A. Omitted Details for the Algorithm

In this section, we provide omitted details on how to implement our algorithm efficiently.

A.1. Updating Occupancy Measure

This subsection explains how to implement the update defined in Eq. (7) efficiently. We use almost the same approach as in (Rosenberg & Mansour, 2019a) with the only difference being the choice of confidence set. We provide details of the modification here for completeness. It has been shown in (Rosenberg & Mansour, 2019a) that Eq. (7) can be decomposed into two steps: (1) compute $\tilde{q}_{t+1}(x,a,x') = \widehat{q}_t(x,a,x')\exp\{-\eta\widehat{\ell}_t(x,a)\}$ for any (x,a,x'), which is the optimal solution of the unconstrained problem; (2) compute the projection step:

$$\widehat{q}_{t+1} = \operatorname*{argmin}_{q\in\Delta(\mathcal{P}_i)} \; D(q \parallel \tilde{q}_{t+1}), \tag{11}$$

Since our choice of confidence set $\Delta(\mathcal{P}_i)$ is different, the main change lies in the second step, whose constraint set can be written explicitly using the following set of linear equations:

$$\begin{aligned}
&\forall k: && \sum_{x\in X_k, a\in A, x'\in X_{k+1}} q(x,a,x') = 1,\\
&\forall k,\ \forall x\in X_k: && \sum_{a\in A, x'\in X_{k+1}} q(x,a,x') = \sum_{x'\in X_{k-1}, a\in A} q(x',a,x),\\
&\forall k,\ \forall (x,a,x')\in X_k\times A\times X_{k+1}: && q(x,a,x') \le \left[\bar{P}_i(x'|x,a) + \epsilon_i(x'|x,a)\right]\sum_{y\in X_{k+1}} q(x,a,y),\\
& && q(x,a,x') \ge \left[\bar{P}_i(x'|x,a) - \epsilon_i(x'|x,a)\right]\sum_{y\in X_{k+1}} q(x,a,y),\\
& && q(x,a,x') \ge 0.
\end{aligned} \tag{12}$$

Therefore, the projection step Eq. (11) is a convex optimization problem with linear constraints, which can be solved in polynomial time. This optimization problem can be further reformulated into a dual problem, which is a convex optimization problem with only non-negativity constraints, and thus can be solved more efficiently.

Lemma 7. *The dual problem of Eq.(11) is to solve*

$$\mu_t, \beta_t = \operatorname*{argmin}_{\mu,\beta\ge 0} \sum_{k=0}^{L-1} \ln Z_t^k(\mu,\beta)$$

where $\beta := \{\beta(x)\}_x$ *and* $\mu := \{\mu^+(x,a,x'), \mu^-(x,a,x')\}_{(x,a,x')}$ *are dual variables and*

$$\begin{aligned}
Z_t^k(\mu,\beta) &= \sum_{x\in X_k, a\in A, x'\in X_{k+1}} \widehat{q}_t(x,a,x')\exp\left\{B_t^{\mu,\beta}(x,a,x')\right\},\\
B_t^{\mu,\beta}(x,a,x') &= \beta(x') - \beta(x) + \left(\mu^- - \mu^+\right)(x,a,x') - \eta\widehat{\ell}_t(x,a)\\
&\quad + \sum_{y\in X_{k(x)+1}} \left(\mu^+ - \mu^-\right)(x,a,y)\,\bar{P}_i(y|x,a) + \left(\mu^+ + \mu^-\right)(x,a,y)\,\epsilon_i(y|x,a).
\end{aligned}$$

Furthermore, the optimal solution to Eq.(11) is given by

$$\widehat{q}_{t+1}(x,a,x') = \frac{\widehat{q}_t(x,a,x')}{Z_t^{k(x)}(\mu_t,\beta_t)}\exp\left\{B_t^{\mu_t,\beta_t}(x,a,x')\right\}.$$

Proof. In the following proof, we omit the non-negativity constraint Eq. (12). This is without loss of generality, since the optimal solution for the modified version of Eq.(11) without the non-negativity constraint Eq. (12) turns out to always satisfy the non-negativity constraint.

We write the Lagrangian as:

$$\begin{aligned}\mathcal{L}\left(q,\lambda,\beta,\mu\right)=&D\left(q\|\tilde{q}_{t+1}\right)+\sum_{k=0}^{L-1}\lambda_k\left(\sum_{x\in X_k,a\in A,x'\in X_{k+1}}q\left(x,a,x'\right)-1\right)\\&+\sum_{k=1}^{L-1}\sum_{x\in X_k}\beta\left(x\right)\left(\sum_{a\in A,x'\in X_{k+1}}q\left(x,a,x'\right)-\sum_{x'\in X_{k-1},a\in A}q\left(x',a,x\right)\right)\\&+\sum_{k=0}^{L-1}\sum_{x\in X_k,a\in A,x'\in X_{k+1}}\mu^{+}\left(x,a,x'\right)\left(q\left(x,a,x'\right)-\left[\bar{P}_i\left(x'|x,a\right)+\epsilon_i\left(x'|x,a\right)\right]\sum_{y\in X_{k+1}}q\left(x,a,y\right)\right)\\&+\sum_{k=0}^{L-1}\sum_{x\in X_k,a\in A,x'\in X_{k+1}}\mu^{-}\left(x,a,x'\right)\left(\left[\bar{P}_i\left(x'|x,a\right)-\epsilon_i\left(x'|x,a\right)\right]\sum_{y\in X_{k+1}}q\left(x,a,y\right)-q\left(x,a,x'\right)\right)\end{aligned}$$

where $\lambda:=\{\lambda_k\}_k$, $\beta:=\{\beta(x)\}_x$ and $\mu:=\{\mu^+(x,a,x'),\mu^-(x,a,x')\}_{(x,a,x')}$ are Lagrange multipliers. We also define $\beta\left(x_0\right)=\beta\left(x_L\right)=0$ for convenience. Now taking the derivative we have

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial q\left(x,a,x'\right)}=&\ln q\left(x,a,x'\right)-\ln\tilde{q}_{t+1}\left(x,a,x'\right)+\lambda_{k(x)}+\beta\left(x\right)-\beta\left(x'\right)+\left(\mu^+-\mu^-\right)\left(x,a,x'\right)\\&-\sum_{y\in X_{k(x)+1}}\left(\mu^+-\mu^-\right)\left(x,a,y\right)\bar{P}_i\left(y|x,a\right)+\left(\mu^++\mu^-\right)\left(x,a,y\right)\epsilon_i\left(y|x,a\right)\\=&\ln q\left(x,a,x'\right)-\ln\tilde{q}_{t+1}\left(x,a,x'\right)+\lambda_{k(x)}-\eta\widehat{\ell}_t\left(x,a\right)-B_t^{\mu,\beta}\left(x,a,x'\right).\end{aligned}$$

Setting the derivative to zero gives the explicit form of the optimal $q^\star$ by

$$\begin{aligned}q^\star\left(x,a,x'\right)&=\tilde{q}_{t+1}\left(x,a,x'\right)\exp\left\{-\lambda_{k(x)}+\eta\widehat{\ell}_t\left(x,a\right)+B_t^{\mu,\beta}\left(x,a,x'\right)\right\}\\&=\widehat{q}_t\left(x,a,x'\right)\exp\left\{-\lambda_{k(x)}+B_t^{\mu,\beta}\left(x,a,x'\right)\right\}.\end{aligned}$$

On the other hand, setting $\partial\mathcal{L}/\partial\lambda_k=0$ shows that the optimal $\lambda^\star$ satisfies

$$\exp\left\{\lambda_k^\star\right\}=\sum_{x\in X_k,a\in A,x'\in X_{k+1}}\widehat{q}_t\left(x,a,x'\right)\exp\left\{B_t^{\mu,\beta}\left(x,a,x'\right)\right\}=Z_t^k\left(\mu,\beta\right).$$

It is straightforward to check that strong duality holds, and thus the optimal dual variables $\mu^\star,\beta^\star$ are given by

$$\mu^\star,\beta^\star=\operatorname*{argmax}_{\mu,\beta\geq0}\max_{\lambda}\min_{q}\mathcal{L}\left(q,\lambda,\beta,\mu\right)=\operatorname*{argmax}_{\mu,\beta\geq0}\mathcal{L}\left(q^\star,\lambda^\star,\beta,\mu\right).$$

Finally, we note the equality

$$\begin{aligned}\mathcal{L}\left(q,\lambda,\beta,\mu\right)=&D\left(q\|\tilde{q}_{t+1}\right)+\sum_{k=0}^{L-1}\sum_{x\in X_k,a\in A,x'\in X_{k+1}}\left(\frac{\partial\mathcal{L}}{\partial q\left(x,a,x'\right)}-\ln q\left(x,a,x'\right)+\ln\tilde{q}_{t+1}\left(x,a,x'\right)\right)q(x,a,x')-\sum_{k=1}^{L-1}\lambda_k\\=&\sum_{k=0}^{L-1}\sum_{x\in X_k,a\in A,x'\in X_{k+1}}\left[\left(\frac{\partial\mathcal{L}}{\partial q\left(x,a,x'\right)}-1\right)q(x,a,x')+\tilde{q}_{t+1}(x,a,x')\right]-\sum_{k=1}^{L-1}\lambda_k.\end{aligned}$$

This, combined with the fact that $q^\star$ has zero partial derivative, gives

$$\mathcal{L}\left(q^\star,\lambda^\star,\beta,\mu\right)=-L+\sum_{k=0}^{L-1}\sum_{x\in X_k,a\in A,x'\in X_{k+1}}\tilde{q}_{t+1}(x,a,x')-\sum_{k=0}^{L-1}\ln Z_t^k\left(\mu,\beta\right).$$

Note that the first two terms in the last expression are independent of (μ, β). We thus have:

$$\mu^\star, \beta^\star = \operatorname*{argmax}_{\mu,\beta \geq 0} \mathcal{L}\left(q^\star, \lambda^\star, \beta, \mu\right) = \operatorname*{argmin}_{\mu,\beta\geq 0} \sum_{k=0}^{L-1} \ln Z_t^k\left(\mu, \beta\right).$$

Combining all equations for $(q^\star, \lambda^\star, \mu^\star, \beta^\star)$ finishes the proof. □

A.2. Computing Upper Occupancy Bounds

This subsection explains how to greedily solve the following optimization problem from Eq. (10):

$$\max_{\widehat{P}(\cdot|\tilde{x},a)} \sum_{x' \in X_{k(\tilde{x})+1}} \widehat{P}(x'|\tilde{x},a) f(x')$$

subject to $\widehat{P}(\cdot|\tilde{x},a)$ being a valid distribution over $X_{k(\tilde{x})+1}$ and for all $x' \in X_{k(\tilde{x})+1}$,

$$\left|\widehat{P}(x'|\tilde{x},a) - \bar{P}_i(x'|\tilde{x},a)\right| \leq \epsilon_i(x'|\tilde{x},a),$$

where $(\tilde{x}, a)$ is some fixed state-action pair, $\epsilon_i(x'|\tilde{x},a)$ is defined in Eq. (6), and the value of $f(x')$ for any $x' \in X_{k(\tilde{x})+1}$ is known. To simplify notation, let $n = |X_{k(\tilde{x})+1}|$, and $\sigma : [n] \to X_{k(\tilde{x})+1}$ be a bijection such that

$$f(\sigma(1)) \leq f(\sigma(2)) \leq \cdots \leq f(\sigma(n)).$$

Further let $\bar{p}$ and ϵ be shorthands of $\bar{P}_i(\cdot|\tilde{x},a)$ and $\epsilon_i(\cdot|\tilde{x},a)$ respectively. With these notations, the problem becomes

$$\max_{\substack{p \in \mathbb{R}_+^n : \sum_{x'} p(x') = 1 \\ |p(x') - \bar{p}(x')| \leq \epsilon(x')}} \sum_{j=1}^{n} p(\sigma(j)) f(\sigma(j)).$$

Clearly, the maximum is achieved by redistributing the distribution $\bar{p}$ so that it puts as much weight as possible on states with large f value under the constraint. This can be implemented efficiently by maintaining two pointers j^- and j^+ starting from 1 and n respectively, and considering moving as much weight as possible from state $x^- = \sigma(j^-)$ to state $x^+ = \sigma(j^+)$. More specifically, the maximum possible weight change for x^- and x^+ are $\delta^- = \min\{\bar{p}(x^-), \epsilon(x^-)\}$ and $\delta^+ = \min\{1 - \bar{p}(x^+), \epsilon(x^+)\}$ respectively, and thus we move $\min\{\delta^-, \delta^+\}$ amount of weight from x^- to x^+. In the case where $\delta^- \leq \delta^+$, no more weight can be decreased from x^- and we increase the pointer j^- by 1 as well as decreasing $\epsilon(x^+)$ by δ^- to reflect the change in maximum possible weight increase for x^+. The situation for the case $\delta^- > \delta^+$ is similar. The procedure stops when the two pointers coincide. See Algorithm 4 for the complete pseudocode.

We point out that the step of sorting the values of f and finding σ can in fact be done only once for each layer (instead of every call of Algorithm 4). For simplicity, we omit this refinement.

B. Omitted Details for the Analysis

In this section, we provide omitted proofs for the regret analysis of our algorithm.

B.1. Auxiliary Lemmas

First, we prove Lemma 2 which states that with probability at least $1 - 4\delta$, the true transition function P is within the confidence set $\mathcal{P}_i$ for all epoch i.

Proof of Lemma 2. By the empirical Bernstein inequality (Maurer & Pontil, 2009, Theorem 4) and union bounds, we have with probability at least $1 - 4\delta$, for all $(x, a, x') \in X_k \times A \times X_{k+1}$, $k = 0, \ldots, L-1$, and any $i \leq T$,

$$\left|P(x'|x,a) - \bar{P}_i(x'|x,a)\right| \leq \sqrt{\frac{2\bar{P}_i(x'|x,a)(1 - \bar{P}_i(x'|x,a)) \ln\left(\frac{T|X|^2|A|}{\delta}\right)}{\max\{1, N_i(x,a) - 1\}}} + \frac{7 \ln\left(\frac{T|X|^2|A|}{\delta}\right)}{3\max\{1, N_i(x,a) - 1\}}$$

Algorithm 4 GREEDY

Input: $f : X \to [0,1]$, a distribution $\bar{p}$ over n states of layer k , positive numbers $\{\epsilon(x)\}_{x \in X_k}$
Initialize: $j^- = 1, j^+ = n$, sort $\{f(x)\}_{x \in X_k}$ and find σ such that $f(\sigma(1)) \leq f(\sigma(2)) \leq \cdots \leq f(\sigma(n))$

```
while j⁻ < j⁺ do
  x⁻ = σ(j⁻), x⁺ = σ(j⁺)
  δ⁻ = min{p̄(x⁻), ε(x⁻)}            ▷maximum weight to decrease for state x⁻
  δ⁺ = min{1 − p̄(x⁺), ε(x⁺)}        ▷maximum weight to increase for state x⁺
  p̄(x⁻) ← p̄(x⁻) − min{δ⁻, δ⁺}
  p̄(x⁺) ← p̄(x⁺) + min{δ⁻, δ⁺}
  if δ₋ ≤ δ₊ then
    ε(x⁺) ← ε(x⁺) − δ⁻
    j⁻ ← j⁻ + 1
  else
    ε(x⁻) ← ε(x⁻) − δ⁺
    j⁺ ← j⁺ − 1
  end if
end while
Return: Σ_{j=1}^n p̄(σ(j)) f(σ(j))
```

$$\leq 2\sqrt{\frac{\bar{P}_i(x'|x,a)\ln\left(\frac{T|X||A|}{\delta}\right)}{\max\{1, N_i(x,a)-1\}}} + \frac{14\ln\left(\frac{T|X||A|}{\delta}\right)}{3\max\{1, N_i(x,a)-1\}} = \epsilon_i(x'|x,a)$$

which finishes the proof. □

Next, we state three lemmas that are useful for the rest of the proof. The first one shows a convenient bound on the difference between the true transition function and any transition function from the confidence set.

Lemma 8. *Under the event of Lemma 2, for all epoch i, all $\widehat{P} \in \mathcal{P}_i$, all $k = 0, \ldots, L-1$ and $(x, a, x') \in X_k \times A \times X_{k+1}$, we have*

$$\left|\widehat{P}(x'|x,a) - P(x'|x,a)\right| = \mathcal{O}\left(\sqrt{\frac{P(x'|x,a)\ln\left(\frac{T|X||A|}{\delta}\right)}{\max\{1, N_i(x,a)\}}} + \frac{\ln\left(\frac{T|X||A|}{\delta}\right)}{\max\{1, N_i(x,a)\}}\right) \triangleq \epsilon_i^\star(x'|x,a).$$

Proof. Under the event of Lemma 2, we have

$$\bar{P}_i(x'|x,a) \leq P(x'|x,a) + 2\sqrt{\frac{\bar{P}_i(x'|x,a)\ln\left(\frac{T|X||A|}{\delta}\right)}{\max\{1, N_i(x,a)-1\}}} + \frac{14\ln\left(\frac{T|X||A|}{\delta}\right)}{3\max\{1, N_i(x,a)-1\}}.$$

Viewing this as a quadratic inequality of $\sqrt{\bar{P}_i(x'|x,a)}$ and solving for $\bar{P}_i(x'|x,a)$ prove the lemma. □

The next one is a standard Bernstein-type concentration inequality for martingale. We use the version from (Beygelzimer et al., 2011, Theorem 1).

Lemma 9. *Let $Y_1, \ldots, Y_T$ be a martingale difference sequence with respect to a filtration $\mathcal{F}_1, \ldots, \mathcal{F}_T$. Assume $Y_t \leq R$ a.s. for all i. Then for any $\delta \in (0,1)$ and $\lambda \in [0, 1/R]$, with probability at least $1-\delta$, we have*

$$\sum_{t=1}^T Y_t \leq \lambda \sum_{t=1}^T \mathbb{E}_t[Y_t^2] + \frac{\ln(1/\delta)}{\lambda}.$$

The last one is a based on similar ideas used for proving many other optimistic algorithms.

Lemma 10. *With probability at least* $1-2\delta$*, we have for all* $k=0,\ldots,L-1$*,*

$$\sum_{t=1}^{T}\sum_{x\in X_k,a\in A}\frac{q_t(x,a)}{\max\{1,N_{i_t}(x,a)\}}=\mathcal{O}\left(|X_k||A|\ln T+\ln(L/\delta)\right) \tag{13}$$

and

$$\sum_{t=1}^{T}\sum_{x\in X_k,a\in A}\frac{q_t(x,a)}{\sqrt{\max\{1,N_{i_t}(x,a)\}}}=\mathcal{O}\left(\sqrt{|X_k||A|T}+|X_k||A|\ln T+\ln(L/\delta)\right). \tag{14}$$

Proof. Let $\mathbb{I}_t(x,a)$ be the indicator of whether the pair (x,a) is visited in episode t so that $\mathbb{E}_t[\mathbb{I}_t(x,a)]=q_t(x,a)$. We decompose the first quantity as

$$\sum_{t=1}^{T}\sum_{x\in X_k,a\in A}\frac{q_t(x,a)}{\max\{1,N_{i_t}(x,a)\}}=\sum_{t=1}^{T}\sum_{x\in X_k,a\in A}\frac{\mathbb{I}_t(x,a)}{\max\{1,N_{i_t}(x,a)\}}+\sum_{t=1}^{T}\sum_{x\in X_k,a\in A}\frac{q_t(x,a)-\mathbb{I}_t(x,a)}{\max\{1,N_{i_t}(x,a)\}}.$$

The first term can be bounded as

$$\sum_{x\in X_k,a\in A}\sum_{t=1}^{T}\frac{\mathbb{I}_t(x,a)}{\max\{1,N_{i_t}(x,a)\}}=\sum_{x\in X_k,a\in A}\mathcal{O}\left(\ln T\right)=\mathcal{O}\left(|X_k||A|\ln T\right).$$

To bound the second term, we apply Lemma 9 with $Y_t=\sum_{x\in X_k,a\in A}\frac{q_t(x,a)-\mathbb{I}_t(x,a)}{\max\{1,N_{i_t}(x,a)\}}\le 1$, $\lambda=1/2$, and the fact

$$\begin{aligned}\mathbb{E}_t[Y_t^2]&\le\mathbb{E}_t\left[\left(\sum_{x\in X_k,a\in A}\frac{\mathbb{I}_t(x,a)}{\max\{1,N_{i_t}(x,a)\}}\right)^2\right]\\&=\mathbb{E}_t\left[\sum_{x\in X_k,a\in A}\frac{\mathbb{I}_t(x,a)}{\max\{1,N_{i_t}^2(x,a)\}}\right]&&(\mathbb{I}_t(x,a)\mathbb{I}_t(x',a')=0\text{ for }x\neq x'\in X_k)\\&\le\sum_{x\in X_k,a\in A}\frac{q_t(x,a)}{\max\{1,N_{i_t}(x,a)\}},\end{aligned}$$

which gives with probability at least $1-\delta/L$,

$$\sum_{t=1}^{T}\sum_{x\in X_k,a\in A}\frac{q_t(x,a)-\mathbb{I}_t(x,a)}{\max\{1,N_{i_t}(x,a)\}}\le\frac{1}{2}\sum_{t=1}^{T}\sum_{x\in X_k,a\in A}\frac{q_t(x,a)}{\max\{1,N_{i_t}(x,a)\}}+2\ln\left(\frac{L}{\delta}\right).$$

Combining these two bounds, rearranging, and applying a union bound over k prove Eq. (13).

Similarly, we decompose the second quantity as

$$\sum_{t=1}^{T}\sum_{x\in X_k,a\in A}\frac{q_t(x,a)}{\sqrt{\max\{1,N_{i_t}(x,a)\}}}=\sum_{t=1}^{T}\sum_{x\in X_k,a\in A}\frac{\mathbb{I}_t(x,a)}{\sqrt{\max\{1,N_{i_t}(x,a)\}}}+\sum_{t=1}^{T}\sum_{x\in X_k,a\in A}\frac{q_t(x,a)-\mathbb{I}_t(x,a)}{\sqrt{\max\{1,N_{i_t}(x,a)\}}}.$$

The first term is bounded by

$$\begin{aligned}\sum_{x\in X_k,a\in A}\sum_{t=1}^{T}\frac{\mathbb{I}_t(x,a)}{\sqrt{\max\{1,N_{i_t}(x,a)\}}}&=\mathcal{O}\left(\sum_{x\in X_k,a\in A}\sqrt{N_{i_T}(x,a)}\right)\\&\le\mathcal{O}\left(\sqrt{|X_k||A|\sum_{x\in X_k,a\in A}N_{i_T}(x,a)}\right)=\mathcal{O}\left(\sqrt{|X_k||A|T}\right),\end{aligned}$$

where the second line uses the Cauchy-Schwarz inequality and the fact $\sum_{x\in X_k, a\in A} N_{i_T}(x,a) \le T$. To bound the second term, we again apply Lemma 9 with $Y_t = \sum_{x\in X_k, a\in A} \frac{q_t(x,a)-\mathbb{I}_t(x,a)}{\sqrt{\max\{1,N_{i_t}(x,a)\}}} \le 1$, $\lambda = 1$, and the fact

$$\mathbb{E}_t[Y_t^2] \le \mathbb{E}_t\left[\left(\sum_{x\in X_k, a\in A} \frac{\mathbb{I}_t(x,a)}{\sqrt{\max\{1,N_{i_t}(x,a)\}}}\right)^2\right] = \sum_{x\in X_k, a\in A} \frac{q_t(x,a)}{\max\{1,N_{i_t}(x,a)\}},$$

which shows with probability at least $1-\delta/L$,

$$\sum_{t=1}^{T}\sum_{x\in X_k, a\in A} \frac{q_t(x,a)-\mathbb{I}_t(x,a)}{\sqrt{\max\{1,N_{i_t}(x,a)\}}} \le \sum_{t=1}^{T}\sum_{x\in X_k, a\in A} \frac{q_t(x,a)}{\max\{1,N_{i_t}(x,a)\}} + \ln\left(\frac{L}{\delta}\right).$$

Combining Eq. (13) and a union bound proves Eq. (14). □

B.2. Proof of the Key Lemma

We are now ready to prove Lemma 4, the key lemma of our analysis which requires using our new confidence set.

Proof of Lemma 4. To simplify notation, let $q_t^x = q^{P_t^x, \pi_t}$. Note that for any occupancy measure q, by definition we have for any (x,a) pair,

$$q(x,a) = \pi^q(x|a) \sum_{\{x_k\in X_k, a_k\in A\}_{k=0}^{k(x)-1}} \prod_{h=0}^{k(x)-1} \pi^q(a_h|x_h) \prod_{h=0}^{k(x)-1} P^q(x_{h+1}|x_h,a_h).$$

where we define $x_{k(x)} = x$ for convenience. Therefore, we have

$$|q_t^x(x,a) - q_t(x,a)| = \pi_t(x|a) \sum_{\{x_k,a_k\}_{k=0}^{k(x)-1}} \prod_{h=0}^{k(x)-1} \pi_t(a_h|x_h) \left(\prod_{h=0}^{k(x)-1} P_t^x(x_{h+1}|x_h,a_h) - \prod_{h=0}^{k(x)-1} P(x_{h+1}|x_h,a_h)\right).$$

By adding and subtracting $k(x)-1$ terms we rewrite the last term in the parentheses as

$$\begin{aligned}
&\prod_{h=0}^{k(x)-1} P_t^x(x_{h+1}|x_h,a_h) - \prod_{h=0}^{k(x)-1} P(x_{h+1}|x_h,a_h) \\
&= \prod_{h=0}^{k(x)-1} P_t^x(x_{h+1}|x_h,a_h) - \prod_{h=0}^{k(x)-1} P(x_{h+1}|x_h,a_h) \pm \sum_{m=1}^{k(x)-1} \prod_{h=0}^{m-1} P(x_{h+1}|x_h,a_h) \prod_{h=m}^{k(x)-1} P_t^x(x_{h+1}|x_h,a_h) \\
&= \sum_{m=0}^{k(x)-1} \left(P_t^x(x_{m+1}|x_m,a_m) - P(x_{m+1}|x_m,a_m)\right) \prod_{h=0}^{m-1} P(x_{h+1}|x_h,a_h) \prod_{h=m+1}^{k(x)-1} P_t^x(x_{h+1}|x_h,a_h),
\end{aligned}$$

which, by Lemma 8, is bounded by

$$\sum_{m=0}^{k(x)-1} \epsilon_{i_t}^\star(x_{m+1}|x_m,a_m) \prod_{h=0}^{m-1} P(x_{h+1}|x_h,a_h) \prod_{h=m+1}^{k(x)-1} P_t^x(x_{h+1}|x_h,a_h).$$

We have thus shown

$$\begin{aligned}
&|q_t^x(x,a) - q_t(x,a)| \\
&\le \pi_t(x|a) \sum_{\{x_k,a_k\}_{k=0}^{k(x)-1}} \prod_{h=0}^{k(x)-1} \pi_t(a_h|x_h) \sum_{m=0}^{k(x)-1} \epsilon_{i_t}^\star(x_{m+1}|x_m,a_m) \prod_{h=0}^{m-1} P(x_{h+1}|x_h,a_h) \prod_{h=m+1}^{k(x)-1} P_t^x(x_{h+1}|x_h,a_h)
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{m=0}^{k(x)-1} \sum_{\{x_k,a_k\}_{k=0}^{k(x)-1}} \epsilon_{i_t}^\star(x_{m+1}|x_m,a_m) \left(\pi_t(a_m|x_m) \prod_{h=0}^{m-1} \pi_t(a_h|x_h) P(x_{h+1}|x_h,a_h)\right) \\
&\qquad \cdot \left(\pi_t(x|a) \prod_{h=m+1}^{k(x)-1} \pi_t(a_h|x_h) P_t^x(x_{h+1}|x_h,a_h)\right) \\
&= \sum_{m=0}^{k(x)-1} \sum_{x_m,a_m,x_{m+1}} \epsilon_{i_t}^\star(x_{m+1}|x_m,a_m) \left(\sum_{\{x_k,a_k\}_{k=0}^{m-1}} \pi_t(a_m|x_m) \prod_{h=0}^{m-1} \pi_t(a_h|x_h) P(x_{h+1}|x_h,a_h)\right) \\
&\qquad \cdot \left(\sum_{a_{m+1}} \sum_{\{x_k,a_k\}_{k=m+2}^{k(x)-1}} \pi_t(x|a) \prod_{h=m+1}^{k(x)-1} \pi_t(a_h|x_h) P_t^x(x_{h+1}|x_h,a_h)\right) \\
&= \sum_{m=0}^{k(x)-1} \sum_{x_m,a_m,x_{m+1}} \epsilon_{i_t}^\star(x_{m+1}|x_m,a_m) q_t(x_m,a_m) q_t^x(x,a|x_{m+1}), \qquad (15)
\end{aligned}
$$

where we use $q_t^x(x,a|x_{m+1})$ to denote the probability of encountering pair (x,a) given that x_{m+1} was visited in layer $m+1$, under policy π_t and transition P_t^x. By the exact same reasoning, we also have

$$
\begin{aligned}
|q_t^x(x,a|x_{m+1}) - q_t(x,a|x_{m+1})| &\le \sum_{h=m+1}^{k(x)-1} \sum_{x'_h,a'_h,x'_{h+1}} \epsilon_{i_t}^\star(x'_{h+1}|x'_h,a'_h) q_t(x'_h,a'_h|x_{m+1}) q_t^x(x,a|x'_{h+1}) \\
&\le \pi_t(a|x) \sum_{h=m+1}^{k(x)-1} \sum_{x'_h,a'_h,x'_{h+1}} \epsilon_{i_t}^\star(x'_{h+1}|x'_h,a'_h) q_t(x'_h,a'_h|x_{m+1}) \qquad (16)
\end{aligned}
$$

Combining Eq. (15) and Eq. (16), summing over all t and (x,a), and using the shorthands $w_m = (x_m,a_m,x_{m+1})$ and $w'_h = (x'_h,a'_h,x'_{h+1})$, we have derived

$$
\begin{aligned}
&\sum_{t=1}^{T} \sum_{x\in X, a\in A} |q_t^x(x,a) - q_t(x,a)| \\
&\le \sum_{t,x,a} \sum_{m=0}^{k(x)-1} \sum_{w_m} \epsilon_{i_t}^\star(x_{m+1}|x_m,a_m) q_t(x_m,a_m) q_t(x,a|x_{m+1}) \\
&\quad + \sum_{t,x,a} \sum_{m=0}^{k(x)-1} \sum_{w_m} \epsilon_{i_t}^\star(x_{m+1}|x_m,a_m) q_t(x_m,a_m) \left(\pi_t(a|x) \sum_{h=m+1}^{k(x)-1} \sum_{w'_h} \epsilon_{i_t}^\star(x'_{h+1}|x'_h,a'_h) q_t(x'_h,a'_h|x_{m+1})\right) \\
&= \sum_t \sum_{k<L} \sum_{m=0}^{k-1} \sum_{w_m} \epsilon_{i_t}^\star(x_{m+1}|x_m,a_m) q_t(x_m,a_m) \sum_{x\in X_k, a\in A} q_t(x,a|x_{m+1}) \\
&\quad + \sum_t \sum_{k<L} \sum_{m=0}^{k-1} \sum_{w_m} \sum_{h=m+1}^{k-1} \sum_{w'_h} \epsilon_{i_t}^\star(x_{m+1}|x_m,a_m) q_t(x_m,a_m) \epsilon_{i_t}^\star(x'_{h+1}|x'_h,a'_h) q_t(x'_h,a'_h|x_{m+1}) \left(\sum_{x\in X_k, a\in A} \pi_t(a|x)\right) \\
&= \sum_{0\le m<k<L} \sum_{t,w_m} \epsilon_{i_t}^\star(x_{m+1}|x_m,a_m) q_t(x_m,a_m) \\
&\quad + \sum_{0\le m<h<k<L} |X_k| \sum_{t,w_m,w'_h} \epsilon_{i_t}^\star(x_{m+1}|x_m,a_m) q_t(x_m,a_m) \epsilon_{i_t}^\star(x'_{h+1}|x'_h,a'_h) q_t(x'_h,a'_h|x_{m+1}) \\
&\le \underbrace{\sum_{0\le m<k<L} \sum_{t,w_m} \epsilon_{i_t}^\star(x_{m+1}|x_m,a_m) q_t(x_m,a_m)}_{\triangleq B_1}
\end{aligned}
$$

$$+ |X| \underbrace{\sum_{0\le m<h<L} \sum_{t,w_m,w'_h} \epsilon^\star_{i_t}(x_{m+1}|x_m,a_m) q_t(x_m,a_m) \epsilon^\star_{i_t}(x'_{h+1}|x'_h,a'_h) q_t(x'_h,a'_h|x_{m+1})}_{\triangleq B_2}.$$

It remains to bound B_1 and B_2 using the definition of $\epsilon^\star_{i_t}$. For B_1, we have

$$\begin{aligned}
B_1 &= \mathcal{O}\left(\sum_{0\le m<k<L} \sum_{t,w_m} q_t(x_m,a_m)\sqrt{\frac{P(x_{m+1}|x_m,a_m)\ln\left(\frac{T|X||A|}{\delta}\right)}{\max\{1,N_{i_t}(x_m,a_m)\}}} + \frac{q_t(x_m,a_m)\ln\left(\frac{T|X||A|}{\delta}\right)}{\max\{1,N_{i_t}(x_m,a_m)\}}\right) \\
&\le \mathcal{O}\left(\sum_{0\le m<k<L} \sum_{t,x_m,a_m} q_t(x_m,a_m)\sqrt{\frac{|X_{m+1}|\ln\left(\frac{T|X||A|}{\delta}\right)}{\max\{1,N_{i_t}(x_m,a_m)\}}} + \frac{q_t(x_m,a_m)\ln\left(\frac{T|X||A|}{\delta}\right)}{\max\{1,N_{i_t}(x_m,a_m)\}}\right) \\
&\le \mathcal{O}\left(\sum_{0\le m<k<L} \sqrt{|X_m||X_{m+1}||A|T\ln\left(\frac{T|X||A|}{\delta}\right)}\right) \\
&\le \mathcal{O}\left(\sum_{0\le m<k<L} (|X_m|+|X_{m+1}|)\sqrt{|A|T\ln\left(\frac{T|X||A|}{\delta}\right)}\right) \\
&= \mathcal{O}\left(L|X|\sqrt{|A|T\ln\left(\frac{T|X||A|}{\delta}\right)}\right),
\end{aligned}$$

where the second line uses the Cauchy-Schwarz inequality, the third line uses Lemma 10, and the fourth line uses the AM-GM inequality.

For B_2, plugging the definition of $\epsilon^\star_{i_t}$ and using trivial bounds (that is, $\epsilon^\star_{i_t}$ and q_t are both at most 1 regardless of the arguments), we obtain the following three terms (ignoring constants)

$$\sum_{0\le m<h<L} \sum_{t,w_m,w'_h} \sqrt{\frac{P(x_{m+1}|x_m,a_m)\ln\left(\frac{T|X||A|}{\delta}\right)}{\max\{1,N_{i_t}(x_m,a_m)\}}} q_t(x_m,a_m)\sqrt{\frac{P(x'_{h+1}|x'_h,a'_h)\ln\left(\frac{T|X||A|}{\delta}\right)}{\max\{1,N_{i_t}(x'_h,a'_h)\}}} q_t(x'_h,a'_h|x_{m+1})$$
$$+ \sum_{0\le m<h<L} \sum_{t,w_m,w'_h} \frac{q_t(x_m,a_m)\ln\left(\frac{T|X||A|}{\delta}\right)}{\max\{1,N_{i_t}(x_m,a_m)\}} + \sum_{0\le m<h<L} \sum_{t,w_m,w'_h} \frac{q_t(x'_h,a'_h)\ln\left(\frac{T|X||A|}{\delta}\right)}{\max\{1,N_{i_t}(x'_h,a'_h)\}}.$$

The last two terms are both of order $\mathcal{O}(\ln T)$ by Lemma 10 (ignoring dependence on other parameters), while the first term can be written as $\ln\left(\frac{T|X||A|}{\delta}\right)$ multiplied by the following:

$$\begin{aligned}
&\sum_{0\le m<h<L} \sum_{t,w_m,w'_h} \sqrt{\frac{q_t(x_m,a_m)P(x'_{h+1}|x'_h,a'_h)q_t(x'_h,a'_h|x_{m+1})}{\max\{1,N_{i_t}(x_m,a_m)\}}}\sqrt{\frac{q_t(x_m,a_m)P(x_{m+1}|x_m,a_m)q_t(x'_h,a'_h|x_{m+1})}{\max\{1,N_{i_t}(x'_h,a'_h)\}}} \\
&\le \sum_{0\le m<h<L} \sqrt{\sum_{t,w_m,w'_h} \frac{q_t(x_m,a_m)P(x'_{h+1}|x'_h,a'_h)q_t(x'_h,a'_h|x_{m+1})}{\max\{1,N_{i_t}(x_m,a_m)\}}}\sqrt{\sum_{t,w_m,w'_h} \frac{q_t(x_m,a_m)P(x_{m+1}|x_m,a_m)q_t(x'_h,a'_h|x_{m+1})}{\max\{1,N_{i_t}(x'_h,a'_h)\}}} \\
&= \sum_{0\le m<h<L} \sqrt{|X_{m+1}|\sum_{t,x_m,a_m} \frac{q_t(x_m,a_m)}{\max\{1,N_{i_t}(x_m,a_m)\}}}\sqrt{|X_{h+1}|\sum_{t,x'_h,a'_h} \frac{q_t(x'_h,a'_h)}{\max\{1,N_{i_t}(x'_h,a'_h)\}}} \\
&= \mathcal{O}\left(|A|\ln\left(\frac{T|X||A|}{\delta}\right)\right)\sum_{0\le m<h<L}\sqrt{|X_m||X_{m+1}||X_h||X_{h+1}|} = \mathcal{O}\left(L^2|X|^2|A|\ln\left(\frac{T|X||A|}{\delta}\right)\right),
\end{aligned}$$

where the second line uses the Cauchy-Schwarz inequality and the last line uses Lemma 10 again. This shows that the entire term B_2 is of order $O(\ln T)$. Finally, realizing that we have conditioned on the events stated in Lemmas 8 and 10, which happen with probability at least $1-6\delta$, finishes the proof. □

B.3. Bounding REG and BIAS$_2$

In this section, we complete the proof of our main theorem by bounding the terms REG and BIAS$_2$. We first state the following useful concentration lemma which is a variant of (Neu, 2015, Lemma 1) and is the key for analyzing the implicit exploration effect introduced by γ. The proof is based on the same idea of the proof for (Neu, 2015, Lemma 1).

Lemma 11. *For any sequence of functions $\alpha_1, \ldots, \alpha_T$ such that $\alpha_t \in [0, 2\gamma]^{X \times A}$ is $\mathcal{F}_t$-measurable for all t, we have with probability at least $1 - \delta$,*

$$\sum_{t=1}^{T} \sum_{x,a} \alpha_t(x,a) \left(\widehat{\ell}_t(x,a) - \frac{q_t(x,a)}{u_t(x,a)} \ell_t(x,a) \right) \le L \ln \tfrac{L}{\delta}.$$

Proof. Fix any t. For simplicity, let $\beta = 2\gamma$ and $\mathbb{I}_{t,x,a}$ be a shorthand of $\mathbb{I}\{x_{k(x)} = x, a_{k(x)} = a\}$. Then for any state-action pair (x, a), we have

$$\widehat{\ell}_t(x,a) = \frac{\ell_t(x,a)\mathbb{I}_{t,x,a}}{u_t(x,a) + \gamma} \le \frac{\ell_t(x,a)\mathbb{I}_{t,x,a}}{u_t(x,a) + \gamma \ell_t(x,a)} = \frac{\mathbb{I}_{t,x,a}}{\beta} \cdot \frac{2\gamma \ell_t(x,a)/u_t(x,a)}{1 + \gamma \ell_t(x,a)/u_t(x,a)} \le \frac{1}{\beta} \ln \left(1 + \frac{\beta \ell_t(x,a)\mathbb{I}_{t,x,a}}{u_t(x,a)} \right), \tag{17}$$

where the last step uses the fact $\frac{z}{1+z/2} \le \ln(1+z)$ for all $z \ge 0$. For each layer $k < L$, further define

$$\widehat{S}_{t,k} = \sum_{x \in X_k, a \in A} \alpha_t(x,a) \widehat{\ell}_t(x,a) \qquad \text{and} \qquad S_{t,k} = \sum_{x \in X_k, a \in A} \alpha_t(x,a) \frac{q_t(x,a)}{u_t(x,a)} \ell_t(x,a).$$

The following calculation shows $\mathbb{E}_t \left[\exp(\widehat{S}_{t,k}) \right] \le \exp(S_{t,k})$:

$$\begin{aligned}
\mathbb{E}_t \left[\exp(\widehat{S}_{t,k}) \right] &\le \mathbb{E}_t \left[\exp \left(\sum_{x \in X_k, a \in A} \frac{\alpha_t(x,a)}{\beta} \ln \left(1 + \frac{\beta \ell_t(x,a)\mathbb{I}_{t,x,a}}{u_t(x,a)} \right) \right) \right] && \text{(by Eq. (17))} \\
&\le \mathbb{E}_t \left[\prod_{x \in X_k, a \in A} \left(1 + \frac{\alpha_t(x,a)\ell_t(x,a)\mathbb{I}_{t,x,a}}{u_t(x,a)} \right) \right] \\
&= \mathbb{E}_t \left[1 + \sum_{x \in X_k, a \in A} \frac{\alpha_t(x,a)\ell_t(x,a)\mathbb{I}_{t,x,a}}{u_t(x,a)} \right] \\
&= 1 + S_{t,k} \le \exp(S_{t,k}).
\end{aligned}$$

Here, the second inequality is due to the fact $z_1 \ln(1 + z_2) \le \ln(1 + z_1 z_2)$ for all $z_2 \ge -1$ and $z_1 \in [0, 1]$, and we apply it with $z_1 = \frac{\alpha_t(x,a)}{\beta}$ which is in $[0, 1]$ by the condition $\alpha_t(x, a) \in [0, 2\gamma]$; the first equality holds since $\mathbb{I}_{t,x,a}\mathbb{I}_{t,x',a'} = 0$ for any $x \ne x'$ or $a \ne a'$ (as only one state-action pair can be visited in each layer for an episode). Next we apply Markov inequality and show

$$\begin{aligned}
\Pr \left[\sum_{t=1}^{T} (\widehat{S}_{t,k} - S_{t,k}) > \ln \left(\frac{L}{\delta} \right) \right] &\le \frac{\delta}{L} \cdot \mathbb{E} \left[\exp \left(\sum_{t=1}^{T} (\widehat{S}_{t,k} - S_{t,k}) \right) \right] \\
&= \frac{\delta}{L} \cdot \mathbb{E} \left[\exp \left(\sum_{t=1}^{T-1} (\widehat{S}_{t,k} - S_{t,k}) \right) \mathbb{E}_T \left[\exp \left(\widehat{S}_{T,k} - S_{T,k} \right) \right] \right] \\
&\le \frac{\delta}{L} \cdot \mathbb{E} \left[\exp \left(\sum_{t=1}^{T-1} (\widehat{S}_{t,k} - S_{t,k}) \right) \right] \\
&\le \cdots \le \frac{\delta}{L}.
\end{aligned} \tag{18}$$

Finally, applying a union bound over $k = 0, \ldots, L - 1$ shows with probability at least $1 - \delta$,

$$\sum_{t=1}^{T} \sum_{x,a} \alpha_t(x,a) \left(\widehat{\ell}_t(x,a) - \frac{q_t(x,a)}{u_t(x,a)} \ell_t(x,a) \right) = \sum_{k=0}^{L-1} \sum_{t=1}^{T} (\widehat{S}_{t,k} - S_{t,k}) \le L \ln \left(\frac{L}{\delta} \right),$$

which completes the proof. □

Bounding REG. To bound $\text{REG} = \sum_{t=1}^{T} \langle \widehat{q}_t - q^*, \widehat{\ell}_t \rangle$, note that under the event of Lemma 2, $q^* \in \cap_i \Delta(\mathcal{P}_i)$, and thus REG is controlled by the standard regret guarantee of OMD. Specifically, we prove the following lemma.

Lemma 12. *With probability at least* $1 - 5\delta$, UOB-REPS *ensures* $\text{REG} = \mathcal{O}\left(\frac{L\ln(|X||A|)}{\eta} + \eta|X||A|T + \frac{\eta L \ln(L/\delta)}{\gamma}\right)$.

Proof. By standard analysis (see Lemma 13 after this proof), OMD with KL-divergence ensures for any $q \in \cap_i \Delta(\mathcal{P}_i)$,

$$\sum_{t=1}^{T} \langle \widehat{q}_t - q, \widehat{\ell}_t \rangle \leq \frac{L\ln(|X|^2|A|)}{\eta} + \eta \sum_{t,x,a} \widehat{q}_t(x,a)\widehat{\ell}_t(x,a)^2.$$

Further note that $\widehat{q}_t(x,a)\widehat{\ell}_t(x,a)^2$ is bounded by

$$\frac{\widehat{q}_t(x,a)}{u_t(x,a)+\gamma}\widehat{\ell}_t(x,a) \leq \widehat{\ell}_t(x,a)$$

by the fact $\widehat{q}_t(x,a) \leq u_t(x,a)$. Applying Lemma 11 with $\alpha_t(x,a) = 2\gamma$ then shows with probability at least $1-\delta$,

$$\sum_{t,x,a} \widehat{q}_t(x,a)\widehat{\ell}_t(x,a)^2 \leq \sum_{t,x,a} \frac{q_t(x,a)}{u_t(x,a)}\ell_t(x,a) + \frac{L\ln\frac{L}{\delta}}{2\gamma}.$$

Finally, note that under the event of Lemma 2, we have $q^* \in \cap_i \Delta(\mathcal{P}_i)$, $q_t(x,a) \leq u_t(x,a)$, and thus $\frac{q_t(x,a)}{u_t(x,a)}\ell_t(x,a) \leq 1$. Applying a union bound then finishes the proof. □

Lemma 13. *The OMD update with* $\widehat{q}_1(x,a,x') = \frac{1}{|X_k||A||X_{k+1}|}$ *for all* $k < L$ *and* $(x,a,x') \in X_k \times A \times X_{k+1}$, *and*

$$\widehat{q}_{t+1} = \operatorname*{argmin}_{q \in \Delta(\mathcal{P}_{i_t})} \eta\langle q, \widehat{\ell}_t \rangle + D(q \,\|\, \widehat{q}_t)$$

where $D(q \,\|\, q') = \sum_{x,a,x'} q(x,a,x') \ln \frac{q(x,a,x')}{q'(x,a,x')} - \sum_{x,a,x'} (q(x,a,x') - q'(x,a,x'))$ *ensures*

$$\sum_{t=1}^{T} \langle \widehat{q}_t - q, \widehat{\ell}_t \rangle \leq \frac{L\ln(|X|^2|A|)}{\eta} + \eta \sum_{t,x,a} \widehat{q}_t(x,a)\widehat{\ell}_t(x,a)^2$$

for any $q \in \cap_i \Delta(\mathcal{P}_i)$, *as long as* $\widehat{\ell}_t(x,a) \geq 0$ *for all* t, x, a.

Proof. Define $\tilde{q}_{t+1}$ such that

$$\tilde{q}_{t+1}(x,a,x') = \widehat{q}_t(x,a,x') \exp\left(-\eta\widehat{\ell}_t(x,a)\right).$$

It is straightforward to verify $\widehat{q}_{t+1} = \operatorname{argmin}_{q \in \Delta(\mathcal{P}_{i_t})} D(q \,\|\, \tilde{q}_{t+1})$ and also

$$\eta\langle \widehat{q}_t - q, \widehat{\ell}_t \rangle = D(q \,\|\, \widehat{q}_t) - D(q \,\|\, \tilde{q}_{t+1}) + D(\widehat{q}_t \,\|\, \tilde{q}_{t+1}).$$

By the condition $q \in \Delta(\mathcal{P}_{i_t})$ and the generalized Pythagorean theorem we also have $D(q \,\|\, \widehat{q}_{t+1}) \leq D(q \,\|\, \tilde{q}_{t+1})$ and thus

$$\begin{aligned}\eta\sum_{t=1}^{T} \langle \widehat{q}_t - q, \widehat{\ell}_t \rangle &\leq \sum_{t=1}^{T} \left(D(q \,\|\, \widehat{q}_t) - D(q \,\|\, \widehat{q}_{t+1}) + D(\widehat{q}_t \,\|\, \tilde{q}_{t+1})\right) \\ &= D(q \,\|\, \widehat{q}_1) - D(q \,\|\, \widehat{q}_{T+1}) + \sum_{t=1}^{T} D(\widehat{q}_t \,\|\, \tilde{q}_{t+1}).\end{aligned}$$

The first two terms can be rewritten as

$$\sum_{k=0}^{L-1} \sum_{x \in X_k} \sum_{a \in A} \sum_{x' \in X_{k+1}} q(x,a,x') \ln \frac{\widehat{q}_{T+1}(x,a,x')}{\widehat{q}_1(x,a,x')}$$

$$\leq \sum_{k=0}^{L-1}\sum_{x\in X_k}\sum_{a\in A}\sum_{x'\in X_{k+1}} q(x,a,x')\ln(|X_k||A||X_{k+1}|) \qquad \text{(by definition of } \widehat{q}_1\text{)}$$
$$= \sum_{k=0}^{L-1} \ln(|X_k||A||X_{k+1}|) \leq L\ln(|X|^2|A|).$$

It remains to bound the term $D(\widehat{q}_t \parallel \tilde{q}_{t+1})$:

$$D(\widehat{q}_t \parallel \tilde{q}_{t+1}) = \sum_{k=0}^{L-1}\sum_{x\in X_k}\sum_{a\in A}\sum_{x'\in X_{k+1}} \left(\eta\widehat{q}_t(x,a,x')\widehat{\ell}_t(x,a) - \widehat{q}_t(x,a,x') + \widehat{q}_t(x,a,x')\exp\left(-\eta\widehat{\ell}_t(x,a)\right)\right)$$
$$\leq \eta^2 \sum_{k=0}^{L-1}\sum_{x\in X_k}\sum_{a\in A}\sum_{x'\in X_{k+1}} \widehat{q}_t(x,a,x')\widehat{\ell}_t(x,a)^2$$
$$= \eta^2 \sum_{x\in X, a\in A} \widehat{q}_t(x,a)\widehat{\ell}_t(x,a)^2$$

where the inequality is due to the fact $e^{-z} \leq 1 - z + z^2$ for all $z \geq 0$. This finishes the proof. □

Bounding $\textsc{Bias}_2$. It remains to bound the term $\textsc{Bias}_2 = \sum_{t=1}^{T}\langle q^*, \widehat{\ell}_t - \ell_t\rangle$, which can be done via a direct application of Lemma 11.

Lemma 14. *With probability at least* $1-5\delta$*,* UOB-REPS *ensures* $\textsc{Bias}_2 = \mathcal{O}\left(\frac{L\ln(|X||A|/\delta)}{\gamma}\right)$.

Proof. For each state-action pair (x,a), we apply Eq. (18) in Lemma 11 with $\alpha_t(x',a') = 2\gamma\mathbb{I}_{\{x'=x,a'=a\}}$, which shows that with probability at least $1 - \frac{\delta}{|X||A|}$,

$$\sum_{t=1}^{T}\left(\widehat{\ell}_t(x,a) - \frac{q_t(x,a)}{u_t(x,a)}\ell_t(x,a)\right) \leq \frac{1}{2\gamma}\ln\left(\frac{|X||A|}{\delta}\right).$$

Taking a union bound over all state-action pairs shows that with probability at least $1-\delta$, we have for all occupancy measure $q \in \Omega$,

$$\sum_{t=1}^{T}\left\langle q, \widehat{\ell}_t - \ell_t\right\rangle \leq \sum_{t,x,a} q(x,a)\ell_t(x,a)\left(\frac{q_t(x,a)}{u_t(x,a)} - 1\right) + \sum_{x,a}\frac{q(x,a)\ln\frac{|X||A|}{\delta}}{2\gamma}$$
$$= \sum_{t,x,a} q(x,a)\ell_t(x,a)\left(\frac{q_t(x,a)}{u_t(x,a)} - 1\right) + \frac{L\ln\frac{|X||A|}{\delta}}{2\gamma}.$$

Note again that under the event of Lemma 2, we have $q_t(x,a) \leq u_t(x,a)$, so the first term of the bound above is nonpositive. Applying a union bound and taking $q = q^\star$ finishes the proof. □