A Proof of Lemma 1

Proof. Let us write $\tilde{\mu}^{t+1} = \Pi_{t+1}(\tilde{\mu}^{t+1/2})$ 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 $\Delta^{t+1}_{i_t, u_t} = \Phi_{i_t, \cdot} - \Phi_{i_t, \cdot} - r_t - M$ and $\Delta^{t+1}_{i_t, u_t} = 0$ 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 $\hat{\epsilon} = \arg\min_{\epsilon \in \mathcal{Y}} \|\Phi \hat{v} - v^*\|_\infty$ and $\hat{\mu} = \arg\min_{\mu \in \mathcal{Y}} \|\Phi \hat{\mu} \Psi^\top - \mu^*\|_{1,1}$. We obtain that

$$
D_{KL}(\hat{\mu} \| \tilde{\mu}^{t+1/2}) - D_{KL}(\hat{\mu} \| \tilde{\mu}^t) = \sum_{i=1}^D \sum_{u=1}^U \tilde{\mu}^t_{i, u} \log \frac{\tilde{\mu}^t_{i, u}}{\mu^{t+1/2}_{i, u}} - \sum_{i=1}^D \sum_{u=1}^U \tilde{\mu}^t_{i, u} \log \frac{\tilde{\mu}^t_{i, u}}{\tilde{\mu}^t_{i, u}}
$$

$$
= \sum_{i=1}^D \sum_{u=1}^U \tilde{\mu}^t_{i, u} \log \frac{\tilde{\mu}^t_{i, u}}{\mu^{t+1/2}_{i, u}}
$$

$$
= \sum_{i=1}^D \sum_{u=1}^U \tilde{\mu}^t_{i, u} \log \frac{Z}{\exp(\beta \Delta^{t+1}_{i, u})}
$$

$$
= \log Z - \beta \sum_{i=1}^D \sum_{u=1}^U \tilde{\mu}^t_{i, u} \Delta^{t+1}_{i, u},
$$

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 $\mathcal{Y}$, we have $|\Phi \hat{v}| \leq 2t_{mix}$ for all state $s$. Combining with our choice of $M = 4t_{mix} + 1$, we have $\Delta^{t+1}_{i, u} \leq 0$ for all $i = 1, \ldots, D$ and $u = 1, \ldots, U$. Consequently, applying the inequalities $e^x \leq 1 + x + \frac{1}{2} x^2$ for all $x \leq 0$ and $\log(1 + x) \leq x$ for all $x > -1$, we have

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

$$
= \log \left(1 + \beta \sum_{i=1}^D \sum_{u=1}^U \tilde{\mu}^t_{i, u} \Delta^{t+1}_{i, u} + \frac{\beta^2}{2} \sum_{i=1}^D \sum_{u=1}^U \tilde{\mu}^t_{i, u} (\Delta^{t+1}_{i, u})^2\right)
$$

$$
\leq \beta \sum_{i=1}^D \sum_{u=1}^U \tilde{\mu}^t_{i, u} \Delta^{t+1}_{i, u} + \frac{\beta^2}{2} \sum_{i=1}^D \sum_{u=1}^U \tilde{\mu}^t_{i, u} (\Delta^{t+1}_{i, u})^2
$$

Combining the above results, we have

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

In order to prove Lemma 1, we now show that $E[\Delta^{t+1}_{i, u} \mid \mathcal{F}^t] = \sum_{a \in A} \Psi_{a, u} \Phi^\top_{a, i} ((P_a - I) \Phi \hat{v} + r_a - M \cdot 1_S)$ and that $\sum_{i=1}^D \sum_{u=1}^U \tilde{\mu}^t_{i, u} E[\Delta^{t+1}_{i, u} \mid \mathcal{F}^t] \leq 100D\mu v_{mix}$. We use $1_S$ 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 $\phi_{i_t}$, $a_t$ is sampled from $\psi_{u_t}$ and $s_t'$
is sampled from $P_{u_i}(s_t, \cdot)$. Hence, for all $(i, u)$, we have
\[
E[\Delta_{i,u}^{t+1} \mid \mathcal{F}_t] = \frac{\tilde{\mu}_{i,u}^t \sum_{a \in A} \sum_{s \in S} \sum_{s' \in S} \Psi_{a,u} \cdot \Phi_{s,i} \cdot P_a(s, s') \cdot \Phi_{s',i} \tilde{v}^t + r_a(s) - \Phi_{ss} \tilde{v}^t - M}{\tilde{\mu}_{i,u}^t}
\]
\[
= \sum_{a \in A} \sum_{s \in S} \sum_{s' \in S} \Psi_{a,u} \Phi_{s,i} \left( P_a(s, \cdot) \Phi \tilde{v}^t + r_a(s) - \Phi_{ss} \tilde{v}^t - M \right)
\]
\[
= \sum_{a \in A} \sum_{u} \sum_{s} \tilde{\mu}_{i,u}^t \left( P_a \Phi \tilde{v}^t + r_a - \Phi \tilde{v}^t - M \cdot 1_S \right).
\]

It remains to prove that $\sum_{i=1}^{D} \sum_{u=1}^{U} \tilde{\mu}_{i,u}^t E[(\Delta_{i,u}^{t+1})^2 \mid \mathcal{F}_t] \leq 100 DU t_{mix}^2$. Expanding the expectation, we have
\[
\sum_{i=1}^{D} \sum_{u=1}^{U} \tilde{\mu}_{i,u}^t E[(\Delta_{i,u}^{t+1})^2 \mid \mathcal{F}_t]
\]
\[
= \sum_{i=1}^{D} \sum_{u=1}^{U} \tilde{\mu}_{i,u}^t \sum_{a \in A} \sum_{s \in S} \sum_{s' \in S} \Psi_{a,u} \cdot \Phi_{s,i} \cdot P_a(s, s') \left( \frac{\Phi_{s',i} \tilde{v}^t + r_a(s) - \Phi_{ss} \tilde{v}^t - M}{\tilde{\mu}_{i,u}^t} \right)^2
\]
\[
= \sum_{i=1}^{D} \sum_{u=1}^{U} \sum_{a \in A} \sum_{s,s'} \Psi_{a,u} \cdot \Phi_{s,i} \cdot P_a(s, s') (\Phi_{s',i} \tilde{v}^t + r_a(s) - \Phi_{ss} \tilde{v}^t - M)^2
\]
\[
\leq \sum_{i=1}^{D} \sum_{u=1}^{U} \sum_{a \in A} \sum_{s,s'} \Psi_{a,u} \cdot \Phi_{s,i} \cdot P_a(s, s') (8 t_{mix} + 2)^2
\]
\[
= DU (8 t_{mix} + 2)^2 \leq 100 DU t_{mix}^2,
\]
where the first inequality uses the relation that $|\Phi_{s',i} \tilde{v}^t + r_a(s) - \Phi_{ss} \tilde{v}^t - M| \leq 8 t_{mix} + 2$, 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 $t_{mix} \geq 1$. Substituting the above bounds in equation (A.1), we obtain that
\[
E[D_{KL}(\tilde{\mu}^{t+1/2} \mid \mathcal{F}_t) - D_{KL}(\tilde{\mu}^{t+1/2})]
\]
\[
\leq \beta \sum_{a \in A} \sum_{i=1}^{D} (\tilde{\mu}_{i,a}^t - \tilde{\mu}_{i,a}) \Psi_{a,u} \Phi_{s,i} ((P_a - I) \Phi \tilde{v}^t + r_a - M \cdot 1_S) + \frac{\beta^2}{2} \cdot 100 DU t_{mix}^2
\]
\[
\leq \beta \sum_{a \in A} \Psi_{a,s} (\tilde{\mu}^t - \tilde{\mu})^T \Phi^T ((P_a - I) \Phi \tilde{v}^t + r_a) + 50 \beta^2 DU t_{mix}^2,
\]
where the last inequality is due to that
\[
\sum_{a \in A} \Psi_{a,s} (\tilde{\mu}^t - \tilde{\mu})^T \Phi^T 1_S = \sum_{a \in A} \Psi_{a,s} (\tilde{\mu}^t - \tilde{\mu})^T \Phi^T 1_S = 1.
\]
Recall that $\tilde{\mu}^{t+1} = \Pi_{U,KL}(\tilde{\mu}^{t+1/2}) = \arg\min_{\mu' \in U} D_{KL}(\mu' \mid |\tilde{\mu}^{t+1/2}|)$ and $U$ 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
\[
E[D_{KL}(\tilde{\mu}^{t+1}) \mid \mathcal{F}_t] \leq E[D_{KL}(\tilde{\mu}^{t+1/2}) \mid \mathcal{F}_t].
\]
Combining the above inequalities, we conclude that
\[
E[D_{KL}(\tilde{\mu}^{t+1}) \mid \mathcal{F}_t] - D_{KL}(\tilde{\mu}^{t+1/2}) \leq E[D_{KL}(\tilde{\mu}^{t+1/2}) \mid \mathcal{F}_t] - D_{KL}(\tilde{\mu}^{t+1/2})
\]
\[
\leq \beta \sum_{a \in A} \Psi_{a,s} (\tilde{\mu}^t - \tilde{\mu})^T \Phi^T ((P_a - I) \Phi \tilde{v}^t + r_a) + 50 \beta^2 DU t_{mix}^2,
\]
Finally, observe that
\[
D_{KL}(\tilde{\mu}||\mu) = \sum_{i=1}^{D} \sum_{u=1}^{U} \tilde{\mu}_{i,u} \log \frac{\tilde{\mu}_{i,u}}{1/(DU)} = \sum_{i=1}^{D} \sum_{u=1}^{U} \mu_{i,u} \log(DU) + \sum_{i=1}^{D} \sum_{u=1}^{U} \tilde{\mu}_{i,u} \log(\tilde{\mu}_{i,u}) \leq \log(DU),
\]
where the last inequality is due to that $\tilde{\mu}_{i,u} \leq 1$ and thus $\log(\tilde{\mu}_{i,u}) \leq 0$ 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 $\Delta^{t+1} \triangleq \Phi_t^{s_{s_t}} - \Phi_t^{s_{s_t}}$. According to the update of Algorithm 1, we have $\tilde{\psi}^{t+1} = \Pi_\nu(\tilde{\psi}^t - \alpha \Delta^{t+1})$. By using the nonexpansive property of $\Pi_\nu$, we obtain that
\[
E \left[ \|\tilde{\psi}^{t+1} - \tilde{\psi}^t\|^2 \right] = E \left[ \|\Pi_\nu(\tilde{\psi}^t - \alpha \Delta^{t+1}) - \tilde{\psi}^t\|^2 \right] \leq E \left[ \|\tilde{\psi}^t - \alpha \Delta^{t+1} - \tilde{\psi}^t\|^2 \right] \leq 2 \alpha E \left[ \|\Delta^{t+1}\|^2 \right].
\]
(\ref{eq1})
Recall that $(i_t, u_t)$ is sampled from $\tilde{\mu}^t$, $a_t$ is sampled from $\tilde{\psi}_{u_t}$, $s_t$ is sampled from $\phi_{i_t}$ and $s'_t$ is sampled from $P_{a_t}(s_{t, \cdot})$. We can expand the expectation of $E[\|\Delta^{t+1}\|^2]$ to obtain that
\[
E[\|\Delta^{t+1}\|^2] = \sum_{a \in A} \sum_{u \in U} \sum_{s \in S} \sum_{s' \in S} \Psi_{a,u} \tilde{\mu}_{i,u}^t \Phi_{s,i} P_{a}(s, s') (\Phi_{s', s} - \Phi_{s,s}) = \sum_{a \in A} \sum_{u \in U} \Psi_{a,u} \tilde{\mu}_{i,u}^t \Phi_{s,i} P_{a}(s, s') \Phi_{s', s} = \sum_{a \in A} \Psi_{a} \tilde{\mu}_{i,u}^t \Phi_{s,i} P_{a}(s, s') \Phi_{s', s}.
\]
Next we prove that $E[\|\Delta^{t+1}\|^2 | \mathcal{F}_t] \leq \|\Phi\|^2_{2,\infty}$. A straightforward calculation yields that
\[
E[\|\Delta^{t+1}\|^2 | \mathcal{F}_t] = \sum_{a \in A} \sum_{u \in U} \sum_{s \in S} \sum_{s' \in S} \Psi_{a,u} \tilde{\mu}_{i,u}^t \Phi_{s,i} P_{a}(s, s') \|\Phi_{s', s} - \Phi_{s,s}\|^2 \leq \sum_{a \in A} \sum_{u \in U} \sum_{s \in S} \sum_{s' \in S} \Psi_{a,u} \tilde{\mu}_{i,u}^t \Phi_{s,i} P_{a}(s, s') (2 \|\Phi_{s', s}\|^2_2 + 2 \|\Phi_{s,s}\|^2_2) \leq \sum_{a \in A} \sum_{u \in U} \sum_{s \in S} \sum_{s' \in S} \Psi_{a,u} \tilde{\mu}_{i,u}^t \Phi_{s,i} P_{a}(s, s') (4 \|\Phi\|^2_{2,\infty}) \leq \|\Phi\|^2_{2,\infty},
\]
where the last equality is due to that $\tilde{\mu}_{i,u}$, $\psi_{u}$ and $\phi_{i}$ are distributions and $\sum_{a_i, u_a, s, s'} \Psi_{a_i, u_a} \tilde{\mu}_{i,u}^t \Phi_{s,i} P_{a}(s, s') = 1$. Substituting the above bounds into equation (\ref{eq1}), we get the first part of Lemma 2.

It remains to show that $\|\tilde{\psi}^{t+1} - \tilde{\psi}^t\|^2 = \|\tilde{\psi}^t\|^2 \leq \frac{4D_{\text{mix}}^2 \|\Phi\|^2_{\infty}}{\lambda_{\min}(\Phi^T \Phi)}$. Let $\nu' = \Phi \nu$. Multiply $\nu'$ by $\Phi^T$ and we get $\Phi^T \nu' = \Phi \Phi \nu$. Hence, by Assumption 1 that $\Phi^T \Phi$ is invertible, we have
\[
\tilde{\psi} = (\Phi^T \Phi)^{-1} \Phi^T \nu'.
\]
By our definition of $\nu$ and $\mathcal{V}$, we have $\|\nu\|_{\infty} \leq 2t_{\text{mix}}$. Using the relation that $\lambda_{\max}((\Phi^T \Phi)^{-1}) = \frac{1}{\lambda_{\min}(\Phi^T \Phi)}$ where $\lambda_{\max}$ and $\lambda_{\min}$ denotes the largest and the smallest eigenvalue, we obtain
\[
\|\tilde{\psi}\|^2 \leq \|\Phi^T \Phi^{-1}\|^2_2 \|\Phi^T \nu'\|^2_2 \leq \frac{1}{\lambda_{\min}(\Phi^T \Phi)} \cdot 4t_{\text{mix}}^2 \cdot \|\Phi\|^2_{1,2} \leq \frac{4t_{\text{mix}}^2 D \|\Phi\|^2_1}{\lambda_{\min}(\Phi^T \Phi)}.
\]

3
As a result, we have \( \| \tilde{v} \|_2^2 \leq \frac{4t_{\text{mix}}^2 D^2 \| \Phi \|_F^2}{\lambda_{\min}^2(\Phi^\top \Phi)}. \) Recall that every column of \( \Phi \) is a distribution and thus \( \| \Phi \|_1 = 1. \) Using this relationship, we obtain that \( \| \tilde{v} \|_2^2 \leq \frac{4t_{\text{mix}}^2 D^2}{\lambda_{\min}^2(\Phi^\top \Phi)}. \)

C Proof of Theorem 4

Proof. All the norms used in the proof of Theorem 4 are matrix norms. For a matrix \( \Phi \) 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

\[
\sum_{a \in A} r_a \mu_a^* + \frac{1}{T} \sum_{t=1}^T \mathbb{E} \left[ \sum_{a \in A} (I - P_a) v^* - r_a \right] \Phi \tilde{\mu}^T \Psi_a^T
- \frac{1}{T} \sum_{t=1}^T \mathbb{E} \left[ \sum_{a \in A} (\Phi \tilde{\mu}^T \Psi_a^T)(I - P_a) \Phi \tilde{v}^* \right] + \sum_{a \in A} (\Phi \tilde{\mu}^T \Psi_a^T - \mu_a^*)^T \tilde{r}_a
+ \frac{1}{T} \sum_{t=1}^T \mathbb{E} \left[ \sum_{a \in A} (\Phi \tilde{\mu}^T \Psi_a^T)(I - P_a)(\Phi \tilde{v} - v^*) \right],
\]

(C.1)

where \( \mu_a^* \) is the \( a \)-th column of \( \mu^* \). Next, we bound (i), (ii), (iii) respectively.

Analysis of (i): Recall that the stationary distribution \( \mu^* \) satisfies the condition \( \sum_{a \in A} (\mu_a^*)^T (I - P_a) = 0_S \). So we can bound (i) by

\[
| (i) | \leq \left\| \sum_{a \in A} (\Phi \tilde{\mu}^T \Psi_a^T - \mu_a^*)^T (I - P_a) \right\|_\infty \leq \frac{1}{T} \sum_{t=1}^T \mathbb{E} \| \Phi \tilde{v}^* \|_\infty
\]

\[
\leq \sum_{a \in A} \left\| (\Phi \tilde{\mu}^T \Psi_a^T - \mu_a^*)^T \right\|_\infty (\| I \|_\infty + \| P_a \|_\infty) \cdot 2t_{\text{mix}}
\]

\[
\leq 4t_{\text{mix}} \| \Phi \tilde{\mu}^T - \mu^* \|_{1,1},
\]

where the first inequality is due to that \( \| \Phi_1 \Phi_2 \|_\infty \leq \| \Phi_1 \|_\infty \| \Phi_2 \|_\infty \) for two matrices \( \Phi_1 \) and \( \Phi_2 \), the second inequality is due to that \( \| \Phi \tilde{v}^* \|_\infty \leq 2t_{\text{mix}} \) for all \( t \) (see Lemma 1 in [2]). In the third inequality, we use the fact that the matrix \( \infty \)-norm of a row vector is the sum of its components. And thus we have

\[
\sum_{a \in A} \left\| (\Phi \tilde{\mu}^T \Psi_a^T - \mu_a^*)^T \right\|_\infty = \| (\Phi \tilde{\mu}^T - \mu^*)^T \|_{1,1}.
\]

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

\[
| (ii) | \leq \sum_{a \in A} \left\| (\Phi \tilde{\mu}^T \Psi_a^T - \mu_a^*)^T \right\|_\infty \| \tilde{r}_a \|_\infty \leq \| \Phi \tilde{\mu}^T - \mu^* \|_{1,1},
\]

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 \( t \), \( \sum_{a \in A} (\Phi \tilde{\mu}^T \Psi_a^T)^T I \) and \( \sum_{a \in A} (\Phi \tilde{\mu}^T \Psi_a^T)^T P_a \) are two row vectors that both sum to 1. Recall that the matrix \( \infty \)-norm of a row vector is the sum of its components. Thus, we have \( \| \sum_{a \in A} (\Phi \tilde{\mu}^T \Psi_a^T)^T I - \sum_{a \in A} (\Phi \tilde{\mu}^T \Psi_a^T)^T P_a \|_\infty \leq 2 \). As a result, we have

\[
| (iii) | \leq \left\| \frac{1}{T} \sum_{t=1}^T \mathbb{E} \left[ \sum_{a \in A} (\Phi \tilde{\mu}^T \Psi_a^T)(I - P_a) \right] \right\|_\infty \| \Phi \tilde{v} - v^* \|_\infty
\]

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

References
