A Value function estimation

A.1 Proof of Lemma 4.1

To see the identity

$$P_\Phi(\Phi - \overline{\Phi}_+ + \mathbf{W})(h - \widehat{h}) = \Phi(I - \Gamma \otimes \Gamma)^\top \mathrm{VEC}(h - \hat{h}), \tag{13}$$

note that a single element of the vector $(\Phi - \overline{\Phi}_+ + W)(h - \widehat{h})$ can be expressed as

$$(\phi - \mathbf{E}(\phi_+) + \mathrm{VEC}(W))^\top (h - \hat{h}) = \mathrm{VEC}(xx^\top - \Gamma x x^\top \Gamma^\top)^\top (h - \hat{h}) = \mathrm{VEC}(xx^\top)^\top (I - \Gamma \otimes \Gamma)^\top (h - \hat{h}), \tag{14}$$

where we have used the Kronecker product identity $\mathrm{VEC}(\Gamma X \Gamma^\top) = (\Gamma \otimes \Gamma)\mathrm{VEC}(X)$. Thus we have that

$$\left\| \Phi(I - \Gamma \otimes \Gamma)^\top (h - \hat{h}) \right\| \le \left\| P_\Phi(\overline{\Phi}_+ - \Phi_+)\hat{h} \right\| . \tag{15}$$

Next we lower-bound $\left\| (I - \Gamma \otimes \Gamma)^\top (h - \hat{h}) \right\|$. Let $L = H^{1/2}\Gamma H^{-1/2}$ and let $\overline{H} = I - H^{-1/2}\widehat{H}H^{-1/2}$. We have the following:

$$\begin{aligned}
\left\| (I - \Gamma \otimes \Gamma)^\top \mathrm{VEC}(H - \widehat{H}) \right\| &= \left\| H - \widehat{H} - \Gamma^\top (H - \widehat{H})\Gamma \right\|_F \\
&= \left\| H^{1/2}(\overline{H} - L^\top \overline{H} L) H^{1/2} \right\|_F \\
&= \sqrt{\mathrm{tr}(H(\overline{H} - L^\top \overline{H} L) H (\overline{H} - L^\top \overline{H} L))} \\
&\ge \lambda_{\min}(H) \left\| \overline{H} - L^\top \overline{H} L \right\|_F \\
&\ge \lambda_{\min}(M) \left\| \overline{H} - L^\top \overline{H} L \right\|_F .
\end{aligned}$$

where the second-last inequality follows from the fact that $\mathrm{tr}(AB) \ge \lambda_{\min}(A)\,\mathrm{tr}(B)$ for p.s.d matrices A and B (Zhang and Zhang, 2006). Furthermore, using the fact that $\|L\|^2 \le 1 - \lambda_{\min}(M)\,\|H\|^{-1}$, [3]

$$\begin{aligned}
\left\| \overline{H} - L^\top \overline{H} L \right\|_F &= \left\| (I - L \otimes L)^\top \mathrm{VEC}(\overline{H}) \right\| \\
&\ge (1 - \|L\|^2) \left\| \overline{H} \right\|_F \\
&\ge \frac{\lambda_{\min}(M)}{\|H\|} \left\| I - H^{-1/2}\widehat{H}H^{-1/2} \right\|_F \\
&= \frac{\lambda_{\min}(M)}{\|H\|} \sqrt{\mathrm{tr}\left(H^{-1}(H - \widehat{H}) H^{-1} (H - \widehat{H}) \right)} \\
&\ge \lambda_{\min}(M)\, \|H\|^{-2} \left\| H - \widehat{H} \right\|_F .
\end{aligned}$$

Hence we get that

$$\left\| (I - \Gamma \otimes \Gamma)^\top (h - \hat{h}) \right\| \ge \lambda_{\min}(M)^2 \, \|H\|^{-2} \left\| H - \widehat{H} \right\|_F . \tag{16}$$

A.2 Proof of Lemma 4.2

Proof. Let $P_\Psi = \Psi(\Psi^\top \Psi)^{-1}\Psi$ be the orthogonal projector onto Ψ. The true parameters $g = \mathrm{VEC}(G)$ and the estimate $\hat{g} = \mathrm{VEC}(\widehat{G})$ satisfy the following:

$$\Psi \hat{g} = P_\Psi(\mathbf{c} + (\Phi_+ - \mathbf{W})\hat{h}) \tag{17}$$

$$\Psi g = \mathbf{c} + (\overline{\Phi}_+ - \mathbf{W}) h \tag{18}$$

Subtracting the above equations, we have

$$\|\Psi g - \Psi \hat{g}\| = \left\| P_\Psi \left((\overline{\Phi}_+ - \mathbf{W})(h - \hat{h}) + (\overline{\Phi}_+ - \Phi_+)\hat{h} \right) \right\|$$

[3] This can be seen by multiplying the equation $H \succ \Gamma^\top H \Gamma + \lambda_{\min}(M) I$ by $H^{-1/2}$ on both sides.

$$\leq \left\| (\overline{\Phi}_+ - \mathbf{W})(h - \hat{h}) \right\| + \left\| P_\Psi (\overline{\Phi}_+ - \Phi_+)\hat{h} \right\| .$$

Using $\|\Psi v\| \geq \sqrt{\lambda_{\min}(\Psi^\top \Psi)}\, \|v\|$ on the l.h.s., and $\|P_\Psi v\| \leq \left\|\Psi^\top v\right\| / \sqrt{\lambda_{\min}(\Psi^\top \Psi)}$ on the r.h.s.,

$$\left\| G - \widehat{G} \right\|_F = \|g - \hat{g}\| \leq \frac{\left\| (\overline{\Phi}_+ - \mathbf{W})(h - \hat{h}) \right\|}{\sqrt{\lambda_{\min}(\Psi^\top \Psi)}} + \frac{\left\| \Psi^\top (\overline{\Phi}_+ - \Phi_+)\hat{h} \right\|}{\lambda_{\min}(\Psi^\top \Psi)}. \tag{19}$$

Using similar arguments as for $\lambda_{\min}(\Phi^\top \Phi)$ and the fact that actions are randomly sampled, it can be shown that $\lambda_{\min}(\Psi^\top \Psi) = O(\tau)$.

Let $\Sigma_{G,\pi} = A \Sigma_\pi A^\top + B \Sigma_a B^\top$. Assuming that we are close to steady state $x \sim \mathcal{N}(0, \Sigma_\pi)$ each time we take a random action $a \sim \mathcal{N}(0, \Sigma_a)$, the next state is distributed as $x_+ \sim \mathcal{N}(0, \Sigma_{G,\pi} + W)$. Therefore each element of $(\overline{\Phi}_+ - \mathbf{W})(h - \hat{h})$ is bounded as:

$$\begin{aligned} |(\mathbf{E}(\phi_+) - \mathrm{VEC}(W))^\top (h - \hat{h})| &= |\operatorname{tr}\left(\Sigma_{G,\pi}(H - \widehat{H})\right)| \\ &\leq \operatorname{tr}(\Sigma_{G,\pi}) \left\| H - \widehat{H} \right\| , \end{aligned}$$

where we have used the fact that $|\operatorname{tr}(M_1 M_2)| \leq \|M_1\| \operatorname{tr}(M_2)$ for real-valued square matrices M_1 and $M_2 \succ 0$ (see e.g. (Zhang and Zhang, 2006)). Thus, the first term of (19) is bounded as

$$\left\| (\overline{\Phi}_+ - \mathbf{W})(h - \hat{h}) \right\| \leq \operatorname{tr}(\Sigma_{G,\pi}) \left\| H - \widehat{H} \right\| \sqrt{\tau}. \tag{20}$$

To bound the second term, we can again use Lemma 4.8 of Tu and Recht (2017), where the only changes are that we bound $\max_t \|\psi_t\|$ as opposed to $\max_t \|\phi_t\|$, and that we have a different distribution of next-state vectors x_+. Thus, with probability at least $1 - \delta$, the second term scales as

$$\left\| \Psi^\top (\overline{\Phi}_+ - \Phi_+)\hat{h} \right\| = O\left(\sqrt{\tau} \left\| W \widehat{H} \right\|_F (\operatorname{tr}(\Sigma_\pi) + \operatorname{tr}(\Sigma_a)) \|\Sigma_{G,\pi}\|_F \operatorname{polylog}(n^2, 1/\delta, \tau) \right). \tag{21}$$

□

B Analysis of the MFLQ algorithm

B.1 Proof of Lemma 5.1

Proof. Let $G^j = \frac{1}{j}\sum_{i=1}^j G_i$ and $\widehat{G}^j = \frac{1}{j}\sum_{i=1}^j \widehat{G}_i$ be the averages of true and estimated state-action value matrices of policies $K_1, \dots, K_j$, respectively. Let H^j and $\widehat{H}^j$ be the corresponding value matrices. The greedy policy with respect to $\widehat{G}^j$ is given by:

$$\begin{aligned} K_{j+1} &= \arg\min_K \operatorname{tr}\left(x^\top \begin{bmatrix} I & -K^\top \end{bmatrix} \widehat{G}^j \begin{bmatrix} I \\ -K \end{bmatrix} x \right) \\ &= \arg\min_K \operatorname{tr}\left(\widehat{G}^j X_K \right) , \end{aligned} \tag{22}$$

$$\text{where } X_K = \begin{bmatrix} I \\ -K \end{bmatrix} x x^\top \begin{bmatrix} I & -K^\top \end{bmatrix} . \tag{23}$$

Let $|X_K|$ be the matrix obtained from X_K by taking the absolute value of each entry. We have the following:

$$\begin{aligned} \operatorname{tr}(G_j X_{K_{j+1}}) &\leq \operatorname{tr}(\widehat{G}_j X_{K_{j+1}}) + \varepsilon_1 \operatorname{tr}(\mathbf{1}\mathbf{1}^\top |X_{K_{j+1}}|) &(24)\\ &\leq \operatorname{tr}(\widehat{G}_j X_{K_j}) + \varepsilon_1 \mathbf{1}^\top |X_{K_{j+1}}| \mathbf{1} &(25)\\ &\leq \operatorname{tr}(G_j X_{K_j}) + \varepsilon_1 \mathbf{1}^\top (|X_{K_j}| + |X_{K_{j+1}}|)\mathbf{1} &(26)\\ &= x^\top H_j x + \varepsilon_1 \mathbf{1}^\top (|X_{K_j}| + |X_{K_{j+1}}|)\mathbf{1} &(27) \end{aligned}$$

Here, (24) and (26) follow from the error bound,[4] and (27) follows from $\mathrm{tr}(G_j X_{K_j}) = x^\top H_j x$. To see (25), note that K_{i+1} is optimal for $\widehat{G}^i$ and we have:

$$\begin{aligned}\mathrm{tr}(\widehat{G}^j X_{K_{j+1}}) &= \frac{j-1}{j}\,\mathrm{tr}(\widehat{G}^{j-1} X_{K_{j+1}}) + \frac{1}{j}\,\mathrm{tr}(\widehat{G}_j X_{K_{j+1}})\\ &\le \frac{j-1}{j}\,\mathrm{tr}(\widehat{G}^{j-1} X_{K_j}) + \frac{1}{j}\,\mathrm{tr}(\widehat{G}_j X_{K_j})\\ &= \mathrm{tr}(\widehat{G}^j X_{K_j})\,.\end{aligned}$$

Since $\mathrm{tr}(\widehat{G}^{j-1} X_{K_j}) \le \mathrm{tr}(\widehat{G}^{j-1} X_{K_{j+1}})$ it follows that $\mathrm{tr}(\widehat{G}_j X_{K_{j+1}}) \le \mathrm{tr}(\widehat{G}_j X_{K_j})$.

Now note that we can rewrite $\mathrm{tr}(G_j X_{K_{j+1}})$ as a function of H_j as follows:

$$\begin{aligned}\mathrm{tr}(G_j X_{K_{j+1}}) &= x^\top \begin{bmatrix} I & -K_{j+1}^\top \end{bmatrix} G_j \begin{bmatrix} I \\ -K_{j+1} \end{bmatrix} x\\ &= x^\top \begin{bmatrix} I & -K_{j+1}^\top \end{bmatrix} \left(\begin{bmatrix} A^\top \\ B^\top \end{bmatrix} H_j \begin{bmatrix} A & B \end{bmatrix} + \begin{bmatrix} M & 0 \\ 0 & N \end{bmatrix} \right) \begin{bmatrix} I \\ -K_{j+1} \end{bmatrix} x\\ &= x^\top \Big((A - BK_{j+1})^\top H_j (A - BK_{j+1}) \Big) x + \mathrm{tr}\left(\begin{bmatrix} M & 0 \\ 0 & N \end{bmatrix} X_{K_{j+1}} \right).\end{aligned}$$

Letting $\Gamma_j = A - BK_j$, we have that

$$x^\top \Big(\Gamma_{j+1}^\top H_j \Gamma_{j+1} \Big) x + \varepsilon_2 \le x^\top H_j x \tag{28}$$

$$\text{where } \; \varepsilon_2 = x^\top (M + K_{j+1}^\top N K_{j+1}) x - \varepsilon_1 \mathbf{1}^\top (|X_{K_j}| + |X_{K_{j+1}}|)\mathbf{1}.$$

If the estimation error ε_1 is small enough so that $\varepsilon_2 > 0$ for any unit-norm x and all policies, then $H_j \succ \Gamma_{j+1}^\top H_j \Gamma_{j+1}$ and K_{j+1} is stable by a Lyapunov theorem. Since K_1 is stable and H_1 bounded, all policies remain stable.

In order to have $\varepsilon_2 > 0$, it suffices to have

$$\varepsilon_1 < \big((\sqrt{n} + \|K_j\|\sqrt{d})^2 + (\sqrt{n} + \|K_{j+1}\|\sqrt{d})^2 \big)^{-1}.$$

This follows since $M \succ I$, and since for any unit norm vector $x \in \mathbb{S}^n$, $\mathbf{1}^\top x x^\top \mathbf{1} \le n$, with equality achieved by $x = \frac{1}{\sqrt{n}}\mathbf{1}$. Similarly, $\mathbf{1}^\top K x x^\top K^\top \mathbf{1} \le \|K\|^2 d$, and $\mathbf{1}^\top (|X_{K_j}|)\mathbf{1} \le (\sqrt{n} + \|K_j\|\sqrt{d})^2$.

As we will see, we need a smaller estimation error in phase j:

$$\varepsilon_1 < \frac{1}{6C_1 S} \big((\sqrt{n} + \|K_j\|\sqrt{d})^2 + (\sqrt{n} + \|K_{j+1}\|\sqrt{d})^2 \big)^{-1}. \tag{29}$$

Here, C_1 is an upper bound on $\|H_1\|$; note that $H_1 \succ M \succ I$, so $C_1 > 1$. The above condition guarantees that

$$\varepsilon_1 \mathbf{1}^\top (|X_{K_j}| + |X_{K_{j+1}}|)\mathbf{1} \le \frac{1}{6C_1 S}\,.$$

We have that $G_{1,22} \succ N \succ I$ and $G_{1,21} = B^\top H_1 A$. Given that the estimation error (10) is small, we have $\|K_2\| \le 2(\|B^\top H_1 A\| + 1) \le C_K$. Then (10) implies (29) for $j = 1$, and the above argument shows that K_2 is stable.

Next, we show a bound on $\|\Gamma_i^k\|$. Let $L_{i+1} = H_i^{1/2} \Gamma_{i+1} H_i^{-1/2}$. By (28), $M \succ I$, and the error bound,

$$\begin{aligned} H_1 &\succ \Gamma_2^\top H_1 \Gamma_2 + (M + K_2^\top N K_2) - (6C_1 S)^{-1} I\\ I &\succ L_2^\top L_2 + H_1^{-1/2}(M + K_2^\top N K_2) H_1^{-1/2} - (6C_1 S)^{-1} H_1^{-1}\\ &\succ L_2^\top L_2 + H_1^{-1} - (6C_1)^{-1} I \end{aligned}$$

[4]Note that the elementwise max norm of a matrix satisfies $\|G\|_{\max} \le \|G\|_F$.

$$\succ L_2^\top L_2 + (3C_1)^{-1} I - (6C_1)^{-1} I \;.$$

Thus, $\|L_2\| \leq \sqrt{1-(6C_1)^{-1}}$ and we have that

$$\left\|\Gamma_2^k\right\| = \left\|(H_1^{-1/2} L_2 H_1^{1/2})^k\right\| \leq \sqrt{C_1}(1-(6C_1)^{-1})^{k/2}.$$

To show a uniform bound on value functions, we first note that

$$H_2 - H_1 \prec \Gamma_2^\top (H_2 - H_1)\Gamma_2 + (6C_1 S)^{-1} I \;.$$

Using the stability of Γ_2,

$$\begin{aligned}
H_2 - H_1 &\prec (6C_1 S)^{-1} \sum_{k=0}^{\infty} (\Gamma_2^\top)^k \Gamma_2^k \\
\|H_2\| &\leq \|H_1\| + \frac{C_1}{6C_1 S(1-\|L_2\|^2)} \leq (1+S^{-1})C_1 \;.
\end{aligned}$$

Thus $C_2 \leq (1+S^{-1})C_1$, and by repeating the same argument,

$$C_i \leq (1+S^{-1})^i C_1 \leq 3C_1 \;. \tag{30}$$

□

C Regret bound

In this section, we prove Lemma 5.2 by bounding β_T, γ_T, and α_T.

C.1 Bounding β_T

Because we use FTL as our expert algorithm and value functions are quadratic, we can use the following regret bound for the FTL algorithm (Theorem 3.1 in (Cesa-Bianchi and Lugosi, 2006)).

Theorem C.1 (FTL Regret Bound)**.** *Assume that the loss function $f_t(\cdot)$ is convex, is Lipschitz with constant F_1, and is twice differentiable everywhere with Hessian $H \succ F_2 I$. Then the regret of the Follow The Leader algorithm is bounded by*

$$B_T \leq \frac{F_1^2}{2F_2}(1+\log T) \;.$$

Because we execute S policies, each for $\tau = T/S$ rounds (where $\tau = T^{2/3+\xi}$ and $\tau = T^{3/4}$ for MFLQV1 and MFLQV2, respectively),

$$\begin{aligned}
\beta_T &= \sum_{i=1}^{S} \tau \mathbf{E}_{x\sim\mu_\pi}(Q_i(x,\pi_i(x)) - Q_i(x,\pi(x))) \\
&= \tau \sum_{i=1}^{S} \Bigg(\mathbf{E}_{x\sim\mu_\pi}(\widehat{Q}_i(x,\pi_i(x)) - \widehat{Q}_i(x,\pi(x))) \\
&\qquad\qquad + \mathbf{E}_{x\sim\mu_\pi}(Q_i(x,\pi_i(x)) - \widehat{Q}_i(x,\pi_i(x))) \\
&\qquad\qquad + \mathbf{E}_{x\sim\mu_\pi}(\widehat{Q}_i(x,\pi(x)) - Q_i(x,\pi(x))) \Bigg) \\
&\leq C'\sqrt{ST}\log T + \tau \sum_{i=1}^{S} \mathbf{E}_{x\sim\mu_\pi}(\widehat{Q}_i(x,\pi_i(x)) - \widehat{Q}_i(x,\pi(x))) \,,
\end{aligned}$$

where the last inequality holds by Lemma 4.2. Consider the remaining term:

$$E_T = \tau \sum_{i=1}^{S} \mathbf{E}_{x\sim\mu_\pi}(\widehat{Q}_i(x,\pi_i(x)) - \widehat{Q}_i(x,\pi(x))) \;.$$

We bound this term using the FTL regret bound. We show that the conditions of Theorem C.1 hold for the loss function $f_i(K) = \mathbf{E}_{x\sim\mu_\pi}(\widehat{Q}_i(x, Kx))$. Let Σ_π be the covariance matrix of the steady-state distribution $\mu_\pi(x)$. We have that

$$
\begin{aligned}
f_i(K) &= \operatorname{tr}\left(\Sigma_\pi\big(\widehat{G}_{i,11} - K^\top\widehat{G}_{i,21} - \widehat{G}_{i,12}K + K^\top\widehat{G}_{i,22}K\big)\right)\\
\nabla_K f_i(K) &= 2\Sigma_\pi\big(K^\top\widehat{G}_{i,22} - \widehat{G}_{i,12}\big)\\
&= 2\textsc{mat}\big((\widehat{G}_{i,22}\otimes\Sigma_\pi)\textsc{vec}(K)\big) - 2\Sigma_\pi\widehat{G}_{i,12}\\
\nabla^2_{\textsc{vec}(K)} f_i(K) &= 2\widehat{G}_{i,22}\otimes\Sigma_\pi\,.
\end{aligned}
$$

Boundedness and Lipschitzness of the loss function $f_i(K_i)$ follow from the boundedness of policies K_i and value matrix estimates $\widehat{G}_i$. By Lemma 5.1, we have that $\|K_i\| \le C_K$. To bound $\|\widehat{G}_i\|$, note that

$$
\begin{aligned}
G_i &= \begin{pmatrix} A^\top \\ B^\top\end{pmatrix} H_i \begin{pmatrix} A & B\end{pmatrix} + \begin{pmatrix} M & 0\\ 0 & N\end{pmatrix}\\
\|G_i\| &\le C_H(\|A\| + \|B\|)^2 + \|M\| + \|N\| && \text{(Lemma 5.1)}\\
\left\|\widehat{G}_i\right\| &\le \|G_i\| + \varepsilon_1\sqrt{n+d} && \text{(Lemma 4.2)}.
\end{aligned}
$$

The Hessian lower bound is $\nabla^2_{\textsc{vec}(K)} f_i(K) \succ F_2 I$, where F_2 is given by two times the product of the minimum eigenvalues of Σ_π and $\widehat{G}_{i,22}$. For any stable policy $\pi(x) = Kx$, the covariance matrix of the stationary distribution satisfies $\Sigma_\pi \succ W$, and we project the estimates $\widehat{G}_i$ onto the constraint $\widehat{G}_i \succeq \left(\begin{smallmatrix} M & 0\\ 0 & N\end{smallmatrix}\right)$. Therefore the Hessian of the loss is lower-bounded by $2\lambda_{\min}(W)I$. By Theorem C.1, $E_T \le \tau\log S = C''\tau\log T$ for an appropriate constant C''.

C.2 Bounding γ_T

In this section, we bound the average cost of following a stable policy, $\gamma_T = \sum_{t=1}^T(\lambda_\pi - c(x_t, \pi(x_t)))$. Recall that the instantaneous and average costs of following a policy $\pi(x) = -Kx$ can be written as

$$
\begin{aligned}
c(x_t, \pi(x_t)) &= x_t^\top(M + K^\top NK)x_t && (31)\\
\lambda_\pi &= \operatorname{tr}(\Sigma_\pi(M + K^\top NK))\,, && (32)
\end{aligned}
$$

where Σ_π is the steady-state covariance of x_t. Let Σ_t be the covariance of x_t, let $D_t = \Sigma_t^{1/2}(M + K^\top NK)\Sigma_t^{1/2}$, and let $\lambda_t = \operatorname{tr}(D_t)$. To bound γ_T, we start by rewriting the cost terms as follows:

$$
\begin{aligned}
\lambda_\pi - c(x_t, \pi(x_t)) &= \lambda_\pi - \lambda_t + \lambda_t - c(x_t^\pi, \pi(x_t^\pi)) && (33)\\
&= \operatorname{tr}((\Sigma_\pi - \Sigma_t)(M + K^\top NK)) + \left(\operatorname{tr}(D_t) - u_t^\top D_t u_t\right) && (34)
\end{aligned}
$$

where $u_t \sim \mathcal{N}(0, I_n)$ is a standard normal vector.

To bound $\operatorname{tr}((\Sigma_\pi - \Sigma_t)(M + K^\top NK))$, note that $\Sigma_\pi = \Gamma\Sigma_\pi\Gamma^\top + W$ and $\Sigma_t = \Gamma\Sigma_{t-1}\Gamma^\top + W$. Subtracting the two equations and recursing,

$$
\Sigma_\pi - \Sigma_t = \Gamma(\Sigma_\pi - \Sigma_{t-1})\Gamma^\top = \Gamma^t(\Sigma_\pi - \Sigma_0)(\Gamma^t)^\top. \tag{35}
$$

Thus,

$$
\begin{aligned}
\sum_{t=0}^T \operatorname{tr}((M + K^\top NK)(\Sigma_\pi - \Sigma_t)) &= \sum_{t=0}^T \operatorname{tr}\left((M + K^\top NK)\Gamma^t(\Sigma_\pi - \Sigma_0)(\Gamma^t)^\top\right) && (36)\\
&\le \operatorname{tr}(\Sigma_\pi - \Sigma_0)\operatorname{tr}\left(\sum_{t=0}^\infty (\Gamma^t)^\top(M + K^\top NK)\Gamma^t\right) && (37)\\
&= \operatorname{tr}(\Sigma_\pi - \Sigma_0)\operatorname{tr}(H_\pi)\,. && (38)
\end{aligned}
$$

Let U be the concatenation of $u_1, \ldots, u_T$, and let D be a block-diagonal matrix constructed from $D_1, \ldots, D_T$. To bound the second term, note that by the Hanson-Wright inequality

$$\begin{aligned}\mathbf{P}\left(\left|\sum_{t=1}^{T} u_t^\top D_t u_t - \operatorname{tr} D_t\right| > s\right) &= \mathbf{P}\left(\left|U^\top D U - \operatorname{tr} D\right| > s\right) \\ &\leq 2\exp\left(-c\min\left(\frac{s^2}{\|D\|_F^2}, \frac{s}{\|D\|}\right)\right). \end{aligned} \tag{39}$$

Thus with probability at least $1-\delta$ we have

$$\begin{aligned}\left|\sum_{t=1}^{T} u_t^\top D_t u_t - \operatorname{tr}(D_t)\right| &\leq \|D\|_F \sqrt{\ln(2/\delta)/c} + \|D\| \ln(2/\delta)/c \\ &\leq \sqrt{\sum_{t=1}^{T} \|D_t\|_F^2} \sqrt{\ln(2/\delta)/c} + \max_t \|D_t\| \ln(2/\delta)/c \end{aligned} \tag{40}$$

where c is a universal constant. Given that for all t,

$$\begin{aligned}\|D_t\| &\leq \operatorname{tr}(D_t) \\ &= \operatorname{tr}((M + K^\top N K)(\Sigma_\pi + \Gamma^t(\Sigma_0 - \Sigma_\pi)(\Gamma^T)^t)) \\ &\leq \lambda_\pi\,,\end{aligned}$$

with probability at least $1-\delta$,

$$\sum_{t=1}^{T} c(x_t^\pi, \pi(x_t^\pi)) - \lambda_t \leq \lambda_\pi\left(\sqrt{T\ln(2/\delta)/c} + \ln(2/\delta)/c\right)\,. \tag{41}$$

Thus, we can bound γ_T as

$$\gamma_T \leq \operatorname{tr}(H_\pi)\operatorname{tr}(\Sigma_\pi) + \lambda_\pi\left(\sqrt{T\ln(2/\delta)/c} + \ln(2/\delta)/c\right)\,. \tag{42}$$

C.3 Bounding α_T

To bound $\alpha_T = \sum_{t=1}^{T}(c(x_t, a_t) - \lambda_{\pi_t})$, in addition to bounding the cost of following a policy, we need to account for having S policy switches, as well as the cost of random actions. Let I_a be the set of time indices of all random actions $a \sim \mathcal{N}(0, \Sigma_a)$. Using the Hanson-Wright inequality, with probability at least $1-\delta$,

$$\sum_{t\in I_a} |a_t^\top N a_t - \operatorname{tr}(\Sigma_a N)| \leq \|\Sigma_a N\|_F \sqrt{|I_a|\ln(2/\delta)/c_1} + \|\Sigma_a N\| \ln(2/\delta)/c_1\,. \tag{43}$$

Let $D_{i,t} = \Sigma_t^{1/2}(M + K_i^\top N K_i)\Sigma_t^{1/2}$, and let $\lambda_{i,t} = \operatorname{tr}(D_{i,t})$. Let I_i be the set of time indices corresponding to following policy π_i in phase i. The corresponding cost can be decomposed similarly to γ_T:

$$\sum_{i=1}^{S}\sum_{t\in I_i} c(x_t, \pi_i(x_t)) - \lambda_{\pi_i} = \sum_{i=1}^{S}\sum_{t\in I_i} \operatorname{tr}((\Sigma_t - \Sigma_{\pi_i})(M + K_i^\top N K_i) + \left(u_t^\top D_{i,t} u_t - \operatorname{tr}(D_{i,t})\right)\,. \tag{44}$$

Let $D_{\max} \geq \max_{i,t}\|D_{i,t}\|$. Similarly to the previous section, with probability at least $1-\delta$ we have

$$\left|\sum_{i=1}^{S}\sum_{t\in I_i} u_t^\top D_{i,t} u_t - \operatorname{tr}(D_{i,t})\right| \leq D_{\max}\sqrt{Tn\ln(2/\delta)/c_2} + D_{\max}\ln(2/\delta)/c_2\,. \tag{45}$$

At the beginning of each phase i, the state covariance is $\Sigma_{\pi_{i-1}}$ (and we define $\Sigma_{\pi_0} = W$). After following π_i for T_v steps,

$$\sum_{i=1}^{S}\sum_{t\in I_i} \operatorname{tr}((\Sigma_t - \Sigma_{\pi_i})(M + K_i^\top N K_i)) = \sum_{i=1}^{S}\sum_{k=0}^{T_v-1} \operatorname{tr}((\Sigma_{\pi_{i-1}} - \Sigma_{\pi_i})({\Gamma_i}^k)^\top (M + K_i^\top N K_i){\Gamma_i}^k)$$

$$
\begin{aligned}
&\leq \sum_{i=1}^{S} \operatorname{tr}(H_i) \operatorname{tr}(\Sigma_{\pi_{i-1}}) \\
&\leq SnC_H \max_i \operatorname{tr}(\Sigma_{\pi_i})
\end{aligned}
$$

Following each random action, the state covariance is $\Sigma_{G,i} = A\Sigma_{\pi_i}A^\top + B\Sigma_a B^\top + W$. After taking a random action and following π_i for T_s steps, we have

$$
\sum_{k=0}^{T_s} \operatorname{tr}((\Sigma_{G,i} - \Sigma_{\pi_i})({\Gamma_i}^k)^\top (M + K_i^\top N K_i){\Gamma_i}^k) \leq \operatorname{tr}(\Sigma_{G,i}) \operatorname{tr}(H_i) \leq nC_H(\operatorname{tr}(B\Sigma_a B^\top) + \|A\|^2 \operatorname{tr}(\Sigma_{\pi_i}))\ .
$$

Putting everything together, we have

$$
\begin{aligned}
\alpha_t &\leq \|\Sigma_a N\|_F \sqrt{|I_a| \ln(2/\delta)/c_1} + \|\Sigma_a N\| \ln(2/\delta)/c_1 \\
&\quad + D_{\max}\sqrt{Tn\ln(2/\delta)/c_2} + D_{\max}\ln(2/\delta)/c_2 \\
&\quad + SnC_H \max_i \operatorname{tr}(\Sigma_{\pi_i}) \\
&\quad + |I_a| nC_H\big(\operatorname{tr}(B\Sigma_a B^\top) + \|A\|^2 \max_i \operatorname{tr}(\Sigma_{\pi_i})\big)
\end{aligned}
$$

where in v1 $S = T^{1/3-\xi}$ and $|I_a| = O(T^{2/3+\xi})$, while in v2 $S = T^{1/4}$ and $|I_a| = T^{3/4+\xi}$. We bound $\max_i \operatorname{tr}(\Sigma_{\pi_i})$ and $\|D_{i,t}\|$ in C.3.1.

C.3.1 State covariance bound

We bound $\max_i \operatorname{tr}(\Sigma_{\pi_i})$ using the following equation for the average cost of a policy:

$$
\begin{aligned}
\operatorname{tr}(\Sigma_{\pi_i}(M + K_i^\top N K_i)) &= \operatorname{tr}(H_i W) \\
\operatorname{tr}(\Sigma_{\pi_i}) &\leq \|H_i\| \operatorname{tr}(W)/\lambda_{\min}(M) \\
\max_i \operatorname{tr}(\Sigma_{\pi_i}) &\leq C_H \operatorname{tr}(W)/\lambda_{\min}(M)\ .
\end{aligned}
$$

To bound $\|D_{i,t}\|$, we note that

$$
\begin{aligned}
\|D_{i,t}\| \leq \operatorname{tr}(D_{i,t}) &= \operatorname{tr}\left(\Sigma_t(M + K_i^\top N K_i)\right) \\
&\leq \operatorname{tr}(\Sigma_t)(\|M\| + C_K^2\|N\|)\ ,
\end{aligned}
$$

and bound the state covariance $\operatorname{tr}(\Sigma_t)$. After starting at distribution $\mathcal{N}(0, \Sigma_0)$ and following a policy π_i for t steps, the state covariance is

$$
\begin{aligned}
\Sigma_t &= \Gamma_i \Sigma_{t-1} \Gamma_i^\top + W \\
&= \Gamma_i^t \Sigma_0 {\Gamma_i^t}^\top + \sum_{k=0}^{t-1} \Gamma_i^k W {\Gamma_i^k}^\top \\
&\prec \Sigma_0 + \Sigma_{\pi_i}\ .
\end{aligned}
$$

The initial covariance Σ_0 is close to $\Sigma_{\pi_{i-1}}$ after a policy switch, and close to $A\Sigma_{\pi_i}A^\top + B\Sigma_a B^\top + W$ after a random action. Therefore we can bound the state covariance in each phase as

$$
\begin{aligned}
\Sigma_t &\preceq \Sigma_{\pi_i} + \Sigma_{\pi_{i-1}} + A\Sigma_{\pi_i}A^\top + B\Sigma_a B^\top \\
\operatorname{tr}(\Sigma_t) &\leq (2 + \|A\|^2) \max_i \operatorname{tr}(\Sigma_{\pi_i}) + \operatorname{tr}(B\Sigma_a B^\top) \\
&\leq (2 + \|A\|^2) C_H \operatorname{tr}(W)/\lambda_{\min}(M) + \operatorname{tr}(B\Sigma_a B^\top)\ .
\end{aligned}
$$