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 $\nu$, 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 $\{v_1, \ldots, v_n\}$ be the normalized left eigenvectors of $H$ corresponding to ordered eigenvalues $\{\lambda_1, \ldots, \lambda_n\}$. 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 $v_i^\top 1 = 0$. 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 $X_1^\top 1 = 1$ and $X_2^\top 1 = 1$, 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 $1^\top v_i = 0$ for all $i \geq 2$. Since

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

the inequality also holds for powers of $H$. $\square$

**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 $\pi_1, \ldots, \pi_k$ for $\tau$ steps each from an initial distribution $\nu_0$. 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 $i \in [T]$,

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

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 $t = 0, \ldots, T - 1$, $\mathbb{E}[X_{i+1} | X_i] = H_i^\top X_i$ and by the Markov property, for any $i \in [T]$,

$$B_i = \sum_{t=1}^i X_t + \sum_{t=i+1}^T \mathbb{E}[X_t | X_i] = \sum_{t=1}^i X_t + \sum_{t=i+1}^T H_i^\top X_i , \quad \text{and} \quad B_0 = \sum_{t=1}^T H_0^\top X_0 .$$

For any $i \in [T]$,

$$B_i - B_{i-1} = \sum_{t=1}^i X_t - \sum_{t=1}^{i-1} X_t + \sum_{t=i+1}^T H_i^\top X_i - \sum_{t=i+1}^T H_{i-1:t}^\top X_{i-1}$$

$$= \sum_{t=i}^T H_t^\top (X_t - H_{i-1:t}^\top X_{i-1}) . \quad (A.1)$$
Since \( X_i \) and \( H_i^\top X_{i-1} \) are distribution vectors, under Assumption 2.1 and using Lemma A.1,
\[
\|B_t - B_{t-1}\|_1 \leq \sum_{t=1}^T \|H_{t-k}(X_i - H_i^\top X_{i-1})\|_1 \leq 2\sum_{j=0}^{T-i} \gamma^j \leq 2(1 - \gamma)^{-1}.
\]

Let \((F_k)_k\) be a filtration and define \( \mathbb{E}_k[\cdot] := \mathbb{E}[\cdot | F_k] \). 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 \((F)_k\)-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 \( A_k^2 \leq A_k^2 \) almost surely. Let \( v = \| \sum_k A_k^2 \| \). Then with probability at least \( 1 - \delta \), \( \| \sum_k X_k \|_2 \leq 2v \log(m/\delta) \).

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 \( X \), \( D(X) = \begin{bmatrix} 0 & X^\top \\ X & 0 \end{bmatrix} \), which satisfies \( \lambda_{\text{max}}(D(X)) = \|X\| \) and \( D(X)^2 = \begin{bmatrix} X^2 & 0 \\ 0 & X^\top \end{bmatrix} \). In this case, we have that \( v = \max(\| \sum_k X_k X_k^\top \|, \| \sum_k X_k X_k^\top \|) \).

**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 \( F_i = \sigma(X_1, \ldots, X_i) \ i \in [n] \), such that \( X_i | F_{i-1} \) is zero-mean norm-subGaussian with \( \sigma_i \in \mathcal{F}_{i-1} \). That is:
\[
\mathbb{E}[X_i | F_{i-1} = 0], \quad P(\|X_i\| > t | F_{i-1}) \leq 2 \exp(-t^2/2\sigma_i^2) \quad \forall t \in \mathbb{R}, \forall i \in [n].
\]

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 \( \tau \geq \frac{\log T}{\pi \log(1/\gamma)} \) and that \( r(x, a) \in [0, 1] \) for all \( x, a \). Then, by choosing \( \eta = \sqrt[8]{8 \log |A| / \log(Q_{\text{max}} \cdot 2K)} \) we have with probability at least \( 1 - \delta \),
\[
\sum_{k=1}^K \sum_{t=(k-1)\tau+1}^{k\tau} (r_t - J_{\pi_k}) \leq 2(1 - \gamma)^{-1} \sqrt{2T \log(2/\delta)} + 2\sqrt{T} + (1 - \gamma)^{-2} \sqrt{8K \log |A|}.
\]

**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 \( \nu_t \) denote the state-action distribution at time \( t \), and \( \nu_{\pi} \) the stationary distribution of policy \( \pi \). Let \( \{B_t\}_{t=0}^\tau \) 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
\[
V_{T1} := \sum_{t=1}^T r^\top (X_t - \nu_t) = r^\top (B_T - B_0).
\]
Improved Regret Bound and Experience Replay in Regularized Policy Iteration

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

\[ V_{T1} \leq 2(1 - \gamma)^{-1}\sqrt{2T \log(2/\delta)}. \]  

(B.1)

For the second term we have

\[ V_{T2} := \sum_{k=1}^{K} \sum_{t=(k-1)\tau + 1}^{k\tau} r^\top (\nu_t - \nu_{\pi_k}) \]

\[ = \sum_{k=1}^{K} \sum_{t=1}^{\tau} (H^k_{\pi_k})^\top (\nu_{(k-1)\tau} - \nu_{\pi_k}) \]

\[ \leq \sum_{k=1}^{K} \|r\|_\infty \sum_{i=1}^{\tau} \| (H^k_{\pi_k})^\top (\nu_{(k-1)\tau} + \nu_{\pi_k}) - \nu_{\pi_k} \|_1 \]

\[ \leq \sum_{k=1}^{K} \|\nu_{(k-1)\tau} - \nu_{\pi_{k-1}}\|_1 + \| (H^k_{\pi_k})^\top \nu_{\pi_{k-1}} - \nu_{\pi_k} \|_1 \]

\[ \leq \sum_{k=1}^{K} \| (H^k_{\pi_{k-1}})^\top \nu_{(k-2)\tau} - \nu_{\pi_{k-1}}\|_1 + \gamma^{\tau} \|\nu_{\pi_{k-1}} - \nu_{\pi_k}\|_1 \]

\[ \leq 2T\gamma^{\tau} + \frac{1}{1 - \gamma} \sum_{k=1}^{K} \|\nu_{\pi_k} - \nu_{\pi_{k-1}}\|_1. \]

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 \(\hat{Q}_{\pi_k}(x, \cdot)\). 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

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

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 \(\|\hat{Q}_{\pi_k}\|_\infty \leq Q_{\text{max}}\), we can obtain

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

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

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

(B.2)

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

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

\[ \square \]
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 $i$:

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

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

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

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

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

Let $b_t = Q^b_\pi(x_t, a_t) - Q_\pi(x_t, a_t)$ denote the truncation bias at time $t$, and let $z_t = \sum_{i=t}^{i+b_t} r_i - X^\top H_\pi^{i-1} r$ denote the reward noise. We will write

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

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

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

We estimate the value function of each policy $\pi_i$ using data from phase $i$ as

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

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

$$\hat{w}_k - w_k = \frac{1}{km} \sum_{i=1}^{k} \tilde{M}_i^{-1} \phi_i (z_k + b_k + b(J_{\pi_i} - \tilde{J}_{\pi_i})) - \frac{\alpha}{km} \sum_{i=1}^{k} \tilde{M}_i^{-1} w_{\pi_i}.$$ 

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

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

$$\frac{\alpha}{km} \left\| \sum_{k=1}^{k} \tilde{M}_i^{-1} w_{\pi_i} \right\|_2 \leq 2\sigma^{-2} C_\omega (km)^{-1/2}. \quad (C.2)$$
Similarly, for
\[ b \geq \frac{\log((1 - \gamma)^{-1}\sqrt{k\bar{m}})}{\log(1/\gamma)}, \]  
(C.3)
the norm of the truncation bias term is upper-bounded as
\[ \frac{1}{k\bar{m}} \sum_{i=1}^{k} \sum_{t \in \mathcal{H}_i} \| \hat{M}_i^{-1} \phi_t b_t \|_2 \leq \frac{2\gamma b}{k\bar{m}(1 - \gamma)} \sum_{i=1}^{k} \sum_{t \in \mathcal{H}_i} \| \hat{M}_i^{-1} \phi_t \|_2 \leq 2\sigma^{-2}C_{\phi}(k\bar{m})^{-1/2}. \]  
(C.4)

To bound the error terms corresponding to reward noise \( z_t \) and average-error noise \( J_{x_t} - \hat{J}_{x_t} \), 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:
\[ \frac{1}{k\bar{m}} \left\| \sum_{i=1}^{k} \sum_{t \in \mathcal{H}_i} \hat{M}_i^{-1} \phi_t z_t \right\|_2 \leq 2c_1C_{\phi}\sigma^{-2} \sqrt{\frac{\log(2d/\delta)}{k\bar{m}}}, \]
\[ \frac{b}{k\bar{m}} \left\| \sum_{i=1}^{k} (J_{x_t} - \hat{J}_{x_t}) \sum_{t \in \mathcal{H}_i} \hat{M}_i^{-1} \phi_t \right\|_2 \leq 2c_2C_{\phi}\sigma^{-2}b \sqrt{\frac{\log(2d/\delta)}{k\bar{m}}}. \]

Thus, putting terms together, we have for an absolute constant \( c \), with probability at least \( 1 - \delta \),
\[ \| \hat{w}_k - w_k \|_2 \leq c\sigma^{-2}(C_w + C_{\phi})b \sqrt{\frac{\log(2d/\delta)}{k\bar{m}}}. \]

Note that this result holds for every \( k \in \lfloor K \rfloor \) and thus also holds for \( k = K \).

### C.1. Bounding \( \sum_{i=1}^{k} \hat{M}_i^{-1} \sum_{t \in \mathcal{H}_i} \phi_t z_t \)

Let \( \| \cdot \|_{tv} \) denote the total variation norm.

**Definition C.1** (\( \beta \)-mixing). Let \( \{Z_t\}_{t=1,2,...} \) be a stochastic process. Denote by \( Z_{1:t} \) the collection \( \{Z_1, \ldots, Z_t\} \), where we allow \( t = \infty \). Let \( \sigma(Z_{i:j}) \) denote the sigma-algebra generated by \( Z_{i:j} \) (\( i \leq j \)). The \( k^{th} \) \( \beta \)-mixing coefficient of \( \{Z_t\} \), \( \beta_k \), is defined by
\[ \beta_k = \sup_{t \geq 1} \mathbb{E} \sup_{B \in \sigma(Z_{t+k-\infty})} \| P(B|Z_{1:t}) - P(B) \|_{tv} \]
\[ = \sup_{t \geq 1} \mathbb{E} \| P(Z_{t+k-\infty}|Z_{1:t}) - P(Z_{t+k-\infty}) \|_{tv}. \]

\( \{Z_t\} \) is said to be \( \beta \)-mixing if \( \beta_k \rightarrow 0 \) as \( k \rightarrow \infty \). In particular, we say that a \( \beta \)-mixing process mixes at an exponential rate with parameters \( \beta, \alpha, \gamma > 0 \) if \( \beta_k \leq \beta \exp(-\alpha k^\gamma) \) 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 \( H_{i:i} = \prod_{j=i}^{t-1} H_j \), 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 \( \mathbb{E}[X_{t+k}] = H_{t:t+k}^\top \nu_0 \), where \( \nu_0 \) is the initial state distribution. Thus, under the uniform mixing Assumption 2.1, the \( k^{th} \) \( \beta \)-mixing coefficient is bounded as:
\[ \beta_k \leq \sup_{t \geq 1} \mathbb{E} \sum_{j=k}^{\infty} \| H_{t:t+j}^\top X_t - H_{t:t+j}^\top \nu_0 \|_1 \leq \sup_{t \geq 1} \mathbb{E} \sum_{j=k}^{\infty} \gamma^j \| X_t - H_{t:t}^\top \nu_0 \|_1 \leq \frac{2\gamma^k}{1 - \gamma}. \]

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 \( I_t \) denote the set of indices in the \( t^{th} \) block, and let \( x_{I_t}, a_{I_t} \) denote the corresponding states and actions. We factorize the joint distribution according to blocks:
\[ \mathbb{P}(x_{I_1}, a_{I_1}, x_{I_3}, a_{I_3}, \ldots, x_{I_{2km-1}}, a_{I_{2km-1}}) = \mathbb{P}_1(x_{I_1}, a_{I_1}) \times \mathbb{P}_3(x_{I_3}, a_{I_3}|x_{I_1}, a_{I_1}) \times \cdots \times \mathbb{P}_{2km-1}(x_{I_{2km-1}}, a_{I_{2km-1}}|x_{I_{2km-3}}, a_{I_{2km-3}}). \]
Let $\tilde{P}_i$ be the marginal distribution over the variables in block $i$, and let $\tilde{P}$ be the product of marginals of odd blocks. Corollary 2.7 of Yu (1994) implies that for any Borel-measurable set $E$,

$$|P(E) - \tilde{P}(E)| \leq (km - 1)\beta_b$$

(C.5)

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

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

$$\tilde{P}\left(\left\|\sum_{i=1}^{k} \tilde{M}_i^{-1}\sum_{t \in H_i} \phi_t z_t\right\|_2 \geq 2c_1C_\Phi\sigma^{-2}\sqrt{bkm \log(2d/\delta)}\right) \leq \delta.$$

Thus, using (C.5),

$$P\left(\left\|\sum_{i=1}^{k} \tilde{M}_i^{-1}\sum_{t \in H_i} \phi_t z_t\right\|_2 \geq 2c_1C_\Phi\sigma^{-2}\sqrt{bkm \log(2d/\delta)}\right) \leq \delta + (km - 1)\beta_b.$$

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

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

(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 $1 - 2\delta$,

$$\left\|\sum_{i=1}^{k} \tilde{M}_i^{-1}\sum_{t \in H_i} \phi_t (J_{\pi_i} - \tilde{J}_{\pi_i})\right\|_2 \leq 2c_1C_\Phi\sigma^{-2}\sqrt{bkm \log(2d/\delta)}.$$

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

Recall that the average-reward estimates $\tilde{J}_{\pi_i}$ are computed using time indices corresponding to the starts of even blocks, $\tilde{T}_i$. Thus this error term is only a function of the indices corresponding to block starts. Now let $P$ denote the distribution over state-action pairs $(x_t, a_t)$ for indices $t$ corresponding to block starts, i.e. $t \in \{1, b + 1, 2b + 1, \ldots, (2km - 1)b + 1\}$. We again factorize the distribution over blocks as $P = P_1 \otimes P_2 \otimes \cdots \otimes P_{2km}$. Let $\tilde{P} = \tilde{P}_1 \otimes \tilde{P}_2 \otimes \cdots \otimes \tilde{P}_{2km}$ be a product-of-marginals distribution defined as follows. For odd $j$, let $\tilde{P}_j$ be the marginal of $P$ over $(x_{jb+1}, a_{jb+1})$. For even $j$ in phase $i$, let $\tilde{P}_j = \nu_{\pi_i}$, correspond to the stationary distribution of the corresponding policy $\pi_i$. Using arguments similar to independent blocks, we show in Appendix E that $||P - \tilde{P}||_1 \leq 2(2km - 1)\gamma^{b^{-1}}$.

Let $\tilde{E}$ denote expectation w.r.t. the product-of-marginals distribution $\tilde{P}$. Then $\tilde{E}[\tilde{M}_i^{-1}\sum_{t \in H_i} \phi_t (J_{\pi_i} - \tilde{J}_{\pi_i})] = 0$, since under $\tilde{P}$, $\tilde{J}_{\pi_i}$ is the sum of rewards for state-action pairs distributed according to $\nu_{\pi_i}$, 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 $1 - 2\delta$,

$$\left\|\sum_{i=1}^{k} \tilde{M}_i^{-1}\sum_{t \in H_i} \phi_t (J_{\pi_i} - \tilde{J}_{\pi_i})\right\|_2 \leq 2c_2C_\Phi\sigma^{-2}\sqrt{km \log(2d/\delta)}.$$
C.3. Bounding $\|\tilde{M}_i^{-1}\|_2$

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

Let $\Phi$ be a $|X||A| \times d$ matrix of all features. Let $D_i = \text{diag}(\nu_i)$, and let $\hat{D}_i = \text{diag}(\sum_{t \in H_i} X_t)$, where $X_t$ is a state-action indicator as in Lemma A.2. Let $M_i = \Phi^T D_i \Phi + \alpha m^{-1} I$. We can write $\tilde{M}_i^{-1}$ as

$$\tilde{M}_i^{-1} = (\Phi^T \hat{D}_i \Phi + \alpha^{-1} I + \Phi^T (D_i - D_i))^{-1}$$

$$= (M_i + \Phi^T (D_i - D_i))^{-1}$$

$$= (I + M_i^{-1} \Phi^T (D_i - D_i))^{-1} M_i^{-1}$$

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 $1 - \delta$,

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

Thus

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

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

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

For any matrix $A$,

$$\|\Phi^T A\|^2 = \left\|\sum_{i,j} A_{ij} \phi_i^T \phi_j^T\right\|_2^2 \leq \sum_{i,j} |A_{ij}| \|\phi_i \phi_j^T\|^2 \leq C_d^2 \sum_{i,j} |A_{ij}| = C_d^2 \|A\|_{1,1}.$$  \hfill (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,

$$\|\Phi^T (\hat{D}_i - D_i)\Phi\|_2 = \frac{1}{m} \sum_{t \in H_i} \Phi^T \text{diag}(X_t - \nu_t + \nu_t - \nu_{\pi_t}) \Phi$$

$$\leq \frac{1}{m} \left\|\sum_{t \in H_i} \Phi^T \text{diag}(X_t - \nu_t) \Phi\right\|_2 + \frac{C_d^2}{m} \sum_{t \in H_i} \|\nu_t - \nu_{\pi_t}\|_1.$$  

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

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 $H_i$. Note that $\sum_{t \in H_i} \Phi^T \text{diag}(X_t - \nu_t) \Phi = \Phi^T \text{diag}(B_m - B_0) \Phi$. Thus we can use matrix-Azuma to bound the difference sequence. Given that

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

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

$$\|\Phi^T (\hat{D}_i - D_i)\Phi\|_2 \leq 4m^{-1/2}C_d^2(1 - \gamma)^{-1}\sqrt{2\log(d/\delta)} + 2m^{-1}C_d^2(1 - \gamma)^{-1}$$

$$\leq 6m^{-1/2}C_d^2(1 - \gamma)^{-1}\sqrt{2\log(d/\delta)}.$$

D. Bounding $\|V_K - \hat{V}_K\|_{\mu_\pi}$

We write the value function error as follows:

$$\mathbb{E}_{x \sim \mu_\pi}[\hat{V}_K(x) - V_K(x)] = \sum_x \mu_\pi(x) \sum_a \phi(x, a) \top \frac{1}{K} \sum_{i=1}^K \pi_i(a|x)(\hat{\omega}_{\pi_i} - \omega_{\pi_i})$$

$$\leq \frac{1}{K} \sum_x \mu_\pi(x) \sum_a \|\phi(x, a)\|_2 \sum_{i=1}^K \|\pi_i(a|x)(\hat{\omega}_{\pi_i} - \omega_{\pi_i})\|_2.$$
Note that for any set of scalars \( \{p_t\}_t^T \) with \( p_t \in [0, 1] \), the term \( \| \sum_t^T p_t (\hat{w}_t - w_t) \|_2 \) has the same upper bound as \( \| \sum_t^T (\hat{w}_t - w_t) \|_2 \). 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 \( \{p_t\}_t^T \) 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 \( \{p_t\}_t^T \) equal to 1. Thus, reusing the results of the previous section, we have

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

**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 \( P \) denote the joint distribution of state-action pairs in odd blocks. Let \( I_i \) denote the set of indices in the \( i \)th block, and let \( x_I, a_I \) denote the corresponding states and actions. We factorize the joint distribution according to blocks:

\[
P(x_{I_1}, a_{I_1}, x_{I_3}, a_{I_3}, \ldots, x_{I_{2km-1}}, a_{I_{2km-1}}) = P_1(x_{I_1}, a_{I_1}) \times P_3(x_{I_3}, a_{I_3}) x_{I_1}, a_{I_1}) \times \cdots \times P_{2km-1}(x_{I_{2km-1}}, a_{I_{2km-1}} | x_{I_{2km-3}}, a_{I_{2km-3}}).
\]

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

\[
P - \bar{P} = P_1 \otimes P_3 \otimes \cdots \otimes P_{2km-1} - P_1 \otimes \bar{P}_3 \otimes \cdots \otimes \bar{P}_{2km-1}
\]

\[
= P_1 \otimes (P_3 - \bar{P}_3) \otimes P_5 \otimes \cdots \otimes P_{2km-1}
\]

\[
+ P_1 \otimes \bar{P}_3 \otimes (P_5 - \bar{P}_5) \otimes P_7 \otimes \cdots \otimes P_{2km-1}
\]

\[
+ \cdots
\]

\[
+ P_1 \otimes \bar{P}_3 \otimes \bar{P}_5 \otimes \cdots \otimes \bar{P}_{2km-3} \otimes (P_{2km-1} - \bar{P}_{2km-1}).
\]

Under \( \beta \)-mixing, since the gap between the blocks is of size \( b \), we have that

\[
\| P_i(x_{I_i}, a_{I_i}, x_{I_i-2}, a_{I_i-2}) - \bar{P}_i(x_{I_i}, a_{I_i}) \|_1 \leq \beta_b = \frac{2\gamma}{1 - \gamma}.
\]

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

\[
\| P - \bar{P} \|_1 \leq (km - 1) \beta_b.
\]

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

\[
P(x_1, a_1, x_{b+1}, a_{b+1}, \ldots, x_{(2km-1)b+1}, a_{(2km-1)b+1}) = P_1(x_1, a_1) \prod_{j=2}^{2km} P_j(x_{jb+1}, a_{jb+1}| x_{(j-1)b+1}, a_{(j-1)b+1}).
\]

Define a product-of-marginals distribution \( \bar{P} = \bar{P}_1 \otimes \bar{P}_2 \otimes \cdots \otimes \bar{P}_{2km} \) over the block-start variables as follows. For odd \( j \), let \( \bar{P}_j = \nu_{\pi_t} \) correspond to the stationary distribution of the policy \( \pi_t \). 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 \( i + 1, \ldots, j \). For odd blocks \( j \), we have

\[
\| P_j(|x_{(j-1)b+1}, a_{(j-1)b+1}) - \bar{P}_j(|\cdot) \|_1 = \| H_{(j-1)b+1;jb}(X_{(j-1)b+1} - \bar{P}_{j-1}) \|_1 \leq 2\gamma^{-b-1}.
\]

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

\[
\| P_j(|x_{(j-1)b+1}, a_{(j-1)b+1}) - \bar{P}_j(|\cdot) \|_1 = \| (H_{\pi_t}^{-1})^T (X_{(j-1)b+1} - \nu_{\pi_t}) \|_1 \leq 2\gamma^{-b-1}.
\]

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