Beam Alignment and User Scheduling in mmWave Networks under Mobility

Jihyun Lee, *Student Member, IEEE* and Eylem Ekici, *Fellow, IEEE*

***Abstract*—In this paper, we study beam alignment in a millimeter wave (mmWave) network with multiple users, and consider an optimal transmission scheduling algorithm. The problem is posed as an infinite horizon average cost constrained Markov decision process (CMDP) with the goal of minimizing the average beam alignment overhead subject to the average rate constraint on each user. By using a structural result derived from the Lagrangian formulation of the CMDP, we show that the optimal policy should keep scheduling the users that are scheduled in the previous time slot unless an abrupt change in the beam direction happens. Using this result, the complexity of the problem decreases to polynomial in the number of users. We also provide a heuristic deterministic algorithm that achieves $(1+\epsilon)$ approximation of the optimal solution, with smaller ϵ at the cost of longer transmission interval of each user.**

I. Introduction

To overcome the high signal attenuation inherent at 30-300 GHz electromagnetic spectrum (which corresponds to10mm to 1mm wavelength), millimeter wave (mmWave) networks must employ highly directional beamforming antennas. However, the use of narrow beams makes link establishment and maintenance much more challenging than traditional omnidirectional antennas as a mmWave link is established only when the transmit and receive antenna beams are steered in the correct directions. Moreover, even a slight misalignment of the beam directions or signal interruption can easily lead to complete link breakage and requires frequent beam re-alignments to maintain seamless connectivity especially under mobility.

To enable beamforming, a BS (base station) and a UE (user equipment) have to go through beam searching procedure, which typically incurs tens to hundreds of milliseconds overhead for the initial link establishment if exhaustive search over all possible combinations of transmission and reception directions is performed through a sequential pilot transmission [1]. To reduce the overhead, current standard activities [2], [3] suggest a two-stage beam search technique, in which a coarse grained sector level sweep is performed, followed by beam-level alignment phase. However, since the mmWave channel frequently varies over time in mobile network, it may lead to unaffordable overhead to perform an exhaustive search from scratch every time. Hence, more efficient schemes exploiting the temporal correlation on the channel are preferred under mobility.

Fast beam search methods in mmWave networks under UE mobility have been extensively studied in the literature. In [7], a smart beam steering algorithm is proposed for fast directional link re-establishment under node mobility, which uses knowledge of the previous feasible antenna sector pair to narrow the sector search space. In [5] a priori aided (PA) channel tracking scheme is proposed to predict the support of beam space in the following time slots without a channel estimation under the assumption that there is no blockage. In [6], Kalman filter based tracking algorithm and an abrupt change detection method based on a threshold test are proposed and evaluated through simulation. In [8], based on linear dynamic model, the authors proposed probing protocols to identify the beam errors caused by link blockage and user movement. Despite of the large volume, however, they either only consider a single UE case or assume no blockage.

Beam alignment (BA) and transmission scheduling in multi-user mmWave networks has also drawn extensive attention lately. In [5], the authors consider point to multipoint channel estimation, but it is assumed that the number of RF chains are equivalent to the number of UEs to guarantee the spatial multiplexing of all UEs, and thus the UE scheduling problem is not considered. In [9], transmission scheduling of multiple links is studied with an objective of optimizing network throughput, but the problem is defined for the multiple point-to-point links. In [17], energy efficient joint beam alignment protocols is addressed, with the goal to minimize the power consumption subject to rate constraints. However, the problem is defined for two users.

It is worth noting that none of the above considers multiple UE scenarios under mobility and the impact of transmission scheduling on BA overhead over time. In this paper, we consider a mmWave network consisting of a fixed BS and multiple mobile UEs communicating with directional antenna patterns. Our objective is to find a UE schedule that minimizes the beam alignment overhead while satisfying minimum data transmission rate constraints for each UE. The main contributions of the paper are as follows:

- Unlike most of the works on mmWave beamforming (or beam tracking) algorithms which focus on the link level performance improvement, we consider the problem of UE scheduling in a mmWave networks (possibly involving a large number of UEs). In our system model, both the abrupt changes and slow variations in beam direction of each link due to UE mobility and environmental changes are taken into account. We formulate the scheduling problem with minimum data rate constraint as CMDP (constrained markov decision process) and its equivalent linear programming formulation is presented.
- To avoid exponential complexity in the number of UEs,

This work has been supported in part by NSF under Grants CNS-1731698.

we use Lagrangian multiplier method to convert the constrained problem to an unconstrained MDP and show that with the optimal policy, in each transmission schedule, the BS should continue to include the UE that is scheduled in the previous time unless an abrupt change happens to the UE. The optimal solution is a randomized mixture of the solutions of unconstrained MDP and the problem reduces to finding the optimal mixture ratio. Using this structural result, the complexity of the problem decreases to polynomial in the number of UEs.

- For a practical use, a deterministic scheduling algorithm is proposed. With this algorithm, the BS schedules the UEs in a circular order but with different consecutive transmission times allocated to each UE. The length of the transmission time at each UE's turn is determined by the rate requirement of the UE. This algorithm ensures that every UE is given a transmission opportunity in a finite time. It is shown that the algorithm achieves $(1+\epsilon)$ approximation of the optimal solution for cases where the minimum data rate requirement of individual user is sufficiently small compared to the channel capacity.

II. System model and Problem formulation

We consider a network consisting of a BS and N UEs labeled $n \in \{1,2,\cdots,N\}$. At the BS, K RF chains are deployed so that it can serve K UEs at a time. As shown in Fig. 1, a time slot consists of two segments: beam alignment and data transmission. We assume that the BS decides which UEs to serve at the beginning of each time slot, based on the information it has on each link. When a UE is selected for data transmission for the time slot, the BS and the UE first decide which beam to use before its data transmission by searching over the possible combinations of beams and finding a best beam with highest SNR.

A. Beam alignment

Beam alignment introduces overhead because it requires time and energy which can otherwise be used for data transmission. The overhead is proportional to the number of directions to be tested and thus depends on the prior information on the correct direction of the beam. For example, if a UE is static, the beam found in the previous time slot can be reused unless some abrupt changes in the environment happen. Let τ and τ_n denote the length of a time slot and the time consumed for beam alignment with UE n respectively. We assume that τ is set to a fixed value such that it can support a seamless connectivity during the time slot under UE mobility (unless an abrupt change such as blockage happens in the environment). To maintain the connectivity, at the beginning of each time slot, beam search algorithm first finds which beam to use for the data transmission by sequentially checking the directions from the one with the highest probability of UE presence to minimize the searching overhead. After the beam alignment, for the remaining time of the time slot, the data is transmitted. We assume that at each time slot t, the channel gain (or path gain) of n-th UE is independent and identically

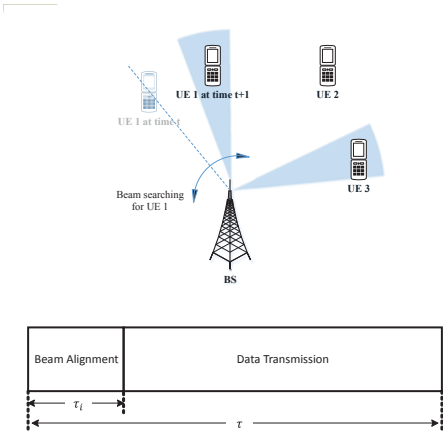

Fig. 1: An example of a multiuser mmWave network and its frame structure.

distributed with the expectation of θ_n. The data rate of UE n at time t is then determined by the channel gain if the beam is aligned and no blockage happens during the time slot. If blockage happens, the best beam direction in the next time slot can become completely independent to the previous one due to the loss of LOS path or change in the path of dominant reflection. In this case, a UE can disappear and re-appear in the following time slot uniform randomly on the entire angular search domain and we assume the search algorithm should scan the entire space. Note that as the time elapsed from the previous beam alignment increases, the information on the correct beam direction becomes more uncertain (i.e., less correlated to the previous direction) and the number of directions to be checked tends to increase and so does τ_n. We assume that the search algorithm should scan the entire space if the latest beam alignment becomes too outdated (e.g., if the time elapsed from the latest BA is larger than L).

B. Problem formulation: weakly coupled CMDP

When the beam alignment procedure in Section II-A is employed, at time t, **the system state space** is defined by a N-tuple $\mathcal{X} = \{(x_1,x_2,\cdots,x_N) : x_n \in \{1,2,\cdots,L\}\}$ and $x^t \in \mathcal{X}$ denote the state of time t, where x_n^t represents the amount of time that has been passed from the last beam alignment of UE n at the time of t. State L includes all time intervals which is greater than or equal to L. Similarly, **the action space** is defined by a N-tuple $\mathcal{A} = \{(a_1,a_2,\cdots,a_N) : a_n \in \{0,1\}, \sum_{n=1}^{N} \mathbb{1}(a_n = 1) \leq K\}$, where 0 stands for no transmission and 1 for transmission. $a^t \in \mathcal{A}$ is the action at time t. If UE n is scheduled for transmission at time t ($a_n^t = 1$), the BS first performs beam alignment with the UE and then transmits data for the remaining period of the time slot. The constraint on the action space is due to the limited number of RF chains in mmWave. At each time t, at most $K(\leq N)$ UEs can be selected for transmission.

Transition probability: The transition probability defines the evolution of the system and reflects the natural independence of UE transitions, i.e. state and action taken for an UE don't influence the transition of the other UEs. Thus, the state

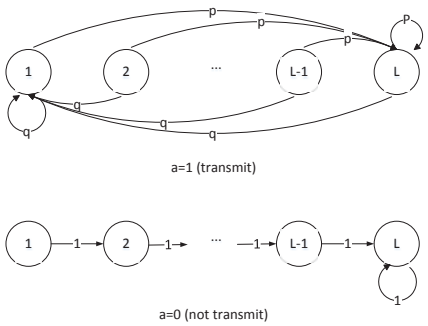

Fig. 2: transition probability graph for each of the two possible actions (transmit or not transmit) for a single user.

of each user transit according to independent homogenous transition law, i.e. the probability that $x^{t+1} = j$ given $x^t = i$ and $a^t = a$ is: $P(x^{t+1} = j | x^t = i, a^t = a) = \prod_{i=1}^{N} P^{a_n}_{i_n,j_n}$, where

$$P^{a_n}_{i_n,j_n} = P(x_n^{t+1} = j_n | x_n^t = i_n, a_n^t = a_n) = \begin{cases} 1 & \text{if } a_n = 0, j_n = \min(L, i_n + 1) \\ q & \text{if } a_n = 1, j_n = 1 \\ p & \text{if } a_n = 1, j_n = L \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

$q = 1 - p$ and $p \in (0,1)$ is the probability of abrupt change (blocking) which is independent across time. When a UE is scheduled for transmission, the state of the UE in the next time slot becomes 1 regardless of the state the UE is in at the current time slot. However, if blockage happens, the UE appears at a completely random direction in the next time slot independent of the current state, which means no prior information on the beam direction (i.e. no correlation with previous beam) and transition to state L. Transition probability graph of (1) is shown in Fig. 2. Since the transition probability is stationary, we can drop the time notation and use $P^a_{i,j} = \prod_{n=1}^{N} P(x_n^{t+1} = j_n | x_n^t = i_n, a_n^t = a_n)$.

Costs: We define data transmission and BA overhead with following assumptions.

(A1) Data transmission of UE n at state x_n and action a_n, $r_n(x_n, a_n)$ is non-increasing in x_n. BA overhead of UE n of state x_n and action a_n, $c_n(x_n, a_n)$ is non-decreasing in x_n.

(A2) The costs (or rewards) are dependent only on individual state and action and additive across UEs. More specifically,

- *Data transmission of m-th UE*: $r_m(x,a) = r_m(x_m, a_m)$.
- *Total power expenditure (BA overhead)*: $c(x,a) = \sum_{m=1}^{N} c_m(x,a) = \sum_{m=1}^{N} c_m(x_m, a_m)$.

It is assumed that fixed per-antenna power and a fixed symmetric antenna configuration for each RF chains are used. (A1) is a reasonable assumption since as more time passes from the last beam alignment, the uncertainty on the correct beam direction increases.

Under the assumption, if $K < N$ the problem is in the form of weakly coupled CMDP [10] with a linkage constraint $\sum_{n=1}^{N} \mathbb{1}(a_n = 1) \leq K$. If $K = N$, the problem can be decomposed to K independent subproblems of a single UE case [11]. The expected average data transmission rate and power cost associated to policy u are given by [13]:

$$C(u) = \lim_{T\to\infty} \frac{1}{T} \mathbb{E}^u \sum_{t=1}^{T} c(x^t, a^t), \quad (2)$$

$$R_m(u) = \lim_{T\to\infty} \frac{1}{T} \mathbb{E}^u \sum_{t=1}^{T} r_m(x^t, a^t),\ m = 1,2,\cdots,N. \quad (3)$$

The problem of optimizing a transmission policy is formally given by

$$C^* = \inf_u C(u) \quad \text{s.t.} R_m(u) \geq \gamma_m,\ m = 1,2,\cdots,N, \quad (4)$$

where γ_m is a constant minimum average data rate required by UE m.

C. Equivalent LP formulation

We note that the MDP of our problem is unichain, i.e. under any deterministic policy, the corresponding Markov chain contains a single ergodic class. Thus, the problem of (4) is equivalent to the following linear programming (LP) [13]:

$$\min_v \sum_{x\in\mathcal{X}} \sum_{a\in\mathcal{A}} c(x,a) v(x,a) \quad (5)$$

$$\text{s.t.} \sum_{x\in\mathcal{X}} \sum_{a\in\mathcal{A}} r_m(x,a) v(x,a) \geq \gamma_m,\ m = 1,2,\cdots,N \quad (6)$$

$$v \in \mathcal{V}, \quad (7)$$

where

$$\mathcal{V} = \left\{ \begin{array}{l} v(x,a), x \in \mathcal{X}, a \in \mathcal{A}: \\ (C1) \sum_{x\in\mathcal{X}} \sum_{a\in\mathcal{A}} v(x,a)(\delta_y(x) - P^a_{x,y}) = 0, y \in \mathcal{X} \\ (C2) \sum_{x\in\mathcal{X}} \sum_{a\in\mathcal{A}} v(x,a) = 1 \\ (C3) v(x,a) \geq 0, \forall x, a \end{array} \right\}$$

$v(x,a) = \lim_{T\to\infty} \frac{1}{T} P^u(x^t = x, a^t = a)$, $a^t \in \mathcal{A}(x^t)$, which can be interpreted as the expected average number of times action a is executed in state x. δ_y is the Dirac probability measure concentrated on y and $P^u(E)$ is the probability of event E under policy u. The constraint set $\mathcal{V}$ can be interpreted as the conservation of flow through each of the states. An optimal policy can be computed from a solution to the LP as:

$$u_x(a) = \frac{v(x,a)}{\sum_{a\in\mathcal{A}} v(x,a)}, \quad (8)$$

where $u_x(a)$ is the probability that the controller executes action a when it encounters state x. This defines the stationary randomized policy u, which maps states to probability distributions over actions, i.e. $u : \mathcal{X} \times \mathcal{A} \to [0,1]$.

Curse of dimensionality In principle, the optimal policy can be found by solving the LP or dynamic programming (DP). However, the complexity of the CMDP is exacerbated for large number of UEs since the state and action spaces for the process typically consists of cross-product of those from individual UE processes, thus exponential in the number of UEs, i.e. $O(L^N)$

states. For heuristic techniques dealing with this problem, see [11] and [10]. In the following section, it will be shown that the complexity of our problem can be reduced to $O(N^K)$ by exploiting its structural property.

III. Optimal scheduling policy

In this section, an optimal UE scheduling algorithm with a polynomial complexity is presented.

A. Optimal scheduling of multiple UEs

We define a set $\bar{\mathcal{V}} \subset \mathcal{V}$ by adding an additional constraint (C4) on the set $\mathcal{V}$ as follows.

$$\bar{\mathcal{V}} = \left\{ \begin{array}{l} v(x,a), x \in \mathcal{X}, a \in \mathcal{A} : \\ (C1),(C2),(C3) \text{ and} \\ (C4) v(x,a) = 0, \forall (x,a) \notin \mathcal{G}, \end{array} \right\} \tag{9}$$

where

$$\mathcal{G} = \left\{ \begin{array}{l} (x,a), x \in \mathcal{X}, a \in \mathcal{A} : \\ \sum_{i=1}^{N} \mathbb{1}(x_i = 1) = m, \sum_{i=1}^{N} \mathbb{1}(x_i = L) = N - m, 0 \le m \le K, \\ \sum_{i=1}^{N} \mathbb{1}(x_i = 1, a_i = 0) = 0 \end{array} \right\}$$

In words, G is a set of state and action pairs $(x,a) \in \mathcal{X} \times \mathcal{A}$ where less than or equal to K UEs are in state 1, the others are in state L and action 1 is assigned to at least one of the UEs in state 1. Note also that $\bar{\mathcal{V}} \subset \mathcal{V}$ with most of the elements in $\mathcal{V}$ set to 0 and the number of unknowns in $\bar{\mathcal{V}}$ is $O(N^K)$.

Theorem 1: If the assumptions (A1)-(A2) are satisfied for the CMDP problem (4), Algorithm 1 achieves an optimal solution.

Proof: See Appendix A. ∎

Algorithm 1 uses the solution of (5)-(7) with the constraint (7) replaced by $v \in \bar{\mathcal{V}}$. Therefore, the complexity of the algorithm is polynomial in N. In the proof of the theorem 1, we use a Lagrangian approach which converts a constrained control problem into an equivalent minmax non-constrained control problem. This approach solves the problem (4) by adding a Lagrangian multiplier per additional constraint while every Lagrangian multiplier results in a separate policy. Then the optimal randomized policy of a CMDP is computed as a mix policy of multiple optimal pure policies for all the Lagrangian multipliers. (See [13], [14] for comprehensive discussions about this topic). The theorem is proved by the structural property (18) showing that (C4) holds for the pure policies of all the Lagrangian multipliers.

B. Application of Algorithm 1 to a single UE network

By Theorem 1, the optimal solution of (5)-(7) can be found by solving (5), (6) and $v \in \bar{\mathcal{V}}$. For a single UE network, $N = 1$, $K = 1$ and $\mathcal{G} = \{(1,1),(L,0),(L,1)\}$. Therefore, there are only three unknowns in $\bar{\mathcal{V}}$ and $v^* \in \bar{\mathcal{V}}$ that minimizes the cost function can be found by a simple calculation. From (8), the optimal policy u is as follows :

$$u_x(1) = \begin{cases} 1 & \text{if } x = 1 \\ \frac{p\gamma}{pr(L,1)+qr(1,1)-q\gamma} & \text{if } x = L \\ \text{arbitrary} & \text{else} \end{cases} \tag{10}$$

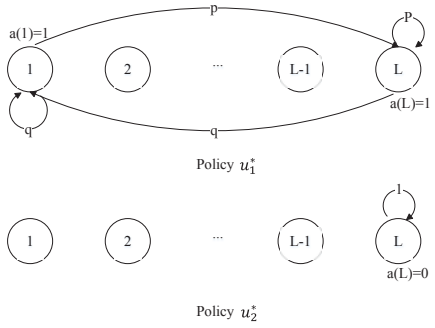

Fig. 3: Optimal solution of a single user case as a mixture of two deterministic policy. Note that the states $2, 3, \cdots, L-1$ are transient and the randomized decision occurs at state L.

and $u_x(0) = 1 - u_x(1)$. The optimal scheduling of the UE is to keep transmitting as long as it succeeds and once a failure (blockage) happens, the BS flips a coin and transmits the data with probability $\frac{p\gamma}{pr(L,1)+qr(1,1)-q\gamma}$ and stays idle with probability $1 - \frac{p\gamma}{pr(L,1)+qr(1,1)-q\gamma}$. Note that in the case of failure, it needs to search the beam from scratch while consecutive successes utilize the information of previous alignments. We note that the optimal solution (10) is a mix policy of u_1^* and u_2^* in Fig. 3.

Algorithm 1 Optimal Transmission Schedule

Data: $\mathcal{X}$, $\mathcal{A}$, $r(x,a)$ $c(x,a)$, $p_{i,j}^a$ for all $x \in \mathcal{X}$ and $a \in \mathcal{A}$, channel capacity θ_i, minimum rate requirement γ_i for each UE i

Result: transmission scheduling $\{a^t\}_{t=1,2,\cdots}$

Initialization: find $v^*(x,a)$ for all $x \in X$ and $a \in A$ by solving (5) s.t. (6) and $v \in \bar{\mathcal{V}}$. $t = 1$.

while *true* **do**
  if $t = 1$ **then**
    $x_i^t = L$ for all $i \in \mathcal{N}$
  end
  choose $a^t = a$ with probability $\frac{v^*(x^t,a)}{\sum_{u \in A} v^*(x^t,u)}$
  for $i = 1$ *to* N **do**
    if $a_i^t = 1$ **then**
      if *an abrupt change occur to user i* **then**
        $x_i^{t+1} = L$
      else
        $x_i^{t+1} = 1$
      end
    else
      $x_i^{t+1} = x_i^t + 1$
    end
  end
  $t = t + 1$
end

IV. Deterministic scheduling

Even though we can find an optimal solution of (4) with polynomial time complexity, the optimal policy is randomized.

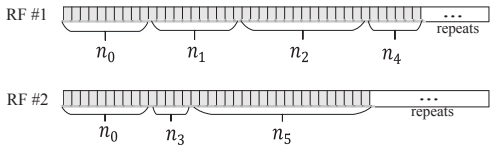

Fig. 4: An example transmission schedule of Algorithm 2, when there are five users ($N = 5$) and the number of RF chains are two ($K = 2$). In this example, the set of users $\mathcal{N} = \{1, 2, 3, 4, 5\}$ are partitioned into two groups such that $\mathcal{N}_1 = \{1, 2, 4\}$ and $\mathcal{N}_2 = \{3, 5\}$.

In practice, it is more desirable to use a deterministic policy [10]. In this section, we propose a deterministic policy with $(1 + \epsilon)$ approximation. In this algorithm, the UEs are first divided into K groups and the UEs in each group are scheduled in a circular order, but with different consecutive transmission time slots assigned to each of them depending on its rate requirement. An example is shown in Fig. 4.

Algorithm 2 Fixed Transmission Time Schedule

Data: $\mathcal{X}$, $\mathcal{A}$, $r(x,a)$ $c(x,a)$, $p^a_{i,j}$ for all $x \in \mathcal{X}$ and $a \in \mathcal{A}$, channel capacity θ_i, minimum rate requirement γ_i for each UE i

Result: transmission scheduling $\{a^t\}_{t=1,2,\cdots}$

initialization: find a parition $\{\mathcal{N}_k\}_{k=1}^K$, set n_0 and find n_{k_i} $\forall i \in \mathcal{N}_k$, $k = 1, \cdots, K$ according to (13), $t = 1$

while *true* **do**
  $a^t_i = 0$, $\forall i \in \mathcal{N}$
  for $k = 1$ *to* K **do**
    $d(k,t) = t \bmod (\sum_{i \in \mathcal{N}_k} n_i + n_0)$
    if $1 \le d(k,t) \le n_{k_1}$ **then**
      $a^t_{k_1} = 1$
    end
    for $j = 2$ *to* N_k **do**
      if $\sum_{l=1}^{j-1} n_{k_l} < d(k,t) \le \sum_{l=1}^{j} n_{k_l}$ **then**
        $a^t_{k_j} = 1$
      end
    end
  end
  $t = t + 1$
end

Theorem 2: If there exists a partition of a set $\mathcal{N}$ into K subsets $\mathcal{N}_1, \mathcal{N}_2, \cdots, \mathcal{N}_K$ such that for all $k \in \{1, 2, \cdots, K\}$, $\gamma_i \le \frac{q\theta_i}{3(|\mathcal{N}_k|-1)}(\tau - q\tau_1 - p\tau_L)$ for all $i \in \mathcal{N}_k$, then for any given $\epsilon > 0$, by setting $n_0 = \max_k \frac{b_k}{\epsilon} - a_k$, Algorithm 2 achieves an average cost function f that is **at most** $(1+\epsilon)f^*$, where f^* is the optimal solution of problem (5)-(7) and a_k and b_k are constant values defined in (14).

Proof: The proof proceeds in four steps.

STEP 1 (Performance bound for single RF): Let us denote the optimal solution of a single RF problem (K=1) with a multiuser set $\mathcal{N}$ by $f^*_{\mathcal{N}/1}$. If $\sum_{i \ne j} n_j + n_0 \ge L, \forall i \in \mathcal{N}$, the expected aggregated reward of user i during its n_i consecutive transmission (action 1) from any time $t_i \ge L$ is

$$\begin{aligned}\sum_{t=t_i}^{t_i+n_i-1} r_i(x^t, a^t) &= r_i(L,1) + \sum_{j=1,L} p^{(1)}_{L,j} r_i(j,1) \\ &+ \sum_{j=1,L} p^{(2)}_{L,j} r_i(j,1) + \cdots + \sum_{j=1,L} p^{(n_i-1)}_{L,j} r_i(j,1) \\ &= r_i(L,1) + (n_i - 1) p r_i(L,1) + (n_i - 1) q r_i(1,1),\end{aligned}$$

where $p^{(n)}_{i,j}$ is n-step transition probability from state i to state j. The last equality follows from that $p^{(n)}_{L,1} = q$ and $p^{(n)}_{L,L} = p$ for all $n \ge 1$. Similarly, the expected aggregate cost of user i during its n_i consecutive transmission is

$$\sum_{t=t_i}^{t_i+n_i-1} c_i(x^t, a^t) = c_i(L,1) + (n_i - 1) p c_i(L,1) + (n_i - 1) q c_i(1,1).$$

Therefore, using this deterministic policy, the problem of minimizing BA overhead subject to the rate constraint on each user can be written as follows.

$$\min_n \hat{f}_{\mathcal{N}/1} = \frac{1^T(c_L - c) + c^T n}{n_0 + 1^T n} \tag{11}$$

$$\text{s.t.}[\text{diag}(r) - \gamma 1^T] n \succcurlyeq (r - r_L) + n_0 \gamma, \tag{12}$$

where c, r, c_L, r_L and γ are $N \times 1$ vectors whose i-th component is $pc_i(L,1) + qc_i(1,1)$, $pr_i(L,1) + qr_i(1,1)$, $c_i(L,1)$, $r_i(L,1)$ and γ_i, respectively. Let $\text{diag}(r) - \gamma 1^T \equiv M$. To meet the rate constraints (12), we set:

$$n = round(M^{-1}[(r - r_L) + \gamma n_0]) + 1, \tag{13}$$

where $round(x)$ rounds x to the nearest integer. Putting this into (11),

$$\begin{aligned}\hat{f}_{\mathcal{N}/1} &= \frac{n_0(c^T M^{-1}\gamma) + 1^T(c_L - c) + c^T M^{-1}(r - r_L) + c^T \epsilon_0}{n_0 + 1^T M^{-1}(r - r_L) + 1^T M^{-1}\gamma n_0} \\ &= \frac{c^T M^{-1}\gamma}{1 + 1^T M^{-1}\gamma}\left(1 + \frac{b}{n_0 + a}\right) = c\delta^T\gamma\left(1 + \frac{b}{n_0 + a}\right) \\ &= f^*_{\mathcal{N}/1}\left(1 + \frac{b}{n_0 + a}\right),\end{aligned}$$

where

$$\begin{aligned}a &= \frac{1^T(M^{-1}(r - r_L)}{1 + 1^T M^{-1}\gamma}, \\ b &= \frac{1^T(c_L - c) + c^T M^{-1}(r - r_L) + c^T\epsilon_0}{c^T M^{-1}\gamma} - \frac{1^T M^{-1}(r - r_L)}{1 + 1^T M^{-1}\gamma},\end{aligned} \tag{14}$$

$\epsilon_0 = n - M^{-1}[(r - r_L) + \gamma n_0]$ and $\delta_i = r_i^{-1}$. Since $\gamma 1^T$ is rank 1, we used $(A + B)^{-1} = A^{-1} - \frac{1}{1+\text{tr}(BA^{-1})} A^{-1}BA^{-1}$ [15] to get $M^{-1} = \delta I - \frac{\gamma^T\delta}{1 - \delta\gamma}$.

Since $b \ge 0$, $\hat{f}_{\mathcal{N}/1}$ is decreasing in n_0 and as n_0 goes to ∞, $\hat{f}_{\mathcal{N}/1}$ approaches $c\delta^T\gamma = (c_1 q + c_L p)\sum_{i=1}^N \frac{\gamma_i}{r_i}$, which is $f^*_{\mathcal{N}/1}$ the optimal solution of the original problem (4) (or its LP formulation (5)-(7)). Therefore,

$$\frac{\hat{f}_{\mathcal{N}/1} - f^*_{\mathcal{N}/1}}{f^*_{\mathcal{N}/1}} \le \frac{b}{n_0 + a}$$

and for any given $\epsilon \geq 0$, we can achieve $\hat{f}_{\mathcal{N}/1} \leq (1+\epsilon) f^*_{\mathcal{N}/1}$ by setting $n_0 \geq \frac{b}{\epsilon} - a$.

STEP 2 (Show $f^*_{\mathcal{N}/1} = \sum_{i\in\mathcal{N}} f^*_i$): Let us denote the optimal solution of a single user-single RF problem by f^*_i. From Theorem 1, the problem (5)-(7) for multiuser $\mathcal{N}$ and a single RF is equivalent to the following:

$$
\begin{aligned}
f^*_{\mathcal{N}/1} = \min_{v} f_{\mathcal{N}/1} = \sum_{i\in\mathcal{N}} [\underbrace{c(L\mathbf{1}_{\mathcal{N}}, \mathbf{1}_{\{i\}})}_{=c_i(L,1)} v(L\mathbf{1}_{\mathcal{N}}, \mathbf{1}_{\{i\}}) \\
+ \underbrace{c(L\mathbf{1}_{\mathcal{N}} - (L-1)\mathbf{1}_{\{i\}}, \mathbf{1}_{\{i\}})}_{=c_i(1,1)} v(L\mathbf{1}_{\mathcal{N}} - (L-1)\mathbf{1}_{\{i\}}, \mathbf{1}_{\{i\}})] \\
\text{s.t.} \quad \underbrace{r(L\mathbf{1}_{\mathcal{N}}, \mathbf{1}_{\{i\}})}_{=r_i(L,1)} v(L\mathbf{1}_{\mathcal{N}}, \mathbf{1}_{\{i\}}) \\
+ \underbrace{r(L\mathbf{1}_{\mathcal{N}} - (L-1)\mathbf{1}_{\{i\}}, \mathbf{1}_{\{i\}})}_{=r_i(1,1)} v(L\mathbf{1}_{\mathcal{N}} - (L-1)\mathbf{1}_{\{i\}}, \mathbf{1}_{\{i\}}) \geq \gamma_i, \\
i = 1,2,\cdots,N. \\
pv(L\mathbf{1}_{\mathcal{N}} - (L-1)\mathbf{1}_{\{i\}}, \mathbf{1}_{\{i\}}) = qv(L\mathbf{1}_{\mathcal{N}}, \mathbf{1}_{\{i\}}), i = 1,2,\cdots,N. \\
\sum_{i\in\mathcal{N}} [v(L\mathbf{1}_{\mathcal{N}} - (L-1)\mathbf{1}_{\{i\}}, \mathbf{1}_{\{i\}}) + v(L\mathbf{1}_{\mathcal{N}}, \mathbf{1}_{\{i\}})] + v(L\mathbf{1}_{\mathcal{N}}, \mathbf{0}) = 1
\end{aligned}
$$

It is easy to see that the above problem is decomposable to N subproblems of the following:

$$
\begin{aligned}
f^*_i = \min_{v} f_i = c_i(L,1)v_i(L,1) + c_i(1,1)v_i(1,1) \\
\text{s.t. } r_i(L,1)v_i(L,1) + r_i(1,1)v_i(1,1) \geq \gamma_i \\
pv_i(1,1) = qv_i(L,1),
\end{aligned}
$$

if $\sum_i [v_i(1,1) + v_i(L,1)] \leq 1$. Since this is satisfied by the condition $\gamma_i \leq \frac{q\theta_i}{3(N-1)}(\tau - q\tau_1 - p\tau_L)$, $f^*_{\mathcal{N}/1} = \sum_{i\in\mathcal{N}} f^*_i$ and $v^*(L\mathbf{1}_{\mathcal{N}} - (L-1)\mathbf{1}_{\{i\}}, \mathbf{1}_{\{i\}}) = v^*_i(1,1)$ and $v^*(L\mathbf{1}_{\mathcal{N}}, \mathbf{1}_{\{i\}}) = v^*_i(L,1)$.

STEP 3 (Show $f^*_{\mathcal{N}/K} = \sum_{k=1}^{K} f^*_{\mathcal{N}_k/1}$): We denote the optimal solution of $\mathcal{N}$ multiuser-K RF problem and $\mathcal{N}_k$ multiuser-single RF problem by $f^*_{\mathcal{N}/K}$ and $f^*_{\mathcal{N}_k/1}$ respectively. First, $f^*_{N/K} \geq f^*_{N/N}$, since the action space $A_{N/K} \subset A_{N/N}$. Also, $f^*_{N/N}$ does not have any linkage constraints between the users $f^*_{N/N} = \sum_{i\in\mathcal{N}} f^*_i$. Therefore, 1) $f^*_{\mathcal{N}/K} \geq \sum_{i\in\mathcal{N}} f^*_i$. On the other hand, 2) $f^*_{\mathcal{N}/K} \leq \sum_{k=1}^{K} f^*_{\mathcal{N}_k/1} = \sum_{k=1}^{K} \sum_{i\in\mathcal{N}_k} f^*_i = \sum_{i\in\mathcal{N}} f^*_i$. The first inequality follows since the $\sum_{k=1}^{K} f^*_{\mathcal{N}_k}$ is optimal for a fixed partition and the first equality follows from the result of STEP 2. Combining 1) and 2), $f^*_{\mathcal{N}/K} = \sum_{k=1}^{K} f^*_{\mathcal{N}_k/1}$.

STEP 4 (Performance bound for multiple RF): We combine the results of STEP 1 through STEP 3.

$$
\begin{aligned}
\hat{f} - f^* = \hat{f}_{N/K} - f^*_{N/K} = \sum_{k=1}^{K} \hat{f}_{N_k/1} - \sum_{k=1}^{K} f^*_{N_k/1} \\
\leq \sum_{k=1}^{K} \frac{b_k}{n_0 + a_k} f^*_{N_k/1} \leq (\max_k \frac{b_k}{n_0 + a_k}) \sum_{k=1}^{K} f^*_{N_k/1} \\
\leq (\max_k \frac{b_k}{n_0 + a_k}) f^*_{N/K} = (\max_k \frac{b_k}{n_0 + a_k}) f^*.
\end{aligned}
$$

This completes the proof. ∎

Remark 1 (K-partition): We note that the assumption on the minimum required data rate for each user in Theorem 3 is necessary to guarantee the feasibility of k-partition problem. However, even though the feasibility is guaranteed, finding the partition is NP-complete since the problem reduces to a set partition problem [16]. In this section, we assume that γ_i are sufficiently small compared to the channel capacity, thus a partition can be easily found. For example, if γ_i is given such that $\gamma_i \leq \frac{q\theta_i}{3(\lceil N/K \rceil - 1)}(\tau - q\tau_1 - p\tau_L)$, for all $i \in \mathcal{N}$, then any partition that allocate less than or equal to $\lceil N/K \rceil$ users to each RF chain will be a feasible solution.

V. Performance Evaluation

In this section, the algorithms in Section III and Section IV are evaluated numerically. We consider a mmWave network consisting of a BS and multiple UEs. We assume the following probabilistic model for temporal changes in the beam orientation [18]:

- Slow change of beam orientation of each UE is modeled as an independent random walk on a polygon with L sides in angular domain. The change occurs at the beginning of each time slot. With probability $\alpha \in (0,1)$ the beam direction does not change and with $(1-2\alpha)$ it changes to the neighbouring directions due to UE mobility.

- An abrupt change in the beam direction occurs with probability $p > 0$. When the abrupt change occurs, the beam direction in the following time slot is determined by a uniform random selection on a polygon with L sides.

- Beam alignment is performed independently for each user and for a given scheduling, orthogonal beamforming exists. The data transmission of UE n is $r_n(x_n, a_n) = (1 - \frac{\tau_p N_p(x_n)}{\tau})\theta_n$ if $a_n = 1$, and 0 otherwise. θ_n is a known constant data rate of UE n when beam-aligned, τ_p is the time required for a pilot transmission and $N_p(x_n)$ is the number of pilot transmission needed for a beam alignment when the UE is at state x_n. The BA overhead of UE n is $c_n(x_n, a_n) = N_p(x_n)e_p$ if $a_n = 1$, and 0 otherwise. e_p is the power consumption of a pilot transmission. Note that $N_p(x_n)$ is determined by the realisation of the probabilistic change in beam orientation of the UE.

Table I shows the average cost of the proposed algorithms for different number of UEs and rate constraints for a given α and p. The BS has 2 RF chains and can serve at most 2 UEs at each time slot. We assume the same channel capacity (bps), i.e., $\theta_n = 1$ for every UE n when the UE is beam-aligned and 0 otherwise. The length of a time slot τ=1. For a pilot transmission, we set the cost $e_p = 0.1$ and $\tau_p = 0.05$. The quantities in Table I and Fig. 5 are averaged over 10^4 time slots and 100 different runs. We consider 5 UEs, i.e., $\mathcal{N} = \{1,2,3,4,5\}$ with different rate constraints. (B) is further divided into two cases in which different UE partition to each RF chain is used. For (C), we consider 2 UEs, i.e., $\mathcal{N} = \{1,2\}$. The minimum rate of UE 1 and UE 2 are set to be the same as the total rate of $\mathcal{N}_1$ and $\mathcal{N}_2$ of (B2), respectively.

(A) γ=[0.3, 0.3 0.3, 0.3, 0.3]
(B1) γ=[0.15, 0.15 0.15, 0.15, 0.15], $\mathcal{N}_1 = \{1,2,3,4\}$, $\mathcal{N}_2 = \{5\}$
(B2) γ=[0.15, 0.15 0.15, 0.15, 0.15], $\mathcal{N}_1 = \{1,2,3\}$, $\mathcal{N}_2 = \{4,5\}$
(C) γ=[0.45, 0.3], $\mathcal{N}_1 = \{1\}$, $\mathcal{N}_2 = \{2\}$

		(A)	(B1)	(B2)	(C)
LP solution		1.0293	0.5147	0.5147	0.5147
Alg. 1 (95 % CI)		1.0297 (±0.0017)	0.5137 (±0.0021)	0.5147 (±0.0017)	0.5126 (±0.0057)
Alg. 2	n_0=11	-	0.8018	0.9198	0.6987
	n_0=44	-	0.5948	0.6327	0.5622

TABLE I: Performance comparison with different rate requirements (γ) for given α and p ($\alpha = 0.5$, $p = 0.1$).

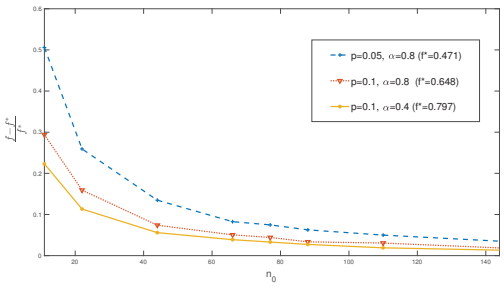

Fig. 5: Performance comparison with different settings of UE mobility (α), blocking probability (p) and n_0 for a given minimum rate requirement γ.

Since there does not exist any partition of $\mathcal{N}$ into two which can support the rate requirement of (A), the γ of (A) is not feasible for Algorithm 2, whereas Algorithm 1 has the average cost very close to the LP solution (The variation from the optimal is due to finite sampling. As the number of runs and time slots increases, this will converge to the LP solution). We note that the action space of Algorithm 2 is a subset of Algorithm 1 since it uses a fixed assignment of UEs to each RF chain. However, when the rate requirement is sufficiently small as in (B), we can readily find a feasible partition $\mathcal{N}_1$ and $\mathcal{N}_2$. As seen in the proof of Theorem 2, the average cost of the optimal policy does not depend on the partition as long as the given γ is feasible for partition. Therefore, the performance of Algorithm 1 is the same for (B1) and (B2). However, depending on the partition, the performance of Algorithm 2 can be different. This is because the constants a_k and b_k of (14) change as we use different partitions. It is shown in Table I. that the cost of (B2) is larger than (B1) for both n_0 =11 and n_0 =44. Regardless of which partition we use, however, the average cost decreases as n_0 increases. The result of (C) shows how the number of users affects the performance of the network. The BS serves UE 1 and UE 2 separately using different RF chain (and its corresponding antenna sets). For each RF, the rate constraint is the same as the total rate of (B2) and thus the optimal cost of (C) is the same as (B2). This is no surprise because for each $k \in \{1,2\}$, if we merge the states $\{(L\mathbf{1}_{\mathcal{N}_k} - (L-1)\mathbf{1}_i)\}_{i\in\mathcal{N}_k}$ of (B2) to a single state (state 1), we can obtain the same MDP as (C). However, the average cost of (C) is lower than (B2) when Algorithm 2 is used. This is because more users with smaller rate requirements causes larger rounding errors in (13).

Fig. 5 shows the performance of Algorithm 2 for different UE mobility (α) and probability of blockage (p) with a given rate requirement γ. 5 UEs with $\gamma_n = 0.1$ are assigned to each RF chain. Obviously, the average cost increases as either the probability of blocking (abrupt change) or UE mobility increases for both the optimal solution and Algorithm 2. It is also shown that the normalized error $\frac{f-f^*}{f^*}$ of Algorithm 2 decreases as n_0 increases as expected by Theorem 2.

VI. Conclusion

This paper explores transmission scheduling algorithms for mmWave networks under user mobility, where the beam alignment is required before each transmission. The problem of minimizing beam alignment overhead under the minimum rate constraints is formulated as CMDP. From the structural result derived from the Lagrangian formulation of the MDP, it is shown that the complexity of the CMDP can be reduced from $O(L^N)$ to $O(N^K)$. In addition, a heuristic deterministic algorithm is proposed and shown to achieve $(1+\epsilon)$ approximation of the optimal solution. In our future work, a joint optimization of transmission scheduling with power allocation and/or beam alignment methods (so that it includes an option of beam tracking for non-scheduled users or reuse of the previous beam alignment for scheduled users) will be investigated along with the study on the non-orthogonal beamforming.

References

[1] Xiao, Ming, et al. "Millimeter wave communications for future mobile networks." IEEE Journal on Selected Areas in Communications 35.9 (2017): 1909-1935.

[2] IEEE standard. "IEEE 802.11ad WLAN enhancements for very high throughput in the 60 GHz Band", 2012

[3] Wang, Junyi. "Beam codebook based beamforming protocol for multi-Gbps millimeter-wave WPAN systems." IEEE Journal on Selected Areas in Communications 27.8 (2009).

[4] Ramirez, David, et al. "On opportunistic mmWave networks with blockage." IEEE Journal on Selected Areas in Communications 35.9 (2017): 2137-2147.

[5] Gao, Xinyu, et al. "Fast channel tracking for terahertz beamspace massive MIMO systems." IEEE Transactions on Vehicular Technology 66.7 (2017): 5689-5696.

[6] Zhang, Chuang, Dongning Guo, and Pingyi Fan. "Tracking angles of departure and arrival in a mobile millimeter wave channel." Communications (ICC), 2016 IEEE International Conference on. IEEE, 2016.

[7] Patra, Avishek and Simic, Ljiljana and Mahonen, Petri. "Smart mm-wave beam steering algorithm for fast link re-establishment under node mobility in 60 GHz indoor WLANs." Proceedings of the 13th ACM International Symposium on Mobility Management and Wireless Access. ACM, 2015.

[8] Tsang, Y. Ming, and Ada SY Poon. "Detecting human blockage and device movement in mmWave communication system." Global Telecommunications Conference (GLOBECOM 2011), 2011 IEEE. IEEE, 2011.

[9] Shokri-Ghadikolaei, Hossein, Lazaros Gkatzikis, and Carlo Fischione. "Beam-searching and transmission scheduling in millimeter wave communications." Communications (ICC), 2015 IEEE International Conference on. IEEE, 2015.

[10] Boutilier, Craig, and Tyler Lu. "Budget Allocation using Weakly Coupled, Constrained Markov Decision Processes." UAI. 2016.

[11] Bertsimas, Dimitris and Nino-Mora, Jose. "Restless bandits, linear programming relaxations, and a primal-dual index heuristic." Operations Research 48.1 (2000): 80-90.

[12] Feyzabadi, Seyedshams. Robot Planning with Constrained Markov Decision Processes. Diss. UNIVERSITY OF CALIFORNIA, MERCED, 2017.

[13] Altman, Eitan. Constrained Markov decision processes. Vol. 7. CRC Press, 1999.

[14] Krishnamurthy, Vikram. Partially Observed Markov Decision Processes. Cambridge University Press, 2016.

[15] Miller, Kenneth S. "On the inverse of the sum of matrices." Mathematics magazine 54.2 (1981): 67-72.
[16] Fairbrother, Jamie, and Adam N. Letchford. "Projection results for the k-partition problem." Discrete Optimization 26 (2017): 97-111.
[17] Hassan, Rana A., and Nicolo Michelusi. "Multi-user Beam-Alignment for Millimeter-Wave Networks." arXiv preprint arXiv:1802.06148 (2018).
[18] Seo, Junyeong, et al. "Training beam sequence design for millimeter-wave MIMO systems: A POMDP framework." submitted for publication (2015).

APPENDIX

A. Proof of Theorem 1

Proof: In this proof, we use $1 \times N$ vector notation to present states and actions of multiple users. When S is a set of indices, $S \subseteq \{1,2,\cdots,N\} = \mathcal{N}$, we let $\mathbf{1}_S$ denote a $1 \times N$ vector whose components on S are 1 and 0 elsewhere, e.g. if $N = 5$ and $S = \{2,3\}$, then $\mathbf{1}_S = (0,1,1,0,0)$. $\mathbf{1}_S$ multiplied by a constant c is $c\mathbf{1}_S$, e.g., $L\mathbf{1}_\mathcal{N} = \{L,L,\cdots,L\}$. Similarly, for some $\mathbf{x} \in \mathcal{X}$, we let $\mathbf{x}_S$ denote $\mathbf{x} \circ \mathbf{1}_S$, where $\circ$ is entry-wise product such that $(\mathbf{x}_S)_j = x_j$ if $j \in S$ and 0 if $j \notin S$. Let $I_a = \{i : a_i = 1\}$. We reformulate the CMDP as a parameterized unconstrained MDP. For each Lagrangian multiplier $\lambda > 0$, define the instantaneous Lagrangian cost by

$$c(x,a;\lambda) = c(x,a) - \sum_{i=1}^{N} \lambda_i r_i(x,a). \tag{15}$$

We note that for any fixed $\lambda > 0$, $c(x,a;\lambda)$ is increasing in x since it is the sum of increasing functions. The Lagrangian average cost for a policy u is then defined as follows:

$$J(u;\lambda) = \lim_{T\to\infty} \frac{1}{T} \mathbb{E}^u \sum_{t=1}^{T} c(x^t, a^t : \lambda). \tag{16}$$

The corresponding unconstrained MDP is to minimize the above Lagrangian average cost:

$$V = \inf_u J(u;\lambda) \quad u^*_\lambda = \arg\inf_u J(u;\lambda). \tag{17}$$

The proof proceeds in two steps.

STEP 1 (Pure policy for a given λ): Bellman Equation with cost function (15) for a discounted cost MDP is as follows.

$$V_\alpha(x) = \min_{a\in\mathcal{A}} \{c(x,a;\lambda) + \alpha \sum_{y\in\mathcal{X}} P^a_{x,y} V_\alpha(y)\} = \min_{a\in\mathcal{A}} Q_\alpha(x,a).$$

where $0 < \alpha < 1$ is the discount factor. This can be computed by the recursion,

$$V^{t+1}_\alpha = \min_{a\in\mathcal{A}} Q^{t+1}_\alpha(x,a),$$

where $Q^{t+1}_\alpha(x,a) = c(x,a;\lambda) + \alpha \sum_{x\in\mathcal{X}} P^a_{x,y} V^t_\alpha(y)$. We used the Bellman equation for a discounted cost MDP since an average cost MDP inherits the properties of a discounted cost MDP [14]. From now on, we omit subscript α. we first show that for all $\mathbf{a} \in \mathcal{A}$ the following holds for all $I \subseteq I_a$, $\mathbf{b} \neq \mathbf{a}$.

$$\begin{aligned} &Q(L\mathbf{1}_\mathcal{N},\mathbf{a}) - Q(L\mathbf{1}_\mathcal{N} - k\mathbf{1}_I,\mathbf{b}) \\ &\le Q(L\mathbf{1} - k\mathbf{1}_I,\mathbf{a}) - Q(L\mathbf{1}_\mathcal{N} - k\mathbf{1}_I,\mathbf{b}), 1 \le k \le L-1. \end{aligned} \tag{18}$$

(18) implies that if action $\mathbf{a}$ is optimal at state $L\mathbf{1}_\mathcal{N}$, so is it at states $L\mathbf{1}_\mathcal{N} - k\mathbf{1}_I$, for all $I \subseteq I_\mathbf{a}$ and $k = 1,2,\cdots,L-1$.

$$\begin{aligned} &Q(L\mathbf{1}_\mathcal{N},\mathbf{b}) - Q(L\mathbf{1}_\mathcal{N} - k\mathbf{1}_I,\mathbf{b}) \le Q(L\mathbf{1}_\mathcal{N},\mathbf{a}) - Q(L\mathbf{1}_\mathcal{N} - k\mathbf{1}_I,\mathbf{a}) \\ &\Leftrightarrow \sum_{i\in I\cap I_b} [c_i(L,1;\lambda) - c_i(L-k,1;\lambda)] + \sum_{J\subseteq I_b} p^{|I_b|-|J|} q^{|J|} \\ &\quad [V(L\mathbf{1}_\mathcal{N} - (L-1)\mathbf{1}_J) - V(L\mathbf{1}_\mathcal{N} - (L-1)\mathbf{1}_J - (k+1)\mathbf{1}_{I\cap\bar{I}_b})] \\ &\le \sum_{i\in I\cap I_b} [c_i(L,1;\lambda_i) - c_i(L-k,1;\lambda_i)] \\ &\quad + \sum_{i\in I\cap\bar{I}_b} [c_i(L,1;\lambda_i) - c_i(L-k,1;\lambda_i)]. \end{aligned}$$

It suffices to show that for any $S \subset \mathcal{N}$ and $\mathbf{x} \in \mathcal{X}$,

$$\begin{aligned} &V(L\mathbf{1}_S + \mathbf{x}_{\bar{S}}) - V((L-k)\mathbf{1}_S + \mathbf{x}_{\bar{S}}) \le \\ &\sum_{i\in S} \{c_i(L,1;\lambda_i) - c_i(L-k-1,1;\lambda_i)\}, 2 \le k \le L-2. \end{aligned} \tag{19}$$

Where $\bar{S}$ is a complement set of S. Since V converges for any initial condition, we can select $V^1(\mathbf{x}) = 0$, for all $\mathbf{x} \in \mathcal{X}$. Then (19) holds at $t = 1$. Suppose that (19) holds for $t = s$. That is:

$$\begin{aligned} &V^s(L\mathbf{1}_S + \mathbf{x}_{\bar{S}}) - V^s((L-k)\mathbf{1}_S + \mathbf{x}_{\bar{S}}) \\ &\quad \le \sum_{i\in S} \{c_i(L,1;\lambda_i) - c_i(L-k-1,1;\lambda_i)\}, 2 \le k \le L-2. \end{aligned}$$

Then, there exist some action $\mathbf{a}^{(1)}$ and $\mathbf{a}^{(2)}$ such that

$$\begin{aligned} &V^{s+1}(L\mathbf{1}_S + \mathbf{x}_{\bar{S}}) = \min_{\mathbf{a}\in\mathcal{A}} Q^{s+1}(L\mathbf{1}_S + \mathbf{x}_{\bar{S}},\mathbf{a}) = Q^{s+1}(L\mathbf{1}_S + \mathbf{x}_{\bar{S}},\mathbf{a}^{(1)}) \\ &V^{s+1}((L-k)\mathbf{1}_S + \mathbf{x}_{\bar{S}}) = \min_{\mathbf{a}\in\mathcal{A}} Q^{s+1}((L-k)\mathbf{1}_S + \mathbf{x}_{\bar{S}},\mathbf{a}) \\ &= Q^{s+1}((L-k)\mathbf{1}_S + \mathbf{x}_{\bar{S}},\mathbf{a}^{(2)}). \end{aligned}$$

Then, at $t = s+1$,

$$\begin{aligned} &V^{s+1}(L\mathbf{1}_S + \mathbf{x}_{\bar{S}}) - V^{s+1}(L\mathbf{1}_S + \mathbf{x}_{\bar{S}}) \\ &= \underbrace{Q^{s+1}(L1_S + \mathbf{x}_{\bar{S}},\mathbf{a}^{(1)}) - Q^{s+1}(L\mathbf{1}_S + \mathbf{x}_{\bar{S}},\mathbf{a}^{(2)})}_{\le 0 \text{ by optimality}} \\ &\quad + Q^{s+1}(L\mathbf{1}_S + \mathbf{x}_{\bar{S}},\mathbf{a}^{(2)}) - Q^{s+1}((L-k)\mathbf{1}_S + \mathbf{x}_{\bar{S}},\mathbf{a}^{(2)}) \\ &\le \sum_{i\in S\cap S_0} [c_i(L,1;\lambda_i) - c_i(L-k,1;\lambda_i)] + \sum_{J\subseteq S_0} p^{|S_0|-|J|} q^{|J|} \\ &\quad [V^s(L\mathbf{1}_\mathcal{N} - (L-1)\mathbf{1}_J) - V^s(L\mathbf{1}_\mathcal{N} - (L-1)\mathbf{1}_J - (k+1)\mathbf{1}_{S\cap\bar{S}_0})] \\ &\le \sum_{i\in S} [c_i(L,1;\lambda_i) - c_i(L-k,1;\lambda_i)] \qquad (20) \\ &\le \sum_{i\in S} [c_i(L,1;\lambda_i) - c_i(L-k-1,1;\lambda_i)], \qquad (21) \end{aligned}$$

where $S_0 = I_{a^{(2)}} = \{i : a^{(2)}_i = 1\}$. (20) follows from the assumption at $t = s$, and (21) follows since $c_i(x,1;\lambda)$ is increasing in x for $\lambda > 0$.

STEP 2 (Randomized mixture of multiple policies): Now, let $\Lambda_\mathbf{a} = \{\lambda : Q(L\mathbf{1}_\mathcal{N},\mathbf{a}) - Q(L\mathbf{1}_\mathcal{N},\mathbf{b}) < 0\}$, for all $\mathbf{b} \neq \mathbf{a}$, and let $\mathcal{X}_\mathbf{a} = \{x : p^\mathbf{a}_{L\mathbf{1}_\mathcal{N},x} > 0\}$ be the one-step reachable states from $L\mathbf{1}_\mathcal{N}$ by taking the action $\mathbf{a}$. Then by (18), for any $x \in \mathcal{X}_\mathbf{a}$, $Q(x,\mathbf{a}) < Q(x,\mathbf{b})$, for all $\mathbf{b} \neq \mathbf{a}$. Also, $\{y : p^\mathbf{a}_{x,y} > 0, \forall x \in \mathcal{X}_\mathbf{a}\} \subseteq \mathcal{X}_\mathbf{a}$. So, $v(x,a) = 0$ for all $(x,a) \notin \mathcal{G}_\mathbf{a}$, where $\mathcal{G}_\mathbf{a} = \{(x,\mathbf{a}) : x \in \mathcal{X}_\mathbf{a}\}$. Since $\cup_{\mathbf{a}\in\mathcal{A}}\Lambda_\mathbf{a} = \{\lambda : \lambda > 0\}$ and $\cup_{\mathbf{a}\in A}\mathcal{G}_\mathbf{a} = \mathcal{G}$, this completes the proof. ■