A Some Useful Lemmas

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

$$\sum_{t=1}^{T} \min(1, \|V_t^{-1/2}\|_{M_t}^2) \le 2\left\{\log\det(V_{T+1}) - \log\det V\right\}$$
$$\le 2\left\{m \log\left(\frac{\operatorname{trace}(V) + TL^2}{m}\right) - \log\det V\right\} .$$

Proof. On the one hand, we have

$$\begin{aligned}
\det(V_T) &= \det(V_{T-1} + M_{T-1}) = \det(V_{T-1}(I + V_{T-1}^{-\frac{1}{2}} M_{T-1} V_{T-1}^{-\frac{1}{2}})) \\
&= \det(V_{T-1}) \det(I + V_{T-1}^{-\frac{1}{2}} M_{T-1} V_{T-1}^{-\frac{1}{2}}) \\
&\vdots \\
&= \det(V) \prod_{t=1}^{T-1} \det(I + V_t^{-\frac{1}{2}} M_t V_t^{-\frac{1}{2}}) .
\end{aligned}$$

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

$$\begin{aligned}
\sum_{t=1}^{T} \min(1, \|V_t^{-\frac{1}{2}} M_t V_t^{-\frac{1}{2}}\|_2) &\le 2\sum_{t=1}^{T} \log(1 + \|V_t^{-\frac{1}{2}} M_t V_t^{-\frac{1}{2}}\|_2) \\
&\le 2\sum_{t=1}^{T} \log(\det(I + V_t^{-\frac{1}{2}} M_t V_t^{-\frac{1}{2}})) \\
&= 2(\log(\det V_{T+1}) - \log(\det V)) ,
\end{aligned}$$

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 $S \in \mathbb{S}^+(m)$, $\log\det S \le m\log(\operatorname{trace}(S)/m)$. Applying this to V_T and using the condition that $\operatorname{trace}(M_t) \le L^2$, we get $\log\det V_T \le m\log((\operatorname{trace}(V) + TL^2)/m)$. 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 semi-definite matrices such that $A \succ B$. Then, we have*

$$\sup_{X \ne 0} \frac{\|X^\top A X\|_2}{\|X^\top B X\|_2} \le \frac{\det(A)}{\det(B)} .$$

B Proofs

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

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

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 $x \in \mathcal{X}$, $0 \le h(x) = h(x) - h(x_0) \le B\,\|x - x_0\| \le BX$. 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 $\widetilde{x}_{t+1}^a = f(x_t, a, \widetilde{\Theta}_t, z_{t+1})$, where f is the map of Assumption A1 and let $h_t(x) = h(x, \widetilde{\Theta}_t)$ be the solution of the ACOE underlying $p(\cdot|x, a, \widetilde{\Theta}_t)$. 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 $g \in L^1(p(\cdot|x_t, a, \widetilde{\Theta}_t))$, $\int g(dy)p(dy|x_t, a, \widetilde{\Theta}_t) = \mathbb{E}\left[g(\widetilde{x}_{t+1}^a)|\mathcal{F}_t, \widetilde{\Theta}_t\right]$. Hence, from (1) and (2),

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

where $\epsilon_t = \widetilde{x}_{t+1}^{a_t} - x_{t+1}$. As $J(\cdot)$ is a deterministic function and conditioned on $\mathcal{F}_{\tau_t}$, $\widetilde{\Theta}_t$ and Θ_* have the same distribution,

$$
\begin{aligned}
R(T) &= \sum_{t=1}^T \mathbb{E}\left[\ell(x_t, a_t) - J(\Theta_*)\right] = \sum_{t=1}^T \mathbb{E}\left[\mathbb{E}\left[\ell(x_t, a_t) - J(\Theta_*) \,|\, \mathcal{F}_{\tau_t}\right]\right] \\
&= \sum_{t=1}^T \mathbb{E}\left[\mathbb{E}\left[\ell(x_t, a_t) - J(\widetilde{\Theta}_t) \,|\, \mathcal{F}_{\tau_t}\right]\right] = \sum_{t=1}^T \mathbb{E}\left[\ell(x_t, a_t) - J(\widetilde{\Theta}_t)\right] \\
&\leq \sum_{t=1}^T \mathbb{E}\left[h_t(x_t) - \mathbb{E}\left[h_t(x_{t+1} + \epsilon_t) \,|\, \mathcal{F}_t, \widetilde{\Theta}_t\right]\right] + \sum_{t=1}^T \mathbb{E}\left[\sigma_t\right] \\
&= \sum_{t=1}^T \mathbb{E}\left[h_t(x_t) - h_t(x_{t+1} + \epsilon_t)\right] + \sum_{t=1}^T \mathbb{E}\left[\sigma_t\right] \,.
\end{aligned}
$$

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

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

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

$$
\begin{aligned}
R(T) - (\Sigma_T + H) &\leq \sum_{t=1}^T \mathbb{E}\left[h_{t+1}(x_{t+1}) - h_t(x_{t+1} + \epsilon_t)\right] \\
&= \sum_{t=1}^T \mathbb{E}\left[h_{t+1}(x_{t+1}) - h_t(x_{t+1})\right] + \sum_{t=1}^T \mathbb{E}\left[h_t(x_{t+1}) - h_t(x_{t+1} + \epsilon_t)\right] \\
&\leq 2H \sum_{t=1}^T \mathbb{E}\left[\mathbf{1}\left\{A_t\right\}\right] + B \sum_{t=1}^T \mathbb{E}\left[\|\epsilon_t\|\right] \,,
\end{aligned}
$$

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

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

It remains to bound R_2 and to show that the number of switches is small.

Bounding R_2 Let $\tau_t \le t$ be the last round before time step t when the policy is changed. So $\widetilde{\Theta}_t = \widetilde{\Theta}_{\tau_t}$. Letting $M_t = M(x_t, a_t)$, by Assumption A1,

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

Further,

$$\left\|\widetilde{\Theta}_t - \Theta_*\right\|_{M_t} \le \left\|\widetilde{\Theta}_t - \widehat{\Theta}_t\right\|_{M_t} + \left\|\widehat{\Theta}_t - \Theta_*\right\|_{M_t}.$$

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

$$\begin{aligned}
\left\|\Theta - \widehat{\Theta}_{\tau_t}\right\|_{M_t}^2 &= \left\|(\Theta - \widehat{\Theta}_{\tau_t})^\top M_t (\Theta - \widehat{\Theta}_{\tau_t})\right\|_2 \\
&= \left\|(\Theta - \widehat{\Theta}_{\tau_t})^\top V_t^{\frac{1}{2}} V_t^{-\frac{1}{2}} M_t V_t^{-\frac{1}{2}} V_t^{\frac{1}{2}} (\Theta - \widehat{\Theta}_{\tau_t})\right\|_2 \\
&\le \left\|(\Theta - \widehat{\Theta}_{\tau_t})^\top V_t^{\frac{1}{2}}\right\|_2^2 \left\|V_t^{-\frac{1}{2}} M_t V_t^{-\frac{1}{2}}\right\|_2 = \left\|(\Theta - \widehat{\Theta}_{\tau_t})^\top V_t^{\frac{1}{2}}\right\|_2^2 \left\|V_t^{-\frac{1}{2}}\right\|_{M_t}^2,
\end{aligned}$$

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

$$\begin{aligned}
\sum_{t=1}^T \mathbb{E}\left[\left\|\Theta - \widehat{\Theta}_{\tau_t}\right\|_{M_t}\right] &\le \mathbb{E}\left[\sum_{t=1}^T \left\|(\Theta - \widehat{\Theta}_{\tau_t})^\top V_t^{1/2}\right\|_2 \left\|V_t^{-1/2}\right\|_{M_t}\right] \\
&\le \mathbb{E}\left[\sqrt{\sum_{t=1}^T \left\|(\Theta - \widehat{\Theta}_{\tau_t})^\top V_t^{1/2}\right\|_2^2} \sqrt{\sum_{t=1}^T \left\|V_t^{-1/2}\right\|_{M_t}^2}\right] \\
&\le \sqrt{\mathbb{E}\left[\sum_{t=1}^T \left\|(\Theta - \widehat{\Theta}_{\tau_t})^\top V_t^{1/2}\right\|_2^2\right]} \sqrt{\mathbb{E}\left[\sum_{t=1}^T \left\|V_t^{-1/2}\right\|_{M_t}^2\right]},
\end{aligned}$$

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

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

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

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

Thus,

$$\begin{aligned}
\sum_{t=1}^T \mathbb{E}\left[\left\|\Theta - \widehat{\Theta}_{\tau_t}\right\|_{M_t}^2\right] \le & \sqrt{\mathbb{E}\left[2m \max\left(1, \frac{\Phi^2}{\lambda_{\min}(V)}\right) \log\left(\frac{\operatorname{trace}(V) + T\Phi^2}{m}\right)\right]} \\
& \times \sqrt{\mathbb{E}\left[\sum_{t=1}^T \left\|(\Theta - \widehat{\Theta}_{\tau_t})^\top V_t^{1/2}\right\|_2^2\right]}.
\end{aligned}$$

By Lemma 9 of Appendix A and the choice of τ_t, we have that

$$\left\|(\Theta - \widehat{\Theta}_{\tau_t})^\top V_t^{1/2}\right\|_2 \le \sqrt{\frac{\det(V_t)}{\det(V_{\tau_t})}} \left\|(\Theta - \widehat{\Theta}_{\tau_t})^\top V_{\tau_t}^{1/2}\right\|_2 \le \sqrt{2} \left\|(\Theta - \widehat{\Theta}_{\tau_t})^\top V_{\tau_t}^{1/2}\right\|_2. \tag{5}$$

Thus,

$$
\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\left\|(\Theta-\widehat{\Theta}_{\tau_t})^\top V_t^{1/2}\right\|_2^2\right] &\le 2\mathbb{E}\left[\sum_{t=1}^{T}\left\|(\Theta-\widehat{\Theta}_{\tau_t})^\top V_{\tau_t}^{1/2}\right\|_2^2\right] && \text{(by (5))}\\
&= 2\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{E}\left[\left\|(\Theta-\widehat{\Theta}_{\tau_t})^\top V_{\tau_t}^{1/2}\right\|_2^2 \,\middle|\, \mathcal{F}_{\tau_t}\right]\right] && \text{(by the tower rule)}\\
&\le 2CT\,. && \text{(by Assumption A2)}
\end{aligned}
$$

Let $G_T = 2m\max\left(1, \frac{\Phi^2}{\lambda_{\min}(V)}\right)\log\left(\frac{\operatorname{trace}(V)+T\Phi^2}{m}\right)$. Collecting the inequalities, we get

$$
\begin{aligned}
R_2 = B\sum_{t=1}^{T}\mathbb{E}\left[\left\|(\widetilde{\Theta}_{\tau_t}-\Theta_*)^\top\varphi_t\right\|\right] &\le \sqrt{\mathbb{E}\left[G_T\right]}\sqrt{CT}\\
&\le 4B\sqrt{m\max\left(1,\frac{\Phi^2}{\lambda_{\min}(V)}\right)\log\left(\frac{\operatorname{trace}(V)+T\Phi^2}{m}\right)}\sqrt{CT}\,.
\end{aligned}
$$

Bounding R_1 If the algorithm has changed the policy K times up to time T, then we should have that $\det(V_T)\ge 2^K$. On the other hand, from Assumption A4 we have $\lambda_{\max}(V_T)\le \operatorname{trace}(V)+(T-1)\Phi^2$. Thus, it holds that $2^K \le (\operatorname{trace}(V)+\Phi^2T)^m$. Solving for K, we get $K\le m\log_2(\operatorname{trace}(V)+\Phi^2T)$. Thus,

$$
R_1 = H\sum_{t=1}^{T}\mathbb{E}\left[\mathbf{1}\left\{A_t\right\}\right] \le Hm\log_2(\operatorname{trace}(V)+\Phi^2T)\,.
$$

Putting together the bounds obtained for R_1 and R_2, we get the desired result. □

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 $t\ge 1$, $\mathbb{E}\left[\operatorname{trace}(M(x_t,a_t))\right]\le\Phi^2$.

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

$$
\mathbb{E}\left[\log(\operatorname{trace}(V+\textstyle\sum_{s=1}^{T}M_t))\right] \le \log\left(\mathbb{E}\left[\operatorname{trace}(V+\textstyle\sum_{s=1}^{T}M_t)\right]\right) \le \log(\operatorname{trace}(V)+T\Phi^2).
$$

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 $p(\cdot|x,a,\Theta_*)$ results in the same total expected cost as running Lazy PSRL for t time steps in $p'(\cdot|x,a,\Theta_*)$ 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 $y=f'(x,a,\Theta,z)$, $y'=f'(x,a,\Theta',z)$, $\mathbb{E}\left[\|y-y'\|\right]\le\mathbb{E}\left[\|\Theta-\Theta'\|_{M(x,a)}\right]$ if $x\in\mathcal{R}$ and $\mathbb{E}\left[\|y-y'\|\right]\le\mathbb{E}\left[\|\Theta-\Theta'\|_{M(x,\pi_{\text{stab}}(x))}\right]$ otherwise. Hence, $\mathbb{E}\left[\|y-y'\|\right]\le EE\|\Theta-\Theta'\|_{M'(x,a)}$, thus showing that A1 holds for $p'(\cdot|x,a,\Theta)$ when M is replaced by M'. Now, Assumption A6 follows from Assumption A5.

□

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

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

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 $Q = \mathrm{diag}(5, 1)$, $R = \mathrm{diag}(1/5062, 0.1^6)$, taken from Example 6.9 of Aström and Murray (2008), were chosen.