A Proof of Theorem 4.2 Cont.

In the proof of Theorem 4.2 we showed that the following optimization problem

$$q_{t+1} = \arg \min_{q \in \Delta(M, i(t))} D(q||\tilde{q}_{t+1})$$

can be reformulated as the following convex optimization problem ($i = i(t)$):

$$\begin{aligned}
&\min_{q,\epsilon} D(q||\tilde{q}_{t+1}) \\
&s.t. \sum_{x \in X_k} \sum_{a \in A} \sum_{x' \in X_{k+1}} q(x,a,x') = 1 && \forall k = 0, \dots, L-1 \\
&\sum_{x' \in X_{k+1}} \sum_{a \in A} q(x,a,x') = \sum_{x' \in X_{k-1}} \sum_{a \in A} q(x',a,x) && \forall k = 1, \dots, L-1 \quad \forall x \in X_k \\
&q(x,a,x') - \bar{P}_i(x'|x,a) \sum_{y \in X_{k+1}} q(x,a,y) \leq \epsilon(x,a,x') && \forall k = 0, \dots, L-1 \quad \forall (x,a,x') \in X_k \times A \times X_{k+1} \\
&\bar{P}_i(x'|x,a) \sum_{y \in X_{k+1}} q(x,a,y) - q(x,a,x') \leq \epsilon(x,a,x') && \forall k = 0, \dots, L-1 \quad \forall (x,a,x') \in X_k \times A \times X_{k+1} \\
&\sum_{x' \in X_{k+1}} \epsilon(x,a,x') \leq \epsilon_i(x,a) \sum_{x' \in X_{k+1}} q(x,a,x') && \forall k = 0, \dots, L-1 \quad \forall (x,a) \in X_k \times A \\
&q(x,a,x') \geq 0 && \forall k = 0, \dots, L-1 \quad \forall (x,a,x') \in X_k \times A \times X_{k+1}
\end{aligned}$$

Now we will derive the solution to this problem using Lagrange multipliers. First we write the Lagrangian with $\lambda, \beta, \mu, \mu^+, \mu^-$ as Lagrange multipliers. Notice that we omit the non-negativity constraints, which we can justify since the solution will be non-negative anyway.

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

Let $(x,a,x') \in X \times A \times X_{k(x)+1}$ and consider the derivative with respect to $\epsilon(x,a,x')$.

$$\frac{\partial \mathcal{L}}{\partial \epsilon(x,a,x')} = -\mu^+(x,a,x') - \mu^-(x,a,x') + \mu(x,a)$$

So setting the gradient to zero we obtain

$$\mu(x,a) = \mu^+(x,a,x') + \mu^-(x,a,x')$$

Thus, we can discard $\mu(x,a)$ to obtain an equivalent Lagrangian. Notice that this way we also get rid of the $\epsilon(x,a,x')$ variables.

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

Now we consider the derivative with respect to $q(x,a,x')$. We denote $\beta(x_0) = \beta(x_L) = 0$ to avoid addressing the edge cases explicitly.

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial q(x,a,x')} &= \ln q(x,a,x') - \ln \tilde{q}_{t+1}(x,a,x') + \lambda_k + \beta(x) - \beta(x') \\
&\quad + (1-\epsilon_i(x,a))\mu^+(x,a,x') - (1+\epsilon_i(x,a))\mu^-(x,a,x') \\
&\quad + \sum_{y\in X_{k(x)+1}} \bar{P}_i(y|x,a)(\mu^-(x,a,y) - \mu^+(x,a,y))
\end{aligned}$$

We define the following value function v and error function e parameterized by μ and β, and an estimated Bellman error.

$$\begin{aligned}
v^\mu(x,a,x') &= \mu^-(x,a,x') - \mu^+(x,a,x') \\
e^{\mu,\beta}(x,a,x') &= (\mu^+(x,a,x') + \mu^-(x,a,x'))\epsilon_i(x,a) + \beta(x') - \beta(x) \\
B_t^{v,e}(x,a,x') &= e(x,a,x') + v(x,a,x') - \eta z_t(x,a,x') - \sum_{y\in X_{k(x)+1}} \bar{P}_i(y|x,a) v(x,a,y)
\end{aligned}$$

So the derivative becomes

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial q(x,a,x')} &= \ln \frac{q(x,a,x')}{\tilde{q}_{t+1}(x,a,x')} + \lambda_k - e^{\mu,\beta}(x,a,x') - v^\mu(x,a,x') + \sum_{y\in X_{k(x)+1}} \bar{P}_i(y|x,a) v^\mu(x,a,y) \\
&= \ln q(x,a,x') - \ln \tilde{q}_{t+1}(x,a,x') + \lambda_k - \eta z_t(x,a,x') - B_t^{v^\mu, e^{\mu,\beta}}(x,a,x')
\end{aligned}$$

Setting the gradient to zero and using the explicit form of $\tilde{q}_{t+1}(x,a,x')$ we obtain

$$\begin{aligned}
q_{t+1}(x,a,x') &= \tilde{q}_{t+1}(x,a,x') e^{-\lambda_k + \eta z_t(x,a,x') + B_t^{v^\mu, e^{\mu,\beta}}(x,a,x')} \\
&= q_t(x,a,x') e^{-\eta z_t(x,a,x')} e^{-\lambda_k + \eta z_t(x,a,x') + B_t^{v^\mu, e^{\mu,\beta}}(x,a,x')} \\
&= q_t(x,a,x') e^{-\lambda_k + B_t^{v^\mu, e^{\mu,\beta}}(x,a,x')}
\end{aligned}$$

We can use the first constraint to discover that λ_k is a normalizer for every $k = 0, \ldots, L-1$, i.e.

$$\begin{aligned}
1 &= \sum_{x \in X_k} \sum_{a \in A} \sum_{x' \in X_{k+1}} q_{t+1}(x, a, x') \\
1 &= \sum_{x \in X_k} \sum_{a \in A} \sum_{x' \in X_{k+1}} q_t(x, a, x') e^{-\lambda_k + B_t^{v^\mu, e^{\mu,\beta}}(x,a,x')} \\
e^{\lambda_k} &= \sum_{x \in X_k} \sum_{a \in A} \sum_{x' \in X_{k+1}} q_t(x, a, x') e^{B_t^{v^\mu, e^{\mu,\beta}}(x,a,x')}
\end{aligned}$$

so defining $Z_t^k(v, e) = \sum_{x \in X_k} \sum_{a \in A} \sum_{x' \in X_{k+1}} q_t(x, a, x') e^{B_t^{v,e}(x,a,x')}$, we obtain

$$q_{t+1}(x, a, x') = \frac{q_t(x, a, x') e^{B^{v^\mu, e^{\mu,\beta}}(x,a,x')}}{Z_t^{k(x)}(v^\mu, e^{\mu,\beta})}$$

Now to find β and μ we consider the dual problem. Substituting q_{t+1} back into $\mathcal{L}$ we obtain the following dual problem.

$$\max_{\beta,\mu \geq 0} \min_q \mathcal{L}(q) = \max_{\beta,\mu \geq 0} \mathcal{L}(q_{t+1}) = \max_{\beta,\mu \geq 0} -\sum_{k=0}^{L-1} \ln Z_t^k(v^\mu, e^{\mu,\beta}) - 1 + \sum_{x,a,x'} \tilde{q}_{t+1}(x, a, x')$$

So after ignoring constants we observe that

$$\beta_t, \mu_t = \arg \min_{\beta,\mu \geq 0} \sum_{k=0}^{L-1} \ln Z_t^k(v^\mu, e^{\mu,\beta})$$

B Proof of Theorem 5.2

First we reduce bounding $\hat{R}_{1:T}^{APP}$ to bounding the L_1-distance between q^{P_t,π_t} and q^{P,π_t}, where $P_t = P^{q_t}$ and $\pi_t = \pi^{q_t}$.

$$\begin{aligned}
\hat{R}_{1:T}^{APP} &= \sum_{t=1}^{T} \mathcal{C}(\mathbb{E}\left[\ell_t(U)|P, \pi_t\right]) - \mathcal{C}(\mathbb{E}\left[\ell_t(U)|P_t, \pi_t\right]) \\
&= \sum_{t=1}^{T} f^{\mathcal{C}}(q^{P,\pi_t}; \ell_t) - f^{\mathcal{C}}(q^{P_t,\pi_t}; \ell_t) \\
&\leq \sum_{t=1}^{T} \left\langle \bar{z}_t, q^{P,\pi_t} - q^{P_t,\pi_t} \right\rangle \qquad (1) \\
&\leq \sum_{t=1}^{T} \|\bar{z}_t\|_\infty \left\| q^{P,\pi_t} - q^{P_t,\pi_t} \right\|_1 \qquad (2) \\
&\leq F \sum_{t=1}^{T} \left\| q^{P,\pi_t} - q^{P_t,\pi_t} \right\|_1 \qquad (3)
\end{aligned}$$

where $\bar{z}_t \in \partial f^{\mathcal{C}}(q^{P,\pi_t}; \ell_t)$ and (1) follows from the definition of the sub-gradient, (2) follows from Hölder's inequality, and (3) follows because $f^{\mathcal{C}}$ is F-Lipschitz.

Therefore, We are left with bounding $\sum_{t=1}^{T} \left\| q^{P,\pi_t} - q^{P_t,\pi_t} \right\|_1$. From now on, we follow arguments from the regret analysis of UCRL-2, since we just need to bound the distance between occupancy measures that are in the confidence sets, and the performance criterion is not involved anymore.

We introduce some new notations that will simplify some equations. we denote the probability to visit a state-action pair (x, a) (or a state x) under occupancy measure q as $q(x, a)$ (or $q(x)$), i.e.,

$$q(x,a) = \sum_{x' \in X_{k(x)+1}} q(x,a,x')$$
$$q(x) = \sum_{a \in A} q(x,a)$$

In addition, for every $(x, a) \in X \times A$ and every $t = 1, \ldots, T$, denote $\xi_t(x, a) = \|P_t(\cdot|x, a) - P(\cdot|x, a)\|_1$.

Now we show how to use these notations to bound the aforementioned L_1-distance.

Lemma B.1. *Let $\{\pi_t\}_{t=1}^T$ be policies and let $\{P_t\}_{t=1}^T$ be transition functions. Then,*

$$\sum_{t=1}^{T} \left\| q^{P_t,\pi_t} - q^{P,\pi_t} \right\|_1 \leq \sum_{t=1}^{T} \sum_{x \in X} \sum_{a \in A} |q^{P_t,\pi_t}(x,a) - q^{P,\pi_t}(x,a)| + \sum_{t=1}^{T} \sum_{x \in X} \sum_{a \in A} q^{P,\pi_t}(x,a) \xi_t(x,a) \tag{4}$$

Proof. For every $(x, a) \in X \times A$ it holds that

$$\begin{aligned}
\sum_{x' \in X_{k(x)+1}} |q^{P_t,\pi_t}(x,a,x') - q^{P,\pi_t}(x,a,x')| &= \sum_{x' \in X_{k(x)+1}} |q^{P_t,\pi_t}(x,a)P_t(x'|x,a) - q^{P,\pi_t}(x,a)P(x'|x,a)| \\
&\leq \sum_{x' \in X_{k(x)+1}} |q^{P_t,\pi_t}(x,a)P_t(x'|x,a) - q^{P,\pi_t}(x,a)P_t(x'|x,a)| \\
&\quad + |q^{P,\pi_t}(x,a)P_t(x'|x,a) - q^{P,\pi_t}(x,a)P(x'|x,a)| \\
&= \sum_{x' \in X_{k(x)+1}} |q^{P_t,\pi_t}(x,a) - q^{P,\pi_t}(x,a)| P_t(x'|x,a) \\
&\quad + |P_t(x'|x,a) - P(x'|x,a)| q^{P,\pi_t}(x,a) \\
&= |q^{P_t,\pi_t}(x,a) - q^{P,\pi_t}(x,a)| + q^{P,\pi_t}(x,a)\xi_t(x,a)
\end{aligned}$$

Summing this for all $t = 1, \ldots, T$ and all $(x, a) \in X \times A$ gives the result. □

Thus, we need to bound each of the terms on the right hand side of (4). First, we show how to bound the first term on the right hand side of (4) using the second term.

Lemma B.2. *Let $\{\pi_t\}_{t=1}^T$ be policies and let $\{P_t\}_{t=1}^T$ be transition functions. Then, for every $k = 1, \ldots, L-1$ and every $t = 1, \ldots, T$, it holds that*

$$\sum_{x_k \in X_k} \sum_{a_k \in A} |q^{P_t,\pi_t}(x_k,a_k) - q^{P,\pi_t}(x_k,a_k)| \leq \sum_{s=0}^{k-1} \sum_{x_s \in X_s} \sum_{a_s \in A} q^{P,\pi_t}(x_s,a_s)\xi_t(x_s,a_s)$$

Proof. We prove the statement by induction on k. For $k = 1$ we have

$$\begin{aligned}
&\sum_{x_1 \in X_1} \sum_{a_1 \in A} |q^{P_t,\pi_t}(x_1, a_1) - q^{P,\pi_t}(x_1, a_1)| = \\
&\qquad = \sum_{a_0 \in A} \sum_{x_1 \in X_1} \sum_{a_1 \in A} |\pi_t(a_0|x_0) P_t(x_1|x_0, a_0) \pi_t(a_1|x_1) - \pi_t(a_0|x_0) P(x_1|x_0, a_0) \pi_t(a_1|x_1)| \\
&\qquad = \sum_{a_0 \in A} \pi_t(a_0|x_0) \sum_{x_1 \in X_1} |P_t(x_1|x_0, a_0) - P(x_1|x_0, a_0)| \sum_{a_1 \in A} \pi_t(a_1|x_1) \\
&\qquad \leq \sum_{a_0 \in A} \pi_t(a_0|x_0) \xi_t(x_0, a_0) \\
&\qquad = \sum_{a_0 \in A} q^{P,\pi_t}(x_0, a_0) \xi_t(x_0, a_0)
\end{aligned}$$

Now assume that the statement holds for some $k - 1$. We have

$$\begin{aligned}
&\sum_{x_k \in X_k} \sum_{a_k \in A} |q^{P_t,\pi_t}(x_k, a_k) - q^{P,\pi_t}(x_k, a_k)| = \\
&= \sum_{x_{k-1}} \sum_{a_{k-1}} \sum_{x_k} \sum_{a_k} |q^{P_t,\pi_t}(x_{k-1}, a_{k-1}) P_t(x_k|x_{k-1}, a_{k-1}) - q^{P,\pi_t}(x_{k-1}, a_{k-1}) P(x_k|x_{k-1}, a_{k-1})| \pi_t(a_k|x_k) \\
&= \sum_{x_{k-1}} \sum_{a_{k-1}} \sum_{x_k} |q^{P_t,\pi_t}(x_{k-1}, a_{k-1}) P_t(x_k|x_{k-1}, a_{k-1}) - q^{P,\pi_t}(x_{k-1}, a_{k-1}) P(x_k|x_{k-1}, a_{k-1})| \\
&\leq \sum_{x_{k-1}} \sum_{a_{k-1}} \sum_{x_k} |q^{P_t,\pi_t}(x_{k-1}, a_{k-1}) P_t(x_k|x_{k-1}, a_{k-1}) - q^{P,\pi_t}(x_{k-1}, a_{k-1}) P_t(x_k|x_{k-1}, a_{k-1})| \\
&\qquad\qquad + |q^{P,\pi_t}(x_{k-1}, a_{k-1}) P_t(x_k|x_{k-1}, a_{k-1}) - q^{P,\pi_t}(x_{k-1}, a_{k-1}) P(x_k|x_{k-1}, a_{k-1})| \\
&\leq \sum_{x_{k-1}} \sum_{a_{k-1}} |q^{P_t,\pi_t}(x_{k-1}, a_{k-1}) - q^{P,\pi_t}(x_{k-1}, a_{k-1})| + \sum_{x_{k-1}} \sum_{a_{k-1}} q^{P,\pi_t}(x_{k-1}, a_{k-1}) \xi_t(x_{k-1}, a_{k-1})
\end{aligned}$$

Finally, we use the induction hypothesis to obtain

$$\begin{aligned}
&\sum_{x_k \in X_k} \sum_{a_k \in A} |q^{P_t,\pi_t}(x_k, a_k) - q^{P,\pi_t}(x_k, a_k)| \leq \\
&\qquad \leq \sum_{s=0}^{k-2} \sum_{x_s \in X_s} \sum_{a_s \in A} q^{P,\pi_t}(x_s, a_s) \xi_t(x_s, a_s) + \sum_{x_{k-1} \in X_{k-1}} \sum_{a_{k-1} \in A} q^{P,\pi_t}(x_{k-1}, a_{k-1}) \xi_t(x_{k-1}, a_{k-1}) \\
&\qquad = \sum_{s=0}^{k-1} \sum_{x_s \in X_s} \sum_{a_s \in A} q^{P,\pi_t}(x_s, a_s) \xi_t(x_s, a_s)
\end{aligned}$$

□

The following lemma will show how to bound the second term on the right hand side of (4), and therefore obtain the bound on $\hat{R}^{APP}_{1:T}$. The proof follows the proof of Lemma 5 in Neu et al. (2012).

Lemma B.3. *Let $\{\pi_t\}_{t=1}^T$ be policies and let $\{P_t\}_{t=1}^T$ be transition functions such that $q^{P_t,\pi_t} \in \Delta(M, i(t))$ for every t. Then, with probability at least $1 - 2\delta$,*

$$\sum_{t=1}^{T} \sum_{k=0}^{L-1} \sum_{s=0}^{k-1} \sum_{x_s \in X_s} \sum_{a_s \in A} q^{P,\pi_t}(x_s, a_s) \xi_t(x_s, a_s) \leq 2L|X| \sqrt{2T \ln \frac{L}{\delta}} + 3L|X| \sqrt{2T|A| \ln \frac{T|X||A|}{\delta}}$$

Proof. We start by some arguments from the regret analysis of UCRL-2 (Auer et al., 2008). Let $n_i(x,a)$ be the number of times state-action pair (x,a) has been visited in epoch E_i. Therefore, we have

$$N_i(x,a) = \sum_{j=1}^{i-1} n_j(x,a)$$

We denote by m the number of epochs, and by Auer et al. (2008), we have

$$\sum_{i=1}^{m} \frac{n_i(x,a)}{\sqrt{N_i(x,a)}} \leq 3\sqrt{N_m(x,a)}$$

Now by Jensen's inequality,

$$\sum_{x\in X}\sum_{a\in A}\sum_{i=1}^{m} \frac{n_i(x,a)}{\sqrt{N_i(x,a)}} \leq 3\sqrt{|X||A|T}$$

Fix arbitrary $1 \leq t \leq T$ and $0 \leq k \leq L-1$. We have

$$\begin{aligned}
&\sum_{s=0}^{k-1}\sum_{x_s\in X_s}\sum_{a_s\in A} q^{P,\pi_t}(x_s,a_s)\xi_t(x_s,a_s) \leq \qquad (5)\\
&\leq \sum_{s=0}^{k-1}\xi_t(x_s^{(t)},a_s^{(t)}) + \sum_{s=0}^{k-1}\sum_{x_s\in X_s}\sum_{a_s\in A}\left(q^{P,\pi_t}(x_s,a_s) - \mathbb{I}\{x_s^{(t)}=x_s, a_s^{(t)}=a_s\}\right)\xi_t(x_s,a_s)
\end{aligned}$$

Now, by Lemma 4.1, we have with probability at least $1-\delta$ simultaneously for all s that

$$\begin{aligned}
\sum_{t=1}^{T}\xi_t(x_s^{(t)},a_s^{(t)}) &\leq \sum_{t=1}^{T}\sqrt{\frac{2|X_{s+1}|\ln\frac{T|X||A|}{\delta}}{\max\{1,N_{i(t)}(x_s^{(t)},a_s^{(t)})\}}}\\
&\leq \sum_{x_s\in X_s}\sum_{a_s\in A}\sum_{i=1}^{m} n_i(x_s,a_s)\sqrt{\frac{2|X_{s+1}|\ln\frac{T|X||A|}{\delta}}{\max\{1,N_i(x_s,a_s)\}}}\\
&\leq 3\sqrt{2T|X_s||X_{s+1}||A|\ln\frac{T|X||A|}{\delta}}
\end{aligned}$$

For the second term on the right hand side of (5), notice that $\left(q^{P,\pi_t}(x_s) - \mathbb{I}\{x_s^{(t)}=x_s\}\right)$ form a martingale difference sequence with respect to $\{U_t\}_{t=1}^{T}$ and thus by Hoeffding-Azuma inequality and $\xi_t(x,a)\leq 2$, we have

$$\begin{aligned}
&\sum_{t=1}^{T}\sum_{a_s\in A}\left(q^{P,\pi_t}(x_s,a_s) - \mathbb{I}\{x_s^{(t)}=x_s,a_s^{(t)}=a_s\}\right)\xi_t(x_s,a_s) \leq\\
&\qquad\leq 2\sum_{t=1}^{T}\left(\sum_{a_s\in A} q^{P,\pi_t}(x_s,a_s) - \sum_{a_s\in A}\mathbb{I}\{x_s^{(t)}=x_s,a_s^{(t)}=a_s\}\right)\\
&\qquad= 2\sum_{t=1}^{T}\left(q^{P,\pi_t}(x_s) - \mathbb{I}\{x_s^{(t)}=x_s\}\right)\\
&\qquad\leq 2\sqrt{2T\ln\frac{L}{\delta}}
\end{aligned}$$

with probability at least $1-\delta/L$. Putting everything together, the union bound implies that we have, with probability at least $1-2\delta$ simultaneously for all $k=1,\dots,L-1$,

$$\begin{aligned}
\sum_{t=1}^{T}\sum_{s=0}^{k-1}\sum_{x_s\in X_s}\sum_{a_s\in A} q^{P,\pi_t}(x_s,a_s)\xi_t(x_s,a_s) &\le \sum_{s=0}^{k-1} 3\sqrt{2T|X_s||X_{s+1}||A|\ln\frac{T|X||A|}{\delta}} + \sum_{s=0}^{k-1} 2|X_s|\sqrt{2T\ln\frac{L}{\delta}} \\
&\le 3L\sum_{s=0}^{k-1}\frac{1}{L}\sqrt{2T|X_s||X_{s+1}||A|\ln\frac{T|X||A|}{\delta}} + \sum_{s=0}^{k-1} 2|X_s|\sqrt{2T\ln\frac{L}{\delta}} \\
&\le 3L\sqrt{2T|A|\left(\frac{|X|}{L}\right)^2\ln\frac{T|X||A|}{\delta}} + 2|X|\sqrt{2T\ln\frac{L}{\delta}} \\
&= 3|X|\sqrt{2T|A|\ln\frac{T|X||A|}{\delta}} + 2|X|\sqrt{2T\ln\frac{L}{\delta}}
\end{aligned}$$

where in the last step we used Jensen's inequality for the concave function $f(x,y)=\sqrt{xy}$ and the fact that $\sum_{s=0}^{k-1}|X_s|\le|X|$.

Summing up for all $k=0,\dots,L-1$ finishes the proof. □