Online Appendix for Coupled Experts and Mobile Service Offloading

Cem Tekin, Mihaela van der Schaar
Electrical Engineering Department, University of California, Los Angeles
Email: cmtkn@ucla.edu, mihaela@ee.ucla.edu

I. Proof of Theorem 1

We will first prove a lemma related to $P(\mathcal{D}_m)$.

Lemma 1: With probability

$$P(\mathcal{D}_m) := \left(1 - \frac{(m_T)^{2\alpha+1}(n_{\max})^2 K \log a\tau_{\text{cov}}}{L^2 T^{2\alpha+1}}\right) \times \left(1 - \frac{2(n_{\max})^K |A_k|}{a^2}\right),$$

the true set of transition probabilities in round m will lie in $\mathcal{D}_m$.

Proof: Let X_1 be the covering time of the Markov chain P^t, $t = (m-1)T/m_T + 1, \ldots, mT/m_T$, i.e., the time all states are visited starting from the beginning of round $m-1$. Similarly let X_i be the ith covering time, that is the time it takes to visit all states after visiting all states for the $i-1$th time. Let Y be the time when all the states of the Markov chain is observed at least $D := (\log(a) m_T^{2\alpha} (n_{\max})^{2K})/(L^2 T^{2\alpha})$ times. For simplicity, we assume that this number is an integer. If this is not an integer, then it can be rounded up to the smallest integer that is greater than it. Then we have $Y \leq X_1 + \ldots + X_D$. Using Markov inequality, we get $P(Y > T/m_T) \leq P(X_1 + \ldots + X_D > T/m_T) \leq (D\tau_{\text{cov}} m_T)/T$. Therefore, the probability that all the states are observed at least D times in a round is greater than or equal to $1 - D\tau_{\text{cov}} m_T/T$. Using a Chernoff bound, it can be shown that given all states are observed at least D times, the probability that the true transition probability matrix will lie in $\mathcal{D}_m$ is $(1 - 2(n_{\max})^K |A_k|/a^2)$. ∎

Even when the true transition probabilities lies in $\mathcal{D}_m$, the solution of the robust dynamic program is suboptimal. Next we bound the regret by bounding the suboptimality due to uncertainty and the suboptimality due to the correct transition probabilities not being in the uncertainty region $\mathcal{D}_m$.

The proof is similar to the proof of Theorem IV.1 in [1]. The difference is that we use m_T to control the variation in uncertainty in a round. When we have more rounds, this means that the variation in the transition probabilities will be small. However, the shorter each round, less explorations will be performed, so the sample mean transition probability estimates may not be accurate. We can balance this tradeoff by a careful choice of m_T. Whenever the true transition probabilities lies in the region of uncertainty $\mathcal{D}_m(\tau, \tau_{\text{cov}})$, using a similar analysis to [1], we can bound the regret by $K(n_{\max} - c)\left((z+1)L\left(\frac{2T}{m_T}\right)^{\alpha}\right)$. Whenever the true transition probabilities are not in

$\mathcal{D}_m(\tau, \tau_{\text{cov}})$, since in the worst case each machine can process c tasks at each time slot, the regret is bounded by $K(n_{\max} - c)$. We get the result by combining these two.

REFERENCES

[1] J. Y. Yu and S. Mannor, "Online learning in markov decision processes with arbitrarily changing rewards and transitions," in *Game Theory for Networks, 2009. GameNets' 09. International Conference on*. IEEE, 2009, pp. 314–322.