde{\Theta}_t \right] \right\} \\
\geq \ell(x_t, a_t) + \mathbb{E} \left[ h_t(\overline{x}_{t+1}^a) | \mathcal{F}_t, \tilde{\Theta}_t \right] - \sigma_t \\
= \ell(x_t, a_t) + \mathbb{E} \left[ h_t(x_t+1+\epsilon_t) | \mathcal{F}_t, \tilde{\Theta}_t \right] - \sigma_t,
\]

where \( \epsilon_t = \overline{x}_{t+1}^a - x_{t+1} \). As \( J(\cdot) \) is a deterministic function and conditioned on \( \mathcal{F}_1, \tilde{\Theta}_t \) and \( \Theta_1 \) have the same distribution,

\[
R(T) = \sum_{t=1}^{T} \mathbb{E} [\ell(x_t, a_t) - J(\Theta_1)] = \sum_{t=1}^{T} \mathbb{E} \left[ \mathbb{E} [\ell(x_t, a_t) - J(\Theta_1) | \mathcal{F}_t] \right] \\
= \sum_{t=1}^{T} \mathbb{E} \left[ \ell(x_t, a_t) - J(\tilde{\Theta}_t) | \mathcal{F}_t \right] = \sum_{t=1}^{T} \mathbb{E} \left[ \ell(x_t, a_t) - J(\tilde{\Theta}_t) \right] \\
\leq \sum_{t=1}^{T} \mathbb{E} \left[ h_t(x_t) - \mathbb{E} \left[ h_t(x_t+1+\epsilon_t) | \mathcal{F}_t, \tilde{\Theta}_t \right] \right] + \sum_{t=1}^{T} \mathbb{E} [\sigma_t] \\
= \sum_{t=1}^{T} \mathbb{E} [h_t(x_t) - h_t(x_t+1+\epsilon_t)] + \sum_{t=1}^{T} \mathbb{E} [\sigma_t].
\]

Let \( \Sigma_T = \sum_{t=1}^{T} \mathbb{E} [\sigma_t] \) be the total error due to the approximate optimal control oracle. Thus, we can bound the regret using

\[
R(T) \leq \Sigma_T + \mathbb{E} \left[ h_1(x_1) - h_{T+1}(x_{T+1}) \right] + \sum_{t=1}^{T} \mathbb{E} \left[ h_{t+1}(x_{t+1}) - h_t(x_t+1+\epsilon_t) \right] \\
\leq \Sigma_T + H + \sum_{t=1}^{T} \mathbb{E} \left[ h_{t+1}(x_{t+1}) - h_t(x_t+1+\epsilon_t) \right],
\]

where the second inequality follows because \( h_1(x_1) \leq H \) and \( -h_{T+1}(x_{T+1}) \leq 0 \). Let \( A_t \) denote the event that the algorithm has changed its policy at time \( t \). We can write

\[
R(T) - (\Sigma_T + H) \leq \sum_{t=1}^{T} \mathbb{E} \left[ h_{t+1}(x_{t+1}) - h_t(x_t+1+\epsilon_t) \right] \\
= \sum_{t=1}^{T} \mathbb{E} \left[ h_{t+1}(x_{t+1}) - h_t(x_{t+1}) \right] + \sum_{t=1}^{T} \mathbb{E} \left[ h_t(x_{t+1}) - h_t(x_{t+1}+\epsilon_t) \right] \\
\leq 2H \sum_{t=1}^{T} \mathbb{E} \left[ 1 \{ A_t \} \right] + B \sum_{t=1}^{T} \mathbb{E} \left[ \epsilon_t \right],
\]

where we used again that \( 0 \leq h_t(x) \leq H \), and also Assumption A3 (ii). Define

\[
R_1 = H \sum_{t=1}^{T} \mathbb{E} \left[ 1 \{ A_t \} \right] , \quad R_2 = B \sum_{t=1}^{T} \mathbb{E} \left[ \epsilon_t \right].
\]

It remains to bound \( R_2 \) and to show that the number of switches is small.
Bounding $R_2$ Let $\tau_t \leq t$ be the last round before time step $t$ when the policy is changed. So $\tilde{\Theta}_t = \tilde{\Theta}_{\tau_t}$. Letting $M_t = M(x_t, a_t)$, by Assumption A1,

$$\mathbb{E}[\|\epsilon_t\|] \leq \mathbb{E}\left[\|\tilde{\Theta}_t - \Theta_*\|_{M_t}\right].$$

Further,

$$\|\tilde{\Theta}_t - \Theta_*\|_{M_t} \leq \|\tilde{\Theta}_t - \tilde{\Theta}_{\tau_t}\|_{M_t} + \|\tilde{\Theta}_{\tau_t} - \Theta_*\|_{M_t}.$$ 

For $\Theta \in \{\tilde{\Theta}_{\tau_t}, \Theta_*\}$ we have that

$$\|\Theta - \tilde{\Theta}_{\tau_t}\|_{M_t}^2 = \|(\Theta - \tilde{\Theta}_{\tau_t})^\top M_t(\Theta - \tilde{\Theta}_{\tau_t})\|_2^2 = \|\left(\Theta - \tilde{\Theta}_{\tau_t}\right)^\top V_t^{-1/2} V_t^{-1/2} M_t V_t^{-1/2} V_t^{-1/2} (\Theta - \tilde{\Theta}_{\tau_t})\|_2^2 \leq \|\left(\Theta - \tilde{\Theta}_{\tau_t}\right)^\top V_t^{-1/2} M_t V_t^{-1/2}\|_2^2 = \|\Theta - \tilde{\Theta}_{\tau_t}\|_2^2 \|V_t^{-1/2}\|_{M_t}^2,$$

where the last inequality follows because $\|\cdot\|_2$ is an induced norm and induced norms are sub-multiplicative. Hence, we have that

$$\sum_{t=1}^T \mathbb{E}\left[\|\Theta - \tilde{\Theta}_{\tau_t}\|_{M_t}\right] \leq \mathbb{E}\left[\sum_{t=1}^T \left\|\left(\Theta - \tilde{\Theta}_{\tau_t}\right)^\top V_t^{-1/2}\right\|_2 \left\|V_t^{-1/2}\right\|_{M_t}\right] \leq \mathbb{E}\left[\sum_{t=1}^T \left\|\left(\Theta - \tilde{\Theta}_{\tau_t}\right)^\top V_t^{-1/2}\right\|_2^2 \|V_t^{-1/2}\|_{M_t}\right] \leq \mathbb{E}\left[\sum_{t=1}^T \left\|\Theta - \tilde{\Theta}_{\tau_t}\right\|_2^2 \left\|V_t^{-1/2}\right\|_{M_t}\right],$$

where the first inequality uses Hölder’s inequality, and the last two inequalities use Cauchy-Schwarz. By Lemma 8 in Appendix A, using Assumption A4, we have that

$$\sum_{t=1}^T \min\left(1, \|V_t^{-1/2}\|_{M_t}^2\right) \leq 2m \log \left(\frac{\text{trace}(V) + T\Phi^2}{m}\right).$$

Denoting by $\lambda_{\text{min}}(V)$ the minimum eigenvalue of $V$, a simple argument shows $\left\|V_t^{-1/2}\right\|_{M_t}^2 \leq \|M_t\|_2 / \lambda_{\text{min}}(V) \leq \Phi^2 / \lambda_{\text{min}}(V)$, where in the second inequality we used Assumption A4 again. Hence,

$$\sum_{t=1}^T \|V_t^{-1/2}\|_{M_t}^2 \leq \sum_{t=1}^T \min\left(\Phi^2 / \lambda_{\text{min}}(V), \left\|V_t^{-1/2}\right\|_{M_t}^2\right) \leq \sum_{t=1}^T \max\left(1, \Phi^2 / \lambda_{\text{min}}(V)\right) \min\left(1, \|V_t^{-1/2}\|_{M_t}^2\right).$$

Thus,

$$\sum_{t=1}^T \mathbb{E}\left[\|\Theta - \tilde{\Theta}_{\tau_t}\|_{M_t}^2\right] \leq \mathbb{E}\left[2m \max\left(1, \frac{\Phi^2}{\lambda_{\text{min}}(V)}\right) \log \left(\frac{\text{trace}(V) + T\Phi^2}{m}\right)\right] \times \mathbb{E}\left[\sum_{t=1}^T \left\|\left(\Theta - \tilde{\Theta}_{\tau_t}\right)^\top V_t^{-1/2}\right\|_2^2\right].$$

By Lemma 9 of Appendix A and the choice of $\tau_t$, we have that

$$\left\|\left(\Theta - \tilde{\Theta}_{\tau_t}\right)^\top V_t^{-1/2}\right\|_2 \leq \sqrt{\frac{\det(V_t)}{\det(V_{\tau_t})}} \left\|\left(\Theta - \tilde{\Theta}_{\tau_t}\right)^\top V_{\tau_t}^{-1/2}\right\|_2 \leq \sqrt{2} \left\|\left(\Theta - \tilde{\Theta}_{\tau_t}\right)^\top V_{\tau_t}^{-1/2}\right\|_2.$$

(5)
Thus,
\[
\mathbb{E} \left[ \sum_{t=1}^{T} \left\| (\Theta - \tilde{\Theta}_{\tau_t})^T V_{t+1/2} \right\|_2^2 \right] \leq 2 \mathbb{E} \left[ \sum_{t=1}^{T} \left\| (\Theta - \tilde{\Theta}_{\tau_t})^T V_{t+1/2} \right\|_2^2 \right] \tag{by (5)}
= 2 \mathbb{E} \left[ \sum_{t=1}^{T} \mathbb{E} \left[ \left\| (\Theta - \tilde{\Theta}_{\tau_t})^T V_{t+1/2} \right\|_2^2 \mid F_{\tau_t} \right] \right] \tag{by the tower rule}
\leq 2CT. \tag{by Assumption A2}
\]

Let \( G_T = 2m \max \left( 1, \frac{\Phi^2}{\lambda_{\text{min}}(V_T)} \right) \log \left( \frac{\text{trace}(V) + T\Phi^2}{m} \right) \). Collecting the inequalities, we get
\[
R_2 = B \sum_{t=1}^{T} \mathbb{E} \left[ \left\| (\tilde{\Theta}_{\tau_t} - \Theta_{\star})^T \varphi_t \right\| \right] \leq \sqrt{\mathbb{E}[G_T]} \sqrt{CT}
\leq 4B \sqrt{m \max \left( 1, \frac{\Phi^2}{\lambda_{\text{min}}(V_T)} \right) \log \left( \frac{\text{trace}(V) + T\Phi^2}{m} \right)} \sqrt{CT}.
\]

Bounding \( R_1 \) If the algorithm has changed the policy \( K \) times up to time \( T \), then we should have that \( \det(V_T) \geq 2^K \). On the other hand, from Assumption A4 we have \( \lambda_{\text{max}}(V_T) \leq \text{trace}(V) + (T - 1)\Phi^2 \). Thus, it holds that \( 2^K \leq (\text{trace}(V) + \Phi^2 T)^m \). Solving for \( K \), we get \( K \leq m \log_2(\text{trace}(V) + \Phi^2 T) \). Thus,
\[
R_1 = H \sum_{t=1}^{T} \mathbb{E} [1 \{ A_t \}] \leq H m \log_2(\text{trace}(V) + \Phi^2 T).
\]

Putting together the bounds obtained for \( R_1 \) and \( R_2 \), we get the desired result. \( \qed \)

Proof of Theorem 3. First notice that Theorem 2 continues to hold if Assumption A4 is replaced by the following weaker assumption:

**Assumption A6 (Boundedness Along Trajectories)** There exist \( \Phi > 0 \) such that for all \( t \geq 1 \), \( \mathbb{E} [\text{trace}(M(x_t, a_t))] \leq \Phi^2 \).

The reason this is true is because A4 is used only in a context where \( \mathbb{E} \left[ \log(\text{trace}(V + \sum_{s=1}^{T} M_s)) \right] \) needs to be bounded. Using that log is concave, we get
\[
\mathbb{E} \left[ \log(\text{trace}(V + \sum_{s=1}^{T} M_s)) \right] \leq \log \left( \mathbb{E} \left[ \text{trace}(V + \sum_{s=1}^{T} M_s) \right] \right) \leq \log(\text{trace}(V) + T\Phi^2).
\]

With this observation, the result follows from Theorem 2 applied to Lazy PSRL and \( \{p'(\cdot| x, a, \Theta)\} \) as running Stabilized Lazy PSRL for \( t \) time steps in \( p(\cdot| x, a, \Theta_{\star}) \) results in the same total expected cost as running Lazy PSRL for \( t \) time steps in \( p'(\cdot| x, a, \Theta_{\star}) \) thanks to the definition of Stabilized Lazy PSRL and \( p' \).

Hence, all what remains is to show that the conditions of Theorem 2 are satisfied when it is used with \( \{p'(\cdot| x, a, \Theta)\} \). In fact, A3 and A2 hold true by our assumptions. Let us check Assumption A3 next. Defining \( f'(x, a, \Theta, z) = f(x, a, \Theta, z) \) if \( x \in \mathcal{R} \) and \( f'(x, a, \Theta, z) = f(x, \pi_{\text{stab}}(x), \Theta, z) \) otherwise, we see that \( x_{t+1} = f'(x_t, a_t, \Theta, z_{t+1}) \). Further, defining \( M'(x, a) = M(x, a) \) if \( x \in \mathcal{R} \) and \( M'(x, a) = M(x, \pi_{\text{stab}}(x)) \) otherwise, we see that, thanks to the second part that of A1 applied to \( p(\cdot| x, a, \Theta) \), for \( y = f'(x, a, \Theta, z), y' = f'(x, a, \Theta', z), \) \( \mathbb{E} [\|y - y'\|] \leq \mathbb{E} [\|\Theta - \Theta'\|_{M(x,a)}] \) if \( x \in \mathcal{R} \) and \( \mathbb{E} [\|y - y'\|] \leq \mathbb{E} [\|\Theta - \Theta'\|_{M(x,\pi_{\text{stab}}(x))}] \) otherwise. Hence, \( \mathbb{E} [\|y - y'\|] \leq EE \|\Theta - \Theta'\|_{M'(x,a)} \), thus showing that A1 holds for \( p'(\cdot| x, a, \Theta) \) when \( M \) is replaced by \( M' \). Now, Assumption A6 follows from Assumption A5. \( \qed \)
C Choice of the matrices in the web-server application

Hellerstein et al. (2004) fitted the linear model detailed earlier to an Apache HTTP server and obtained the parameters

\[ A = \begin{pmatrix} 0.54 & -0.11 \\ -0.026 & 0.63 \end{pmatrix}, \quad B = \begin{pmatrix} -85 & 4.4 \\ -2.5 & 2.8 \end{pmatrix} \times 10^{-4}, \]

while the noise standard deviation was measured to be 0.1. Hellerstein et al. found that these parameters provided a reasonable fit to their data. For control purposes, the cost matrices \( Q = \text{diag}(5, 1) \), \( R = \text{diag}(1/5062, 0.1^6) \), taken from Example 6.9 of Aström and Murray (2008), were chosen.