A Proof of Lemma 1

Proof. Let us write $\tilde{\mu}^{t+1} = \Pi_{\mathcal{U},KL}\left(\tilde{\mu}^{t+1/2}\right)$ where $\tilde{\mu}^{t+1/2}$ is the update vector prior to the projection step. Denote by $(i_t, u_t, s_t, a_t, s_t', r_t)$ the sample at iteration t. Define the vector $\Delta^{t+1} \in \mathbb{R}^{D\times U}$ to be $\Delta^{t+1}_{i_t,u_t} = \frac{\Phi_{s_t'*}\tilde{v}^t - \Phi_{s_t*}\tilde{v}^t + r_t - M}{\tilde{\mu}^t_{i_t,u_t}}$ and $\Delta^{t+1}_{i,u} = 0$ for all $(i,u) \neq (i_t,u_t)$. Then the vector $\tilde{\mu}^{t+1/2}$ can be equivalently written as

$$\tilde{\mu}^{t+1/2}_{i,u} = \frac{\tilde{\mu}^t_{i,u} \cdot \exp(\beta\Delta^{t+1}_{i,u})}{\sum_{i',u'} \tilde{\mu}^t_{i'u'} \cdot \exp(\beta\Delta^{t+1}_{i',u'})}, \quad \forall i \in 1,\ldots,D, u \in 1,\ldots,U.$$

Recall that $\check{v} = \mathrm{argmin}_{\tilde{v}\in\mathcal{V}}\|\Phi\tilde{v} - v^*\|_\infty$ and $\check{\mu} = \mathrm{argmin}_{\tilde{\mu}\in\mathcal{U}}\|\Phi\tilde{\mu}\Psi^\top - \mu^*\|_{1,1}$. We obtain that

$$\begin{aligned}
D_{KL}(\check{\mu}\|\tilde{\mu}^{t+1/2}) - D_{KL}(\check{\mu}\|\tilde{\mu}^t) &= \sum_{i=1}^D\sum_{u=1}^U \check{\mu}_{i,u}\log\frac{\check{\mu}_{i,u}}{\tilde{\mu}^{t+1/2}_{i,u}} - \sum_{i=1}^D\sum_{u=1}^U \check{\mu}_{i,u}\log\frac{\check{\mu}_{i,u}}{\tilde{\mu}^t_{i,u}} \\
&= \sum_{i=1}^D\sum_{u=1}^U \check{\mu}_{i,u}\log\frac{\tilde{\mu}^t_{i,u}}{\tilde{\mu}^{t+1/2}_{i,u}} \\
&= \sum_{i=1}^D\sum_{u=1}^U \check{\mu}_{i,u}\log\frac{Z}{\exp\left(\beta\Delta^{t+1}_{i,u}\right)} \\
&= \log Z - \beta\sum_{i=1}^D\sum_{u=1}^U \check{\mu}_{i,u}\Delta^{t+1}_{i,u},
\end{aligned}$$

where we let $Z = \sum_{i=1}^D\sum_{u=1}^U \tilde{\mu}^t_{i,u}\cdot\exp(\beta\Delta^{t+1}_{i,u})$. According to the definition of $\mathcal{V}$, we have $|\Phi_{s*}\tilde{v}^t| \le 2t_{mix}$ for all state s. Combining with our choice of $M = 4t_{mix}+1$, we have $\Delta^{t+1}_{i,u} \le 0$ for all $i = 1,\ldots,D$ and $u = 1,\ldots,U$. Consequently, applying the inequalities $e^x \le 1 + x + \frac{1}{2}x^2$ for all $x \le 0$ and $\log(1+x) \le x$ for all $x > -1$, we have

$$\begin{aligned}
\log Z = \log\sum_{i=1}^D\sum_{u=1}^U \tilde{\mu}^t_{i,u}\cdot\exp(\beta\Delta^{t+1}_{i,u}) &\le \log\sum_{i=1}^D\sum_{u=1}^U \tilde{\mu}^t_{i,u}\left(1 + \beta\Delta^{t+1}_{i,u} + \frac{\beta^2}{2}(\Delta^{t+1}_{i,u})^2\right) \\
&= \log\left(1 + \beta\sum_{i=1}^D\sum_{u=1}^U \tilde{\mu}^t_{i,u}\Delta^{t+1}_{i,u} + \frac{\beta^2}{2}\sum_{i=1}^D\sum_{u=1}^U \tilde{\mu}^t_{i,u}(\Delta^{t+1}_{i,u})^2\right) \\
&\le \beta\sum_{i=1}^D\sum_{u=1}^U \tilde{\mu}^t_{i,u}\Delta^{t+1}_{i,u} + \frac{\beta^2}{2}\sum_{i=1}^D\sum_{u=1}^U \tilde{\mu}^t_{i,u}(\Delta^{t+1}_{i,u})^2
\end{aligned}$$

Combining the above results, we have

$$D_{KL}(\check{\mu}\|\tilde{\mu}^{t+1/2}) - D_{KL}(\check{\mu}\|\tilde{\mu}^t) \le \beta\sum_{i=1}^D\sum_{u=1}^U (\tilde{\mu}^t_{i,u} - \check{\mu}_{i,u})\Delta^{t+1}_{i,u} + \frac{\beta^2}{2}\sum_{i=1}^D\sum_{u=1}^U \tilde{\mu}^t_{i,u}(\Delta^{t+1}_{i,u})^2. \tag{A.1}$$

In order to prove Lemma 1, we now show that $\mathbf{E}[\Delta^{t+1}_{i,u} \mid \mathcal{F}_t] = \sum_{a\in\mathcal{A}}\Psi_{a,u}\Phi^\top_{*i}((P_a - I)\Phi\tilde{v}^t + r_a - M\cdot\mathbf{1}_S)$ and that $\sum_{i=1}^D\sum_{u=1}^U \tilde{\mu}^t_{i,u}\mathbf{E}[\left(\Delta^{t+1}_{i,u}\right)^2 \mid \mathcal{F}_t] \le 100DUt^2_{mix}$. We use $\mathbf{1}_S$ to denote the all one column vector with dimension S. Recall that (i_t,u_t) is sampled from $\tilde{\mu}^t$, s_t is sampled from ϕ_{i_t}, a_t is sampled from ψ_{u_t} and s_t'

is sampled from $P_{u_t}(s_t,\cdot)$. Hence, for all (i,u), we have

$$\begin{aligned}\mathbf{E}[\Delta_{i,u}^{t+1}\mid\mathcal{F}_t] &= \tilde{\mu}_{i,u}^t\sum_{a\in\mathcal{A}}\sum_{s\in\mathcal{S}}\sum_{s'\in\mathcal{S}}\Psi_{a,u}\cdot\Phi_{s,i}\cdot P_a(s,s')\cdot\frac{\Phi_{s'*}\tilde{v}^t+r_a(s)-\Phi_{s*}\tilde{v}^t-M}{\tilde{\mu}_{i,u}^t}\\ &=\sum_{a\in\mathcal{A}}\sum_{s\in\mathcal{S}}\Psi_{a,u}\Phi_{s,i}\left(P_a(s,\cdot)\Phi\tilde{v}^t+r_a(s)-\Phi_{s*}\tilde{v}^t-M\right)\\ &=\sum_{a\in\mathcal{A}}\Psi_{a,u}\Phi_{*i}^\top\left(P_a\Phi\tilde{v}^t+r_a-\Phi\tilde{v}^t-M\cdot\mathbf{1}_S\right).\end{aligned}$$

It remains to prove that $\sum_{i=1}^D\sum_{u=1}^U\tilde{\mu}_{i,u}^t\mathbf{E}[\left(\Delta_{i,u}^{t+1}\right)^2\mid\mathcal{F}_t]\le 100DUt_{mix}^2$. Expanding the expectation, we have

$$\begin{aligned}&\sum_{i=1}^D\sum_{u=1}^U\tilde{\mu}_{i,u}^t\mathbf{E}[\left(\Delta_{i,u}^{t+1}\right)^2\mid\mathcal{F}_t]\\ &=\sum_{i=1}^D\sum_{u=1}^U\tilde{\mu}_{i,u}^t\sum_{a\in\mathcal{A}}\sum_{s\in\mathcal{S}}\sum_{s'\in\mathcal{S}}\Psi_{a,u}\cdot\tilde{\mu}_{i,u}^t\cdot\Phi_{s,i}\cdot P_a(s,s')\left(\frac{\Phi_{s'*}\tilde{v}^t+r_a(s)-\Phi_{s*}\tilde{v}^t-M}{\tilde{\mu}_{i,u}^t}\right)^2\\ &=\sum_{i=1}^D\sum_{u=1}^U\sum_{a,s,s'}\Psi_{a,u}\cdot\Phi_{s,i}\cdot P_a(s,s')(\Phi_{s'*}\tilde{v}^t+r_a(s)-\Phi_{s*}\tilde{v}^t-M)^2\\ &\le\sum_{i=1}^D\sum_{u=1}^U\sum_{a,s,s'}\Psi_{a,u}\cdot\Phi_{s,i}\cdot P_a(s,s')(8t_{mix}+2)^2\\ &=DU(8t_{mix}+2)^2\le 100DUt_{mix}^2,\end{aligned}$$

where the first inequality uses the relation that $|\Phi_{s'*}\tilde{v}^t+r_a(s)-\Phi_{s*}\tilde{v}^t-M|\le 8t_{mix}+2$, the third equality is due to that $\sum_{a,s,s'}\Psi_{a,u}\cdot\Phi_{s,i}\cdot P_a(s,s')=1$ and the last inequality is because $t_{mix}\ge 1$. Substituting the above abounds in equation (A.1), we obtain that

$$\begin{aligned}&\mathbf{E}[D_{KL}(\check{\mu}\|\tilde{\mu}^{t+1/2})\mid\mathcal{F}_t]-D_{KL}(\check{\mu}\|\tilde{\mu}^t)\\ &\le\beta\sum_{a\in\mathcal{A}}\sum_{i=1}^D\sum_{u=1}^U(\tilde{\mu}_{i,u}^t-\check{\mu}_{i,u})\Psi_{a,u}\Phi_{*i}^\top((P_a-I)\Phi\tilde{v}^t+r_a-M\cdot\mathbf{1}_S)+\frac{\beta^2}{2}\cdot 100DUt_{mix}^2\\ &\le\beta\sum_{a\in\mathcal{A}}\Psi_{a*}(\tilde{\mu}^t-\check{\mu})^\top\Phi^\top((P_a-I)\Phi\tilde{v}^t+r_a)+50\beta^2DUt_{mix}^2,\end{aligned}$$

where the last inequality is due to that

$$\sum_{a\in\mathcal{A}}\Psi_{a*}(\tilde{\mu}^t)^\top\Phi^\top\mathbf{1}_S=\sum_{a\in\mathcal{A}}\Psi_{a*}(\check{\mu})^\top\Phi^\top\mathbf{1}_S=1.$$

Recall that $\tilde{\mu}^{t+1}=\Pi_{\mathcal{U},KL}\left(\tilde{\mu}^{t+1/2}\right)=\mathrm{argmin}_{\mu'\in\mathcal{U}}D_{KL}(\mu'\|\tilde{\mu}^{t+1/2})$ and $\mathcal{U}$ is a convex set. By the property of information projection with regard to KL divergence (see [1] Theorem 11.6.1 on page 367), we have

$$\mathbf{E}[D_{KL}(\check{\mu}\|\tilde{\mu}^{t+1})\mid\mathcal{F}_t]\le\mathbf{E}[D_{KL}(\check{\mu}\|\tilde{\mu}^{t+1/2})\mid\mathcal{F}_t].$$

Combining the above inequalities, we conclude that

$$\begin{aligned}&\mathbf{E}[D_{KL}(\check{\mu}\|\tilde{\mu}^{t+1})\mid\mathcal{F}_t]-D_{KL}(\check{\mu}\|\tilde{\mu}^t)\le\mathbf{E}[D_{KL}(\check{\mu}\|\tilde{\mu}^{t+1/2})\mid\mathcal{F}_t]-D_{KL}(\check{\mu}\|\tilde{\mu}^t)\\ &\le\beta\sum_{a\in\mathcal{A}}\Psi_{a*}(\tilde{\mu}^t-\check{\mu})^\top\Phi^\top((P_a-I)\Phi\tilde{v}^t+r_a)+50\beta^2DUt_{mix}^2,\end{aligned}$$

Finally, observe that

$$D_{KL}(\check{\mu}\|\tilde{\mu}^1) = \sum_{i=1}^{D}\sum_{u=1}^{U}\check{\mu}_{i,u}\log\frac{\check{\mu}_{i,u}}{1/(DU)} = \sum_{i=1}^{D}\sum_{u=1}^{U}\check{\mu}_{i,u}\log(DU) + \sum_{i=1}^{D}\sum_{u=1}^{U}\check{\mu}_{i,u}\log(\check{\mu}_{i,u}) \le \log(DU),$$

where the last inequality is due to that $\check{\mu}_{i,u} \le 1$ and thus $\log(\check{\mu}_{i,u}) \le 0$ for all i, u. To this point, we complete the proof of Lemma 1. $\square$

B Proof of Lemma 2

Proof. Let $(i_t, u_t, s_t, a_t, s'_t, r_t)$ be the sample at iteration t. Throughout the proof, we use the shorthand $\Delta^{t+1} \triangleq \Phi_{s'_t*}^\top - \Phi_{s_t*}^\top$. According to the update of Algorithm 1, we have $\tilde{v}^{t+1} = \Pi_{\mathcal{V}}(\tilde{v}^t - \alpha\Delta^{t+1})$. By using the nonexpansize property of $\Pi_{\mathcal{V}}$, we obtain that

$$\begin{aligned}\mathbf{E}\left[\|\tilde{v}^{t+1} - \check{v}\|_2^2 \mid \mathcal{F}_t\right] &= \mathbf{E}\left[\|\Pi_{\mathcal{V}}(\tilde{v}^t - \alpha\Delta^{t+1}) - \check{v}\|_2^2 \mid \mathcal{F}_t\right] \le \mathbf{E}\left[\|\tilde{v}^t - \alpha\Delta^{t+1} - \check{v}\|_2^2 \mid \mathcal{F}_t\right] \\ &= \|\tilde{v}^t - \check{v}\|_2^2 - 2\alpha\mathbf{E}[(\Delta^{t+1})^\top \mid \mathcal{F}_t](\tilde{v}^t - \check{v}) + \alpha^2\mathbf{E}[\|\Delta^{t+1}\|_2^2 \mid \mathcal{F}_t].\end{aligned} \quad \text{(B.1)}$$

Recall that (i_t, u_t) is sampled from $\tilde{\mu}^t$, a_t is sampled from ψ_{u_t}, s_t is sampled from ϕ_{i_t} and s'_t is sampled from $P_{a_t}(s_t, \cdot)$. We can expand the expectation of $\mathbf{E}[(\Delta^{t+1}) \mid \mathcal{F}_t]$ to obtain that

$$\begin{aligned}\mathbf{E}[(\Delta^{t+1})^\top \mid \mathcal{F}_t] &= \sum_{a\in\mathcal{A}}\sum_{i=1}^{D}\sum_{u=1}^{U}\sum_{s\in\mathcal{S}}\sum_{s'\in\mathcal{S}}\Psi_{a,u}\tilde{\mu}_{i,u}^t\Phi_{s,i}P_a(s,s')(\Phi_{s'*} - \Phi_{s*}) \\ &= \sum_{a\in\mathcal{A}}\sum_{i=1}^{D}\sum_{u=1}^{U}\sum_{s\in\mathcal{S}}\Psi_{a,u}\tilde{\mu}_{i,u}^t\Phi_{s,i}(P_a(s,\cdot)\Phi - \Phi_{s*}) = \sum_{a\in\mathcal{A}}\sum_{i=1}^{D}\sum_{u=1}^{U}\Psi_{a,u}\tilde{\mu}_{i,u}^t\Phi_{*i}^\top(P_a\Phi - \Phi) \\ &= \sum_{a\in\mathcal{A}}\sum_{u=1}^{U}\Psi_{a,u}(\tilde{\mu}_{*u}^t)^\top\Phi^\top(P_a\Phi - \Phi) = \sum_{a\in\mathcal{A}}\Psi_{a*}(\tilde{\mu}^t)^\top\Phi^\top(P_a\Phi - \Phi).\end{aligned}$$

Next we prove that $\mathbf{E}[\|\Delta^{t+1}\|_2^2 \mid \mathcal{F}_t] \le \|\Phi\|_{2,\infty}^2$. A straightforward calculation yields that

$$\begin{aligned}\mathbf{E}[\|\Delta^{t+1}\|_2^2 \mid \mathcal{F}_t] &= \sum_{a\in\mathcal{A}}\sum_{i=1}^{D}\sum_{u=1}^{U}\sum_{s\in\mathcal{S}}\sum_{s'\in\mathcal{S}}\Psi_{a,u}\tilde{\mu}_{i,u}^t\Phi_{s,i}P_a(s,s')\|\Phi_{s'*} - \Phi_{s*}\|_2^2 \\ &\le \sum_{a\in\mathcal{A}}\sum_{i=1}^{D}\sum_{u=1}^{U}\sum_{s\in\mathcal{S}}\sum_{s'\in\mathcal{S}}\Psi_{a,u}\tilde{\mu}_{i,u}^t\Phi_{s,i}P_a(s,s')(2\|\Phi_{s'*}\|_2^2 + 2\|\Phi_{s*}\|_2^2) \\ &\le \sum_{a\in\mathcal{A}}\sum_{i=1}^{D}\sum_{u=1}^{U}\sum_{s\in\mathcal{S}}\sum_{s'\in\mathcal{S}}\Psi_{a,u}\tilde{\mu}_{i,u}^t\Phi_{s,i}P_a(s,s')(4\|\Phi\|_{2,\infty}^2) = 4\|\Phi\|_{2,\infty}^2,\end{aligned}$$

where the last equality is due to that $\tilde{\mu}$, ψ_u and ϕ_i are distributions and $\sum_{i,u,a,s,s'}\Psi_{a,u}\tilde{\mu}_{i,u}^t\Phi_{s,i}P_a(s,s') = 1$. Substituting the above bounds into equation (B.1), we get the first part of Lemma 2.

It remains to show that $\|\tilde{v}^1 - \check{v}\|_2^2 = \|\check{v}\|_2^2 \le \frac{4Dt_{mix}^2\|\Phi\|_1^2}{\lambda_{\min}^2(\Phi^\top\Phi)}$. Let $v' \triangleq \Phi\check{v}$. Multiply v' by $\Phi^\top$ and we get $\Phi^\top v' = \Phi^\top\Phi\check{v}$. Hence, by Assumption 1 that $\Phi^\top\Phi$ is invertible, we have

$$\check{v} = (\Phi^\top\Phi)^{-1}\Phi^\top v'$$

By our definition of $\check{v}$ and $\mathcal{V}$, we have $\|v'\|_\infty \le 2t_{mix}$. Using the relation that $\lambda_{\max}((\Phi^\top\Phi)^{-1}) = \frac{1}{\lambda_{\min}(\Phi^\top\Phi)}$ where $\lambda_{\max}$ and $\lambda_{\min}$ denotes the largest and the smallest eigenvalue, we obtain

$$\begin{aligned}\|\check{v}\|_2^2 &\le \|(\Phi^\top\Phi)^{-1}\|_2^2\|\Phi^\top v'\|_2^2 \le \frac{1}{\lambda_{\min}^2(\Phi^\top\Phi)}\cdot 4t_{mix}^2\cdot\|\Phi^\top\|_{1,2}^2 \\ &\le \frac{4t_{mix}^2}{\lambda_{\min}^2(\Phi^\top\Phi)}\cdot D\cdot\|\Phi^\top\|_{1,\infty}^2 = \frac{4t_{mix}^2 D\|\Phi\|_1^2}{\lambda_{\min}^2(\Phi^\top\Phi)}.\end{aligned}$$

As a result, we have $\|\breve{v}\|_2^2 \leq \frac{4t_{mix}^2 D\|\Phi\|_1^2}{\lambda_{\min}^2(\Phi^\top\Phi)}$. Recall that every column of Φ is a distribution and thus $\|\Phi\|_1 = 1$. Using this relationship, we obtain that $\|\breve{v}\|_2^2 \leq \frac{4t_{mix}^2 D}{\lambda_{\min}^2(\Phi^\top\Phi)}$. □

C Proof of Theorem 4

Proof. All the norms used in the proof of Theorem 4 are matrix norms. For a matrix Φ of size $m \times n$, the matrix p-norm for $1 \leq p \leq \infty$ is defined as $\|\Phi\|_p = \max\{\|\Phi v\|_p : v \in \mathbb{R}^n \text{ with } \|v\|_p = 1\}$. Especially, $\|\Phi\|_1$ is the maximum absolute column sum and $\|\Phi\|_\infty$ is the maximum absolute row sum.

We begin by analyzing the behavior of the duality gap in Theorem 2. By some algebra, we can rewrite the LFS of equation (9) as

$$\begin{aligned}
&\sum_{a\in\mathcal{A}} r_a^\top \mu_{*a}^* + \frac{1}{T}\sum_{t=1}^T \mathbf{E}\Big[\sum_{a\in\mathcal{A}} ((I-P_a)v^* - r_a)^\top \Phi\tilde{\mu}^t\Psi_{a*}^\top\Big] \\
&\underbrace{-\frac{1}{T}\sum_{t=1}^T \mathbf{E}\left[\sum_{a\in\mathcal{A}} (\Phi\breve{\mu}\Psi_{a*}^\top)^\top (I-P_a)\Phi\tilde{v}^t\right]}_{\text{(i)}} + \underbrace{\sum_{a\in\mathcal{A}} (\Phi\breve{\mu}\Psi_{a*}^\top - \mu_{*a}^*)^\top r_a}_{\text{(ii)}} \\
&\underbrace{+\frac{1}{T}\sum_{t=1}^T \mathbf{E}\Big[\sum_{a\in\mathcal{A}} (\Phi\tilde{\mu}^t\Psi_{a*}^\top)^\top (I-P_a)(\Phi\breve{v} - v^*)\Big]}_{\text{(iii)}},
\end{aligned} \tag{C.1}$$

where μ_{*a}^* is the a-th column of μ^*. Next, we bound (i), (ii), (iii) respectively.

Analysis of (i): Recall that the stationary distribution μ^* satisfies the condition $\sum_{a\in\mathcal{A}} (\mu_{*a}^*)^\top (I-P_a) = \mathbf{0}_S$. So we can bound (i) by

$$\begin{aligned}
|\text{(i)}| &\leq \left\|\sum_{a\in\mathcal{A}} (\Phi\breve{\mu}\Psi_{a*}^\top - \mu_{*a}^*)^\top (I-P_a)\right\|_\infty \left\|\frac{1}{T}\sum_{t=1}^T \mathbf{E}[\Phi\tilde{v}^t]\right\|_\infty \\
&\leq \sum_{a\in\mathcal{A}} \left\|(\Phi\breve{\mu}\Psi_{a*}^\top - \mu_{*a}^*)^\top\right\|_\infty (\|I\|_\infty + \|P_a\|_\infty)\cdot 2t_{mix} \\
&\leq 4t_{mix}\|\Phi\breve{\mu}\Psi^\top - \mu^*\|_{1,1},
\end{aligned}$$

where the first inequality is due to that $\|\Phi_1\Phi_2\|_\infty \leq \|\Phi_1\|_\infty\|\Phi_2\|_\infty$ for two matrices Φ_1 and Φ_2, the second inequality is due to that $\|\Phi\tilde{v}^t\|_\infty \leq 2t_{mix}$ for all t (see Lemma 1 in [2]). In the third inequality, we use the fact that the matrix ∞-norm of a row vector is the sum of its components. And thus we have $\sum_{a\in\mathcal{A}} \left\|(\Phi\breve{\mu}\Psi_{a*}^\top - \mu_{*a}^*)^\top\right\|_\infty = \|\Phi\breve{\mu}\Psi^\top - \mu^*\|_{1,1}$.

Analysis of (ii): Using the inequality that $\|\Phi_1\Phi_2\|_\infty \leq \|\Phi_1\|_\infty\|\Phi_2\|_\infty$ for two *matrices* Φ_1, Φ_2, we have

$$|\text{(ii)}| \leq \sum_{a\in\mathcal{A}} \|(\Phi\breve{\mu}\Psi_{a*}^\top - \mu_{*a}^*)^\top\|_\infty \|r_a\|_\infty \leq \|\Phi\breve{\mu}\Psi^\top - \mu^*\|_{1,1},$$

where the last inequality is due to that all the rewards are bounded between 0 and 1.

Analysis of (iii): We note that for any iteration t, $\sum_{a\in\mathcal{A}} (\Phi\tilde{\mu}^t\Psi_{a*}^\top)^\top I$ and $\sum_{a\in\mathcal{A}} (\Phi\tilde{\mu}^t\Psi_{a*}^\top)^\top P_a$ are two row vectors that both sum to 1. Recall that the matrix ∞-norm of a row vector is the sum of its components. Thus, we have $\|\sum_{a\in\mathcal{A}} (\Phi\tilde{\mu}^t\Psi_{a*}^\top)^\top I - \sum_{a\in\mathcal{A}} (\Phi\tilde{\mu}^t\Psi_{a*}^\top)^\top P_a\|_\infty \leq 2$. As a result, we have

$$\begin{aligned}
|\text{(iii)}| &\leq \left\|\frac{1}{T}\sum_{t=1}^T \mathbf{E}\Big[\sum_{a\in\mathcal{A}} (\Phi\tilde{\mu}^t\Psi_{a*}^\top)^\top (I-P_a)\Big]\right\|_\infty \|\Phi\breve{v} - v^*\|_\infty \\
&\leq 2\|\Phi\breve{v} - v^*\|_\infty,
\end{aligned}$$

By Theorem 2, we have the relation that (C.1) $= \mathcal{O}\left(t_{mix}\left(c_\Phi + \sqrt{U\log(DU)}\right)\sqrt{\frac{D}{T}}\right)$. By equation (13), the first two terms of (C.1) is larger than $\frac{1}{\tau}(\bar{v}^* - \mathbf{E}[\bar{v}^{\hat{\pi}}])$. Combining the above results and the bounds on (i), (ii) and (iii), we draw the conclusion of Theorem 4. □

References

[1] Thomas M. Cover and Joy A. Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[2] Mengdi Wang. Primal-dual π learning: Sample complexity and sublinear run time for ergodic markov decision problems. *CoRR*, abs/1710.06100, 2017.