Appendix

A. Preliminaries

We state some useful definitions and lemmas in this section.

Lemma A.1. Let X_1 and X_2 be a pair of distribution vectors. Let H be the transition matrix of an ergodic Markov chain with a stationary distribution ν, and ergodicity coefficient (defined in Assumption 2.1) upper-bounded by $\gamma < 1$. Then

$$\|(H^m)^\top (X_1 - X_2)\|_1 \leq \gamma^m \|X_1 - X_2\|_1 \, .$$

Proof. Let $\{v_1, ..., v_n\}$ be the normalized left eigenvectors of H corresponding to ordered eigenvalues $\{\lambda_1, ..., \lambda_n\}$. Then $v_1 = \nu$, $\lambda_1 = 1$, and for all $i \geq 2$, we have that $\lambda_i < 1$ (since the chain is ergodic) and $v_i^\top \mathbf{1} = 0$. Write X_1 in terms of the eigenvector basis as:

$$X_1 = \alpha_1 \nu + \sum_{i=2}^{n} \alpha_i v_i \quad \text{and} \quad X_2 = \beta_1 \nu + \sum_{i=2}^{n} \beta_i v_i \, .$$

Since $X_1^\top \mathbf{1} = 1$ and $X_2^\top \mathbf{1} = 1$, it is easy to see that $\alpha_1 = \beta_1 = 1$. Thus we have

$$\|H^\top (X_1 - X_2)\|_1 = \|H^\top \sum_{i=2}^{n} (\alpha_i - \beta_i) v_i\|_1 \leq \gamma \| \sum_{i=2}^{n} (\alpha_i - \beta_i) v_i\|_1 = \gamma \|X_1 - X_2\|_1$$

where the inequality follows from the definition of the ergodicity coefficient and the fact that $\mathbf{1}^\top v_i = 0$ for all $i \geq 2$. Since

$$\mathbf{1}^\top H^\top \sum_{i=2}^{n} (\alpha_i - \beta_i) v_i = \mathbf{1}^\top \sum_{i=2}^{n} \lambda_i (\alpha_i - \beta_i) v_i = 0,$$

the inequality also holds for powers of H. □

Lemma A.2 (Doob martingale)**.** Let Assumption 2.1 hold, and let $\{(x_t, a_t)\}_{t=1}^T$ be the state-action sequence obtained when following policies $\pi_1, ..., \pi_k$ for τ steps each from an initial distribution ν_0. For $t \in [T]$, let X_t be a binary indicator vector with a non-zero element at the linear index of the state-action pair (x_t, a_t). Define for $i \in [T]$,

$$B_i = \mathbb{E}\left[\sum_{t=1}^{T} X_t | X_1, ..., X_i\right], \quad \text{and} \quad B_0 = \mathbb{E}\left[\sum_{t=1}^{T} X_t\right].$$

Then, $\{B_i\}_{i=0}^T$ is a vector-valued martingale: $\mathbb{E}[B_i - B_{i-1} | B_0, \ldots, B_{i-1}] = 0$ for $i = 1, \ldots, T$, and $\|B_i - B_{i-1}\|_1 \leq 2(1-\gamma)^{-1}$ holds for $i \in [T]$.

The constructed martingale is known as the Doob martingale underlying the sum $\sum_{t=1}^T X_t$.

Proof. That $\{B_i\}_{i=0}^T$ is a martingale follows from the definition. We now bound its difference sequence. Let H_t be the state-action transition matrix at time t, and let $H_{i:t} = \prod_{j=i}^{t-1} H_j$, and define $H_{i:i} = I$. Then, for $t = 0, \ldots, T-1$, $\mathbb{E}[X_{t+1} | X_t] = H_t^\top X_t$ and by the Markov property, for any $i \in [T]$,

$$B_i = \sum_{t=1}^{i} X_t + \sum_{t=i+1}^{T} \mathbb{E}[X_t | X_i] = \sum_{t=1}^{i} X_t + \sum_{t=i+1}^{T} H_{i:t}^\top X_i, \quad \text{and} \quad B_0 = \sum_{t=1}^{T} H_{0:t}^\top X_0.$$

For any $i \in [T]$,

$$\begin{aligned} B_i - B_{i-1} &= \sum_{t=1}^{i} X_t - \sum_{t=1}^{i-1} X_t + \sum_{t=i+1}^{T} H_{i:t}^\top X_i - \sum_{t=i}^{T} H_{i-1:t}^\top X_{i-1} \\ &= \sum_{t=i}^{T} H_{i:t}^\top (X_i - H_{i-1}^\top X_{i-1}). \qquad \text{(A.1)} \end{aligned}$$

Since X_i and $H_{i-1}^\top X_{i-1}$ are distribution vectors, under Assumption 2.1 and using Lemma A.1,

$$\|B_i - B_{i-1}\|_1 \leq \sum_{t=i}^{T} \|H_{i:t}^\top (X_i - H_{i-1}^\top X_{i-1})\|_1 \leq 2 \sum_{j=0}^{T-i} \gamma^j \leq 2(1-\gamma)^{-1} .$$

□

Let $(\mathcal{F}_k)_k$ be a filtration and define $\mathbb{E}_k[\cdot] := \mathbb{E}[\cdot|\mathcal{F}_k]$. We will make use of the following concentration results for the sum of random matrices and vectors.

Theorem A.3 (Matrix Azuma, Tropp (2012) Thm 7.1)**.** Consider a finite $(\mathcal{F})_k$-adapted sequence $\{X_k\}$ of Hermitian matrices of dimension m, and a fixed sequence $\{A_k\}$ of Hermitian matrices that satisfy $\mathbb{E}_{k-1} X_k = 0$ and $X_k^2 \preceq A_k^2$ almost surely. Let $v = \|\sum_k A_k^2\|$. Then with probability at least $1-\delta$, $\|\sum_k X_k\|_2 \leq 2\sqrt{2v\ln(m/\delta)}$.

A version of Theorem A.3 for non-Hermitian matrices of dimension $m_1 \times m_2$ can be obtained by applying the theorem to a Hermitian dilation of X, $\mathcal{D}(X) = \left[\begin{smallmatrix} 0 & X \\ X^* & 0 \end{smallmatrix}\right]$, which satisfies $\lambda_{\max}(\mathcal{D}(X)) = \|X\|$ and $\mathcal{D}(X)^2 = \left[\begin{smallmatrix} XX^* & 0 \\ 0 & X^*X \end{smallmatrix}\right]$. In this case, we have that $v = \max\left(\|\sum_k X_k X_k^*\|, \|\sum_k X_k^* X_k\|\right)$.

Lemma A.4 (Hoeffding-type inequality for norm-subGaussian random vectors, Jin et al. (2019))**.** Consider random vectors $X_1, \ldots, X_n \in \mathbb{R}^d$ and corresponding filtrations $\mathcal{F}_i = \sigma(X_1, \ldots, X_i)$ $i \in [n]$, such that $X_i|\mathcal{F}_{i-1}$ is zero-mean norm-subGaussian with $\sigma_i \in \mathcal{F}_{i-1}$. That is:

$$\mathbb{E}[X_i|\mathcal{F}_i] = 0, \quad P(\|X_i\| \geq t|\mathcal{F}_{i-1}) \leq 2\exp(-t^2/2\sigma_i^2) \quad \forall t \in \mathbb{R}, \forall i \in [n].$$

If the condition is satisfied for fixed $\{\sigma_i\}$, there exists a constant c such that for any $\delta > 0$, with probability at least $1-\delta$,

$$\|\sum_{i=1}^{n} X_i\| \leq c\sqrt{\sum_{i=1}^{n} \sigma_i^2 \log(2d/\delta)} \, .$$

B. Bounding the Difference Between Empirical and Average Rewards

In this section, we bound the second term in Equation 5.1, corresponding to the difference between empirical and average rewards.

Lemma B.1. Let Assumption 2.1 hold, and assume that $\tau \geq \frac{\log T}{2\log(1/\gamma)}$ and that $r(x,a) \in [0,1]$ for all x, a. Then, by choosing $\eta = \frac{\sqrt{8\log|\mathcal{A}|}}{Q_{\max}\sqrt{K}}$, we have with probability at least $1-\delta$,

$$\sum_{k=1}^{K} \sum_{t=(k-1)\tau+1}^{k\tau} (r_t - J_{\pi_k}) \leq 2(1-\gamma)^{-1}\sqrt{2T\log(2/\delta)} + 2\sqrt{T} + (1-\gamma)^{-2}\sqrt{8K\log|\mathcal{A}|} \, .$$

Proof. Let r denote the vector of rewards, and recall that $J_\pi = \nu_\pi^\top r$. Let X_t be the indicator vector for the state-action pair at time t, as in Lemma A.2, and let $\nu_t = \mathbb{E}[X_t]$. We have the following:

$$V_T := \sum_{k=1}^{K} \sum_{t=(k-1)\tau+1}^{k\tau} (r_t - J_{\pi_k}) = \sum_{k=1}^{K} \sum_{t=(k-1)\tau+1}^{k\tau} r^\top (X_t - \nu_t + \nu_t - \nu_{\pi_k})$$

We slightly abuse the notation above by letting ν_t denote the state-action distribution at time t, and ν_π the stationary distribution of policy π. Let $\{B_i\}_{i=0}^T$ be the Doob martingale in Lemma A.2. Then $B_0 = \sum_{t=1}^T \nu_t$ and $B_T = \sum_{t=1}^T X_t$, and the first term can be expressed as

$$V_{T1} := \sum_{t=1}^{T} r^\top (X_t - \nu_t) = r^\top (B_T - B_0).$$

By Lemma A.2, $|\langle B_i - B_{i-1}, r\rangle| \leq \|B_i - B_{i-1}\|_1 \|r\|_\infty \leq 2(1-\gamma)^{-1}$. Hence by Azuma's inequality, with probability at least $1-\delta$,

$$V_{T1} \leq 2(1-\gamma)^{-1}\sqrt{2T\log(2/\delta)}. \tag{B.1}$$

For the second term we have

$$\begin{aligned}
V_{T2} &:= \sum_{k=1}^{K} \sum_{t=(k-1)\tau+1}^{k\tau} r^\top(\nu_t - \nu_{\pi_k}) \\
&= \sum_{k=1}^{K} r^\top \left(\sum_{i=1}^{\tau} (H^i_{\pi_k})^\top \nu_{(k-1)\tau} - \nu_{\pi_k} \right) \\
&\leq \sum_{k=1}^{K} \|r\|_\infty \sum_{i=1}^{\tau} \left\| (H^i_{\pi_k})^\top (\nu_{(k-1)\tau} - \nu_{\pi_{k-1}} + \nu_{\pi_{k-1}}) - \nu_{\pi_k} \right\|_1 \\
&\leq \sum_{k=1}^{K} \sum_{i=1}^{\tau} \|\nu_{(k-1)\tau} - \nu_{\pi_{k-1}}\|_1 + \|(H^i_{\pi_k})^\top \nu_{\pi_{k-1}} - \nu_{\pi_k}\|_1 \\
&\leq \sum_{k=1}^{K} \sum_{i=1}^{\tau} \|(H^\tau_{\pi_{(k-1)}})^\top \nu_{(k-2)\tau} - \nu_{\pi_{k-1}}\|_1 + \gamma^i \|\nu_{\pi_{k-1}} - \nu_{\pi_k}\|_1 \\
&\leq 2T\gamma^\tau + \frac{1}{1-\gamma} \sum_{k=1}^{K} \|\nu_{\pi_k} - \nu_{\pi_{k-1}}\|_1 .
\end{aligned}$$

For $\tau \geq \frac{\log T}{2\log(1/\gamma)}$, the first term is upper-bounded by $2\sqrt{T}$.

Using results on perturbations of Markov chains (Seneta, 1988; Cho & Meyer, 2001), we have that

$$\|\nu_{\pi_k} - \nu_{\pi_{k-1}}\|_1 \leq \frac{1}{1-\gamma}\|H_{\pi_k} - H_{\pi_{k-1}}\|_\infty \leq \frac{1}{1-\gamma} \max_x \|\pi_k(\cdot|x) - \pi_{k-1}(\cdot|x)\|_1$$

Note that the policies $\pi_k(\cdot|x)$ are generated by running mirror descent on reward functions $\widehat{Q}_{\pi_k}(x,\cdot)$. A well-known property of mirror descent updates with entropy regularization (or equivalently, the exponentially-weighted-average algorithm) is that the difference between consecutive policies is bounded as

$$\|\pi_{k+1}(\cdot|x) - \pi_k(\cdot|x)\|_1 \leq \eta \|\widehat{Q}_{\pi_k}(x,\cdot)\|_\infty .$$

See e.g. Neu et al. (2014) Section V.A for a proof, which involves applying Pinsker's inequality and Hoeffding's lemma (Cesa-Bianchi & Lugosi (2006) Section A.2 and Lemma A.6). Since we assume that $\|\widehat{Q}_{\pi_k}\|_\infty \leq Q_{\max}$, we can obtain

$$V_{T2} \leq 2\sqrt{T} + (1-\gamma)^{-2} K \eta Q_{\max}.$$

By choosing $\eta = \frac{\sqrt{8\log|\mathcal{A}|}}{Q_{\max}\sqrt{K}}$, we can bound the second term as

$$V_{T2} \leq 2\sqrt{T} + (1-\gamma)^{-2}\sqrt{8K\log|\mathcal{A}|}. \tag{B.2}$$

Putting Eq. (B.1) and (B.2) together, we obtain that with probability at least $1-\delta$,

$$V_T \leq 2(1-\gamma)^{-1}\sqrt{2T\log(2/\delta)} + 2\sqrt{T} + (1-\gamma)^{-2}\sqrt{8K\log|\mathcal{A}|} .$$

□

C. Proof of Lemma 6.3

Proof. Recall that we split each phase into $2m$ blocks of size b and let $\mathcal{H}_i$ and $\mathcal{T}_i$ denote the starting indices of odd and even blocks, respectively. We let R_t denote the empirical b-step returns from the state action pair (x_t, a_t) in phase i:

$$R_t = \sum_{i=t}^{t+b} (r_i - \widehat{J}_{\pi_i}), \quad \widehat{J}_{\pi_i} = \frac{1}{|\mathcal{T}_i|} \sum_{t \in \mathcal{T}_i} r_t.$$

We start by bounding the error in R_t. Let X be a binary indicator vector for a state-action pair (x, a). Let H_π be the state-action transition kernel for policy π, and let ν_π be the corresponding stationary state-action distribution. We can write the action-value function at (x, a) as

$$\begin{aligned} Q_\pi(x, a) &= r(x, a) - J_\pi + X^\top H_\pi Q_\pi \\ &= (X - \nu_\pi)^\top r + X^\top H_\pi (r - J_\pi \mathbf{1} + H_\pi Q_\pi) \\ &= \sum_{i=0}^{\infty} (X - \nu_\pi)^\top H_\pi^i r \,. \end{aligned}$$

Let $Q_\pi^b(x, a) = \sum_{i=0}^{b} (X - \nu_\pi)^\top H_\pi^i r$ be a version of Q_π truncated to b steps. Under uniform mixing, the difference to the true Q_π is bounded as

$$|Q_\pi(x, a) - Q_\pi^b(x, a)| \leq \sum_{i=1}^{\infty} \left|(X - \nu_\pi)^\top H_\pi^{i+b} r\right| \leq \frac{2\gamma^{b+1}}{1 - \gamma} \,. \tag{C.1}$$

Let $b_t = Q_{\pi_i}^b(x_t, a_t) - Q_{\pi_i}(x_t, a_t)$ denote the truncation bias at time t, and let $z_t = \sum_{i=t}^{t+b} r_i - X_t^\top H_{\pi_i}^{(i-t)} r$ denote the reward noise. We will write

$$R_t = Q_{\pi_i}(x_t, a_t) + b(J_{\pi_i} - \widehat{J}_{\pi_i}) + z_t + b_t.$$

Note that $m = |\mathcal{H}_i|$ and let

$$\widehat{M}_i = \frac{1}{m} \sum_{t \in \mathcal{H}_i} \phi_t \phi_t^\top + \frac{\alpha}{m} I \,.$$

We estimate the value function of each policy π_i using data from phase i as

$$\begin{aligned} \widehat{w}_{\pi_i} &= \widehat{M}_i^{-1} m^{-1} \sum_{t \in \mathcal{H}_i} \phi_t R_t \\ &= \widehat{M}_i^{-1} m^{-1} \sum_{t \in \mathcal{H}_i} \phi_t (\phi_t^\top w_{\pi_i} + b_t + z_t + b(J_{\pi_i} - \widehat{J}_{\pi_i})) + \widehat{M}_i^{-1} \frac{\alpha}{m} (w_{\pi_i} - w_{\pi_i}) \\ &= w_{\pi_i} + \widehat{M}_i^{-1} m^{-1} \sum_{t \in \mathcal{H}_i} \phi_t (z_t + b_t + b(J_{\pi_i} - \widehat{J}_{\pi_i})) - \widehat{M}_i^{-1} m^{-1} \alpha w_{\pi_i} \end{aligned}$$

Our estimate $\widehat{w}_k$ of $w_k = \frac{1}{k} \sum_{i=1}^{k} w_{\pi_i}$ can thus be written as follows:

$$\widehat{w}_k - w_k = \frac{1}{km} \sum_{i=1}^{k} \sum_{t \in \mathcal{H}_i} \widehat{M}_i^{-1} \phi_t (z_t + b_t + b(J_{\pi_i} - \widehat{J}_{\pi_i})) - \frac{\alpha}{km} \sum_{i=1}^{k} \widehat{M}_i^{-1} w_{\pi_i}.$$

We proceed to upper-bound the norm of the RHS.

Set $\alpha = \sqrt{m/k}$. Let C_w be an upper-bound on the norm of the true value-function weights $\|w_{\pi_i}\|_2$ for $i = 1, ..., K$. In Appendix C.3, we show that with probability at least $1 - \delta$, for $m \geq 72 C_\Phi^4 \sigma^{-2} (1 - \gamma)^{-2} \log(d/\delta)$, $\|\widehat{M}_i^{-1}\|_2 \leq 2\sigma^{-2}$. Thus with probability at least $1 - \delta$, the last error term is upper-bounded as

$$\frac{\alpha}{km} \left\| \sum_{k=1}^{k} \widehat{M}_i^{-1} w_{\pi_i} \right\|_2 \leq 2\sigma^{-2} C_w (km)^{-1/2}. \tag{C.2}$$

Similarly, for

$$b \geq \frac{\log((1-\gamma)^{-1}\sqrt{km})}{\log(1/\gamma)}, \tag{C.3}$$

the norm of the truncation bias term is upper-bounded as

$$\frac{1}{km}\sum_{i=1}^{k}\sum_{t\in\mathcal{H}_i}\|\widehat{M}_i^{-1}\phi_t b_t\|_2 \leq \frac{2\gamma^b}{km(1-\gamma)}\sum_{i=1}^{k}\sum_{t\in\mathcal{H}_i}\|\widehat{M}_i^{-1}\phi_t\|_2 \leq 2\sigma^{-2}C_\Phi(km)^{-1/2}. \tag{C.4}$$

To bound the error terms corresponding to reward noise z_t and average-error noise $J_{\pi_i} - \widehat{J}_{\pi_i}$, we rely on the independent blocks techniques of Yu (1994). We show in Sections C.1 and C.2 that with probability $1-2\delta$, for constants c_1 and c_2, each of these terms can be bounded as:

$$\frac{1}{km}\left\|\sum_{i=1}^{k}\sum_{t\in\mathcal{H}_i}\widehat{M}_i^{-1}\phi_t z_t\right\|_2 \leq 2c_1 C_\Phi\sigma^{-2}\sqrt{\frac{b\log(2d/\delta)}{km}}$$
$$\frac{b}{km}\left\|\sum_{i=1}^{k}(J_{\pi_i}-\widehat{J}_{\pi_i})\sum_{t\in\mathcal{H}_i}\widehat{M}_i^{-1}\phi_t\right\|_2 \leq 2c_2 C_\Phi\sigma^{-2}b\sqrt{\frac{\log(2d/\delta)}{km}}.$$

Thus, putting terms together, we have for an absolute constant c, with probability at least $1-\delta$,

$$\|\widehat{w}_k - w_k\|_2 \leq c\sigma^{-2}(C_w + C_\Phi)b\sqrt{\frac{\log(2d/\delta)}{km}}.$$

Note that this result holds for every $k \in [K]$ and thus also holds for $k = K$. □

C.1. Bounding $\sum_{i=1}^{k}\widehat{M}_i^{-1}\sum_{t\in\mathcal{H}_i}\phi_t z_t$

Let $\|\cdot\|_{\mathrm{tv}}$ denote the total variation norm.

Definition C.1 (β-mixing)**.** Let $\{Z_t\}_{t=1,2,\dots}$ be a stochastic process. Denote by $Z_{1:t}$ the collection $(Z_1,\dots,Z_t)$, where we allow $t=\infty$. Let $\sigma(Z_{i:j})$ denote the sigma-algebra generated by $Z_{i:j}$ ($i \leq j$). The k^{th} β-mixing coefficient of $\{Z_t\}$, β_k, is defined by

$$\begin{aligned}\beta_k &= \sup_{t\geq 1}\mathbb{E}\sup_{B\in\sigma(Z_{t+k:\infty})}|P(B|Z_{1:t}) - P(B)| \\ &= \sup_{t\geq 1}\mathbb{E}\|P_{Z_{t+k:\infty}|Z_{1:t}}(\cdot|Z_{1:t}) - P_{Z_{t+k:\infty}}(\cdot)\|_{\mathrm{tv}}\,.\end{aligned}$$

$\{Z_t\}$ is said to be β-mixing if $\beta_k \to 0$ as $k\to\infty$. In particular, we say that a β-mixing process mixes at an *exponential* rate with parameters $\overline{\beta},\alpha,\gamma > 0$ if $\beta_k \leq \overline{\beta}\exp(-\alpha k^\gamma)$ holds for all $k \geq 0$.

Let X_t be the indicator vector for the state-action pair (x_t,a_t) as in Lemma A.2. Note that the distribution of (x_{t+1},a_{t+1}) given (x_t,a_t) can be written as $\mathbb{E}[X_{t+1}|X_t]$. Let H_t be the state-action transition matrix at time t, let $H_{i:t} = \prod_{j=i}^{t-1}H_j$, and define $H_{i:i} = I$. Then we have that $\mathbb{E}[X_{t+k}|X_{1:t}] = H_{t:t+k}^\top X_t$ and $\mathbb{E}[X_{t+k}] = H_{1:t+k}^\top\nu_0$, where ν_0 is the initial state distribution. Thus, under the uniform mixing Assumption 2.1, the k^{th} β-mixing coefficient is bounded as:

$$\beta_k \leq \sup_{t\geq 1}\mathbb{E}\sum_{j=k}^{\infty}\|H_{t:t+j}^\top X_t - H_{1:t+j}^\top\nu_0\|_1 \leq \sup_{t\geq 1}\mathbb{E}\sum_{j=k}^{\infty}\gamma^j\|X_t - H_{1:t}^\top\nu_0\|_1 \leq \frac{2\gamma^k}{1-\gamma}\,.$$

We bound the noise terms using the independent blocks technique of Yu (1994). Recall that we partition each phase into $2m$ blocks of size b. Thus, after k phases we have a total of $2km$ blocks. Let $\mathbb{P}$ denote the joint distribution of state-action pairs in *odd* blocks. Let $\mathcal{I}_i$ denote the set of indices in the i^{th} block, and let $x_{\mathcal{I}_i}, a_{\mathcal{I}_i}$ denote the corresponding states and actions. We factorize the joint distribution according to blocks:

$$\begin{aligned}\mathbb{P}(x_{\mathcal{I}_1},a_{\mathcal{I}_1},x_{\mathcal{I}_3},a_{\mathcal{I}_3},\dots,x_{\mathcal{I}_{2km-1}},a_{\mathcal{I}_{2km-1}}) = \; &\mathbb{P}_1(x_{\mathcal{I}_1},a_{\mathcal{I}_1})\times\mathbb{P}_3(x_{\mathcal{I}_3},a_{\mathcal{I}_3}|x_{\mathcal{I}_1},a_{\mathcal{I}_1})\times\cdots \\ &\times\mathbb{P}_{2km-1}(x_{\mathcal{I}_{2km-1}},a_{\mathcal{I}_{2km-1}}|x_{\mathcal{I}_{2km-3}},a_{\mathcal{I}_{2km-3}}).\end{aligned}$$

Let $\widetilde{\mathbb{P}}_i$ be the marginal distribution over the variables in block i, and let $\widetilde{\mathbb{P}}$ be the product of marginals of odd blocks.

Corollary 2.7 of Yu (1994) implies that for any Borel-measurable set E,

$$|\mathbb{P}(E) - \widetilde{\mathbb{P}}(E)| \leq (km-1)\beta_b \tag{C.5}$$

where β_b is the b^{th} β-mixing coefficient of the process. The result follows since the size of the "gap" between successive blocks is b; see Appendix E for more details.

Recall that our estimates $\widehat{w}_{\pi_i}$ are based only on data in odd blocks in each phase. Let $\widetilde{\mathbb{E}}$ denote the expectation w.r.t. the product-of-marginals distribution $\widetilde{\mathbb{P}}$. Then $\widetilde{\mathbb{E}}[\widehat{M}_i^{-1}\sum_{t\in\mathcal{H}_i}\phi_t z_t] = 0$ because for $t \in \mathcal{H}_i$ and under $\widetilde{\mathbb{P}}$, z_t is zero-mean given ϕ_t and is independent of other feature vectors outside of the block. Furthermore, by Hoeffding's inequality $\widetilde{\mathbb{P}}(|z_t|/b \geq a) \leq 2\exp(-2ba^2)$. Since $\|\phi_t\|_2 \leq C_\Phi$ and $\|\widehat{M}_i^{-1}\|_2 \leq 2\sigma^{-2}$ for large enough m, we have that

$$\widetilde{\mathbb{P}}(\|\widehat{M}_i^{-1}\phi_t z_t\|_2 \geq 2b\sigma^{-2}C_\Phi a) \leq 2\exp(-2ba^2).$$

Since $\widehat{M}_i^{-1}\phi_t z_t$ are norm-subGaussian vectors, using Lemma A.4, there exists a constant c_1 such that for any $\delta \geq 0$

$$\widetilde{\mathbb{P}}\left(\left\|\sum_{i=1}^k \widehat{M}_i^{-1}\sum_{t\in\mathcal{H}_i}\phi_t z_t\right\|_2 \geq 2c_1 C_\Phi\sigma^{-2}\sqrt{bkm\log(2d/\delta)}\right) \leq \delta\,.$$

Thus, using (C.5),

$$\mathbb{P}\left(\left\|\sum_{i=1}^k \widehat{M}_i^{-1}\sum_{t\in\mathcal{H}_i}\phi_t z_t\right\|_2 \geq 2c_1 C_\Phi\sigma^{-2}\sqrt{bkm\log(2d/\delta)}\right) \leq \delta + (km-1)\beta_b\,.$$

Under Assumption 2.1, we have that $\beta_b \leq 2\gamma^b(1-\gamma)^{-1}$. Setting $\delta = 2km\gamma^b(1-\gamma)^{-1}$ and solving for b we get

$$b = \frac{\log(2km\delta^{-1}(1-\gamma)^{-1})}{\log(1/\gamma)}. \tag{C.6}$$

Notice that when b is chosen as in Eq. (C.6), the condition (C.3) is also satisfied. Plugging this into the previous display gives that with probability at least $1-2\delta$,

$$\left\|\sum_{i=1}^k \widehat{M}_i^{-1}\sum_{t\in\mathcal{H}_i}\phi_t z_t\right\|_2 \leq 2c_1 C_\Phi\sigma^{-2}\sqrt{bkm\log(2d/\delta)}.$$

C.2. Bounding $\left\|\sum_{i=1}^k \widehat{M}_i^{-1}\sum_{t\in\mathcal{H}_i}\phi_t(J_{\pi_i} - \widehat{J}_{\pi_i})\right\|_2$

Recall that the average-reward estimates $\widehat{J}_{\pi_i}$ are computed using time indices corresponding to the starts of even blocks, $\mathcal{T}_i$. Thus this error term is only a function of the indices corresponding to block starts. Now let $\mathbb{P}$ denote the distribution over state-action pairs (x_t, a_t) for indices t corresponding to block starts, i.e. $t \in \{1, b+1, 2b+1, ..., (2km-1)b+1\}$. We again factorize the distribution over blocks as $\mathbb{P} = \mathbb{P}_1 \otimes \mathbb{P}_2 \otimes \cdots \otimes \mathbb{P}_{2km}$. Let $\widetilde{\mathbb{P}} = \widetilde{\mathbb{P}}_1 \otimes \widetilde{\mathbb{P}}_2 \otimes \cdots \otimes \widetilde{\mathbb{P}}_{2km}$ be a product-of-marginals distribution defined as follows. For odd j, let $\widetilde{\mathbb{P}}_j$ be the marginal of $\mathbb{P}$ over (x_{jb+1}, a_{jb+1}). For even j in phase i, let $\widetilde{\mathbb{P}}_j = \nu_{\pi_i}$ correspond to the stationary distribution of the corresponding policy π_i. Using arguments similar to independent blocks, we show in Appendix E that

$$\|\mathbb{P} - \widetilde{\mathbb{P}}\|_1 \leq 2(2km-1)\gamma^{b-1}.$$

Let $\widetilde{\mathbb{E}}$ denote expectation w.r.t. the product-of-marginals distribution $\widetilde{\mathbb{P}}$. Then $\widetilde{\mathbb{E}}[\widehat{M}_i^{-1}\sum_{t\in\mathcal{H}^i}\phi_t(J_{\pi_i} - \widehat{J}_{\pi_i})] = 0$, since under $\widetilde{\mathbb{P}}$, $\widehat{J}_{\pi_i}$ is the sum of rewards for state-action pairs distributed according to ν_{π_i}, and these state-action pairs are independent of other data. Using a similar argument as in the previous section, for $b = 1 + \frac{\log(4km/\delta)}{\log(1/\gamma)}$, there exists a constant c_2 such that with probability at least $1-2\delta$,

$$\left\|\sum_{i=1}^k \widehat{M}_i^{-1}\sum_{t\in\mathcal{H}_i}\phi_t(J_{\pi_i} - \widehat{J}_{\pi_i})\right\|_2 \leq 2c_2 C_\Phi\sigma^{-2}\sqrt{km\log(2d/\delta)}\,.$$

C.3. Bounding $\|\widehat{M}_i^{-1}\|_2$

In this subsection, we show that with probability at least $1-\delta$, for $m \geq 72C_\Phi^4\sigma^{-2}(1-\gamma)^{-2}\log(d/\delta))$, $\|M_i^{-1}\|_2 \leq 2\sigma^{-2}$.

Let Φ be a $|\mathcal{X}||\mathcal{A}| \times d$ matrix of all features. Let $D_i = \mathrm{diag}(\nu_{\pi_i})$, and let $\widehat{D}_i = \mathrm{diag}(\sum_{t\in\mathcal{H}_i} X_t)$, where X_t is a state-action indicator as in Lemma A.2. Let $M_i = \Phi^\top D_i \Phi + \alpha m^{-1} I$. We can write $\widehat{M}_i^{-1}$ as

$$\begin{aligned}\widehat{M}_i^{-1} &= (\Phi^\top \widehat{D}_i \Phi + \alpha\tau^{-1} I + \Phi^\top (D_i - D_i)\Phi)^{-1} \\ &= (M_i + \Phi^\top (D_i - D_i)\Phi)^{-1} \\ &= (I + M_i^{-1}\Phi^\top (D_i - D_i)\Phi)^{-1} M_i^{-1}\end{aligned}$$

By Assumption 6.2 and 6.1, $\|M_i^{-1}\|_2 \leq \sigma^{-2}$. In Appendix C.4, we show that w.p. at least $1-\delta$,

$$\|\Phi^\top(\widehat{D}_i - D_i)\Phi\|_2 \leq 6m^{-1/2}C_\Phi^2(1-\gamma)^{-1}\sqrt{2\log(d/\delta)}$$

Thus

$$\|\widehat{M}_i^{-1}\|_2 \leq \sigma^{-2}(1 - \sigma^{-2}6m^{-1/2}C_\Phi^2(1-\gamma)^{-1}\sqrt{2\log(d/\delta)})^{-1}$$

For $m \geq 72C_\Phi^4\sigma^{-2}(1-\gamma)^{-2}\log(d/\delta))$, the above norm is upper-bounded by $\|\widehat{M}_i^{-1}\|_2 \leq 2\sigma^{-2}$.

C.4. Bounding $\|\Phi^\top(\widehat{D}_i - D_i)\Phi^\top\|_2$

For any matrix A,

$$\|\Phi^\top A \Phi\|_2 = \left\|\sum_{ij} A_{ij}\phi_i\phi_j^\top\right\|_2 \leq \sum_{i,j}|A_{ij}|\|\phi_i\phi_j^\top\|_2 \leq C_\Phi^2 \sum_{i,j}|A_{ij}| = C_\Phi^2\|A\|_{1,1}\,. \tag{C.7}$$

where $\|A\|_{1,1}$ denotes the sum of absolute entries of A. Using the same notation for X_t as in Lemma A.2,

$$\begin{aligned}\|\Phi^\top(\widehat{D}_i - D_i)\Phi\|_2 &= \frac{1}{m}\sum_{t\in\mathcal{H}_i}\Phi^\top \mathrm{diag}(X_t - \nu_t + \nu_t - \nu_{\pi_i})\Phi \\ &\leq \frac{1}{m}\left\|\sum_{t\in\mathcal{H}_i}\Phi^\top\mathrm{diag}(X_t - \nu_t)\Phi\right\|_2 + \frac{C_\Psi^2}{m}\sum_{t\in\mathcal{H}_i}\|\nu_t - \nu_{\pi_i}\|_1\,.\end{aligned}$$

Under the fast-mixing assumption 2.1, the second term is bounded by $2C_\Psi^2 m^{-1}(1-\gamma)^{-1}$.

For the first term, we can define a martingale $(B_i)_{i=0}^m$ similar to the Doob martingale in Lemma A.2, but defined only on the m indices $\mathcal{H}_i$. Note that $\sum_{t\in\mathcal{H}_i}\Phi^\top\mathrm{diag}(X_t - \nu_t)\Phi = \Phi^\top\mathrm{diag}(B_m - B_0)\Phi$. Thus we can use matrix-Azuma to bound the difference sequence. Given that

$$\|(\Phi^\top(B_i - B_{i-1})\Phi)^2\|_2 \leq 4C_\Phi^4(1-\gamma)^{-2},$$

combining the two terms, we have that with probability at least $1-\delta$,

$$\begin{aligned}\|\Phi^\top(\widehat{D}_i - D_i)\Phi\|_2 &\leq 4m^{-1/2}C_\Phi^2(1-\gamma)^{-1}\sqrt{2\log(d/\delta)} + 2m^{-1}C_\Phi^2(1-\gamma)^{-1} \\ &\leq 6m^{-1/2}C_\Phi^2(1-\gamma)^{-1}\sqrt{2\log(d/\delta)}\,.\end{aligned}$$

D. Bounding $\|V_K - \widehat{V}_K\|_{\mu_*}$

We write the value function error as follows:

$$\begin{aligned}\mathbb{E}_{x\sim\mu*}[\widehat{V}_K(x) - V_K(x)] &= \sum_x \mu_*(x)\sum_a \phi(x,a)^\top \frac{1}{K}\sum_{i=1}^K \pi_i(a|x)(\widehat{w}_{\pi_i} - w_{\pi_i}) \\ &\leq \frac{1}{K}\sum_x \mu_*(x)\sum_a \|\phi(x,a)\|_2 \left\|\sum_{i=1}^K \pi_i(a|x)(\widehat{w}_{\pi_i} - w_{\pi_i})\right\|_2\end{aligned}$$

Note that for any set of scalars $\{p_i\}_{i=1}^K$ with $p_i \in [0,1]$, the term $\left\|\sum_{i=1}^K p_i(\widehat{w}_{\pi_i} - w_{\pi_i})\right\|_2$ has the same upper bound as $\|\sum_{i=1}^K(\widehat{w}_{\pi_i} - w_{\pi_i})\|_2$. The reason is as follows. One part of the error includes bias terms (C.2) and (C.4), whose upper bounds are only smaller when reweighted by scalars in $[0,1]$. Thus we can simply upper-bound the bias by setting all $\{p_i\}_{i=1}^K$ to 1. Another part of the error, analyzed in Appendices C.1 and C.2 involves sums of norm-subGaussian vectors. In this case, applying the weights only results in these vectors potentially having smaller norm bounds. We keep the same bounds for simplicity, again corresponding to all $\{p_i\}_{i=1}^K$ equal to 1. Thus, reusing the results of the previous section, we have

$$\mathbb{E}_{x\sim\mu*}[\widehat{V}_K(x) - V_K(x)] \le C_\Phi|\mathcal{A}|c\sigma^{-2}(C_w + C_\Phi)b\sqrt{\frac{\log(2d/\delta)}{Km}}.$$

E. Independent Blocks

Blocks. Recall that we partition each phase into $2m$ blocks of size b. Thus, after k phases we have a total of $2km$ blocks. Let $\mathbb{P}$ denote the joint distribution of state-action pairs in odd blocks. Let $\mathcal{I}_i$ denote the set of indices in the i^{th} block, and let $x_{\mathcal{I}_i}, a_{\mathcal{I}_i}$ denote the corresponding states and actions. We factorize the joint distribution according to blocks:

$$\begin{aligned}\mathbb{P}(x_{\mathcal{I}_1}, a_{\mathcal{I}_1}, x_{\mathcal{I}_3}, a_{\mathcal{I}_3}, \ldots, x_{\mathcal{I}_{2km-1}}, a_{\mathcal{I}_{2km-1}}) = \;& \mathbb{P}_1(x_{\mathcal{I}_1}, a_{\mathcal{I}_1}) \times \mathbb{P}_3(x_{\mathcal{I}_3}, a_{\mathcal{I}_3}|x_{\mathcal{I}_1}, a_{\mathcal{I}_1}) \times \cdots \\ & \times \mathbb{P}_{2km-1}(x_{\mathcal{I}_{2km-1}}, a_{\mathcal{I}_{2km-1}}|x_{\mathcal{I}_{2km-3}}, a_{\mathcal{I}_{2km-3}}).\end{aligned}$$

Let $\widetilde{\mathbb{P}}_i$ be the marginal distribution over the variables in block i, and let $\widetilde{\mathbb{P}}$ be the product of marginals. Then the difference between the distributions $\widetilde{\mathbb{P}}$ and $\mathbb{P}$ can be written as

$$\begin{aligned}\mathbb{P} - \widetilde{\mathbb{P}} = \;& \mathbb{P}_1 \otimes \mathbb{P}_3 \otimes \cdots \otimes \mathbb{P}_{2km-1} - \mathbb{P}_1 \otimes \widetilde{\mathbb{P}}_3 \cdots \otimes \widetilde{\mathbb{P}}_{2km-1} \\ = \;& \mathbb{P}_1 \otimes (\mathbb{P}_3 - \widetilde{\mathbb{P}}_3) \otimes \mathbb{P}_5 \otimes \cdots \otimes \mathbb{P}_{2km-1} \\ & + \mathbb{P}_1 \otimes \widetilde{\mathbb{P}}_3 \otimes (\mathbb{P}_5 - \widetilde{\mathbb{P}}_5) \otimes \mathbb{P}_7 \otimes \ldots \otimes \mathbb{P}_{2km-1} \\ & + \cdots \\ & + \mathbb{P}_1 \otimes \widetilde{\mathbb{P}}_3 \otimes \widetilde{\mathbb{P}}_5 \otimes \cdots \otimes \widetilde{\mathbb{P}}_{2km-3} \otimes (\mathbb{P}_{2km-1} - \widetilde{\mathbb{P}}_{2km-1}).\end{aligned}$$

Under β-mixing, since the gap between the blocks is of size b, we have that

$$\|\mathbb{P}_i(x_{\mathcal{I}_i}, a_{\mathcal{I}_i}|x_{\mathcal{I}_{i-2}}, a_{\mathcal{I}_{i-2}}) - \widetilde{\mathbb{P}}_i(x_{\mathcal{I}_i}, a_{\mathcal{I}_i})\|_1 \le \beta_b = \frac{2\gamma^b}{1-\gamma}.$$

Thus the difference between the joint distribution and the product of marginals is bounded as

$$\|\mathbb{P} - \widetilde{\mathbb{P}}\|_1 \le (km-1)\beta_b.$$

Block starts. Now let $\mathbb{P}$ denote the distribution over state-action pairs (x_t, a_t) for indices t corresponding to block starts, i.e. $t \in \{1, b+1, 2b+1, ..., (2km-1)b+1\}$. We again factorize the distribution over blocks:

$$\mathbb{P}(x_1, a_1, x_{b+1}, a_{b+1}, \ldots, x_{(2km-1)b+1}, a_{(2km-1)b+1}) = \mathbb{P}_1(x_1, a_1)\prod_{j=2}^{2km}\mathbb{P}_i(x_{jb+1}, a_{jb+1}|x_{(j-1)b+1}, a_{(j-1)b+1}).$$

Define a product-of-marginals distribution $\widetilde{\mathbb{P}} = \widetilde{\mathbb{P}}_1 \otimes \widetilde{\mathbb{P}}_2 \otimes \cdots \otimes \widetilde{\mathbb{P}}_{2km}$ over the block-start variables as follows. For odd j, let $\widetilde{\mathbb{P}}_j$ be the marginal of $\mathbb{P}$ over (x_{jb+1}, a_{jb+1}). For even j in phase i, let $\widetilde{\mathbb{P}}_j = \nu_{\pi_i}$ correspond to the stationary distribution of the policy π_i. Using the same notation as in Appendix A, let X_t be the indicator vector for (x_t, a_t) and let $H_{i:j}$ be the product of state-action transition matrices at times $i+1, ..., j$. For odd blocks j, we have

$$\|\mathbb{P}_j(\cdot|x_{(j-1)b+1}, a_{(j-1)b+1}) - \widetilde{\mathbb{P}}_j(\cdot)\|_1 = \|H_{(j-1)b+1:jb}^\top(X_{(j-1)b+1} - \widetilde{\mathbb{P}}_{j-1})\|_1 \le 2\gamma^{b-1}\,.$$

Slightly abusing notation, let H_{π_i} be the state-action transition matrix under policy π_i. For even blocks j in phase i, since they always follow an odd block in the same phase,

$$\|\mathbb{P}_j(\cdot|x_{(j-1)b+1}, a_{(j-1)b+1}) - \widetilde{\mathbb{P}}_j(\cdot)\|_1 = \|(H_{\pi_i}^{b-1})^\top(X_{(j-b)+1} - \nu_{\pi_i})\|_1 \le 2\gamma^{b-1}\,.$$

Thus, using a similar distribution decomposition as before, we have that $\|\mathbb{P} - \widetilde{\mathbb{P}}\|_1 \le 2(2km-1)\gamma^{b-1}$.