

# Understanding the Impact of Limited Channel State Information on Massive MIMO Network Performances

Paper ID: 1570243333

## ABSTRACT

In recent years, there have been significant efforts on the research and development of Massive MIMO (M-MIMO) technologies at the physical layer. So far, however, the understanding of how M-MIMO could affect the performance of network control and optimization algorithms remains rather limited. In this paper, we focus on analyzing the performance of the queue-length-based joint congestion control and scheduling framework (QCS) over M-MIMO cellular networks with limited channel state information (CSI). Our contributions in this paper are two-fold: i) We characterize the scaling performance of the queue-lengths and show that there exists a phase transitioning phenomenon in the steady-state queue-length deviation respect to the CSI quality (reflected in the number of bits  $B$  that represent CSI); and ii) We characterize the congestion control rate scaling performance and show that there also exists a phase transitioning phenomenon in steady-state congestion control rate deviation respect to the CSI quality. Collectively, the findings in this paper advance our understanding of the trade-offs between delay, throughput, and the accuracy/complexity of CSI acquisition in M-MIMO cellular network systems.

## 1. INTRODUCTION

To allow 5G wireless networks to support multi-Gigabit per second data rates, there have been significant recent efforts on the research and development of massive multiple-input multiple-output systems, or simply Massive MIMO (M-MIMO). In contrast to conventional multi-antenna technologies, the number of antennas in M-MIMO is on the order of hundreds or even thousands. To date, various promising theoretical results on M-MIMO capacity gain and transmit power efficiency have been established (see, e.g., [1–3] for comprehensive overviews). Also, some lab-scale M-MIMO prototypes have been built and favorable field test results have been reported (e.g., [4, 5]). However, in spite of all of this progress, the existing research efforts on M-MIMO are mostly concerned about problems at the physical layer

or signal processing aspects. The understanding of how M-MIMO could affect the performance of network control, scheduling, and resource allocation algorithms remains limited in the literature. In this paper, our goal is to fill this gap by conducting an in-depth theoretical study on the interactions between M-MIMO physical layer and network control and optimization algorithms at higher layers, as well as their impacts on queueing delay and throughput performances.

To this end, in this paper, we focus on the performance analysis of the celebrated queue-length-based congestion control and scheduling framework (QCS) (see, e.g., [6–9], and [10] for a survey) in M-MIMO-based cellular systems, where the M-MIMO data transmissions can rely only on *limited channel state information* (CSI). The fundamental rationale of our work is that, as noted by many researchers [1, 2], CSI acquisition has become one of the most fundamental limiting factors in the design of M-MIMO-based cellular systems. Generally speaking, to leverage the MIMO spatial multiplexing benefits, the transmitter must obtain CSI to perform spatial beamforming so that independent data streams can be simultaneously transmitted. In conventional MIMO-based networks, such CSI is usually learned at each mobile station based on pilot symbols and fed back to the base station (BS). However, due to the constraints on feedback channel capacity and channel coherence time, this traditional CSI feedback approach scales poorly with the increase of antennas in M-MIMO. An alternative CSI acquisition strategy is to let the system operate in time-division duplexing mode and, based on channel reciprocity, use the uplink CSI measured at the BS for downlink transmissions. However, as observed in [1, 11], the channel reciprocity assumption may not hold in practice due to the magnetic properties of the channel environment and transceiver hardware chains.

In this paper, we accept the reality that CSI inaccuracy is unavoidable and we do not require full CSI at the M-MIMO BS. Instead, we assume that the CSI at the BS is limited and accurate only to a certain degree. Such limited CSI can be obtained by a small amount of feedback from each mobile device using a limited number of bits, say  $B$ , to approximately represent its channel instantiation. Alternatively, the BS could use  $B$  bits to approximately represent the downlink CSI based on the channel reciprocity assumption. In such cases, one interesting question naturally arises: *How does the  $B$ -bit limited CSI affect the performance of the QCS framework?* In particular, it is well-known that the QCS framework is throughput-optimal under full CSI and achieves an  $[O(1/K), O(K)]$  utility-delay trade-off, where  $K > 0$  is a system parameter [8]. Also, the average queue-length devia-

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

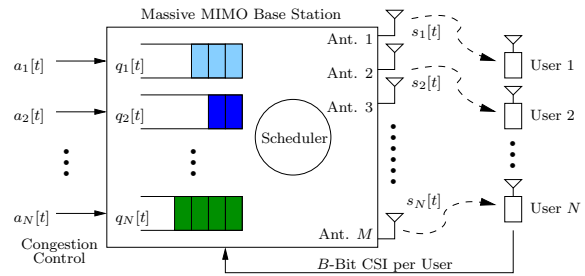
tion and the congestion control rate optimality gap scale as  $O(\sqrt{K})$  [12] and  $O(1/\sqrt{K})$  [7], respectively. However, when the QCS framework is adopted in M-MIMO cellular networks with limited CSI, it begs the following question: *Will the same utility and delay performance scaling laws continue to hold?* As will be seen later, due to the complex cross-layer interactions (e.g., precoder design, choice of channel quantization codebook, power allocations, etc.) in M-MIMO cellular systems, answering this question is challenging.

The main contribution of this paper is that we theoretically characterize the queueing delay and network utility-optimality performance of the QCS framework in M-MIMO cellular networks with  $B$ -bit limited CSI. Our main results and technical contributions are as follows:

- We show that the queues in the network remain stable under QCS for any given  $B$ -bit limited CSI scheme; and the steady-state average queue-lengths still follow an  $O(K)$  linear scaling. However, the slope (i.e., the hidden constant in Big-O) is affected by  $B$ : The larger the value of  $B$  (more accurate CSI), the more gradual the slope becomes. Moreover, the steady-state queue-length deviation from the mean exhibits a phase-transitioning phenomenon: There exists a critical value  $B_{cr}$  such that: i) For all  $0 < B < B_{cr}$ , the steady-state queue-length deviation is bounded by  $O(D_{(B)}K)$ , where  $D_{(B)} > 0$  is a quantity that depends on the specific channel quantization codebook design; and ii) For all  $B \geq B_{cr}$ , the steady-state queue-length deviation scales as  $O(\sqrt{K})$ , i.e., recovering the same scaling law under full CSI.
- For any given  $B$ -bit limited CSI scheme, we show that the steady-state average congestion control rates under the QCS framework increase as  $B$  increases. Interestingly, the same phase transitioning phenomenon also happens in the congestion control rates in the following sense: There exists the same critical value  $B_{cr}$  such that: i) For all  $0 < B < B_{cr}$ , the steady-state congestion control rate deviation scales as  $O(D_{(B)})$  and independent of  $K$ ; and ii) For all  $B \geq B_{cr}$ , the steady-state congestion control rate deviation scales as  $O(1/\sqrt{K})$ , also recovering the same scaling law under the full CSI.
- Collectively, all queue-length and congestion control rate scaling results and their phase transitioning effects advance our understanding of the trade-offs between delay, throughput, and the accuracy/complexity of CSI acquisition in M-MIMO networks. Also, our results suggest that delay and throughput scalings could potentially be employed as useful proxies to control CSI quality and acquisition complexity in M-MIMO networks. More importantly, our work establishes a unifying theoretical framework as well as design guidelines in practice that enable the development of effective CSI quantization schemes for M-MIMO cellular networks.

The remainder of this paper is organized as follows: In Section 2, we introduce network model and the problem formulation. In Section 3, we introduce the queue-length-based congestion control and scheduling framework and present the main results of this work. Section 4 presents the numerical results and Section 5 concludes this paper.

## 2. NETWORK MODEL AND PROBLEM FORMULATION



**Figure 1: A Massive MIMO cellular downlink with  $M$  antennas and  $N$  users, with  $M \gg N$ .**

In this section, we first present the M-MIMO downlink model and the preliminaries of matched-filter beamforming, based on which we introduce the notion of limited CSI in M-MIMO. Next, we will discuss the joint congestion control and scheduling as well as the associated queueing model that operate over the M-MIMO physical layer. Finally, we will introduce the problem formulation.

**Notation:** We use boldface to denote matrices/vectors. We let  $\mathbf{A}^T$  and  $\mathbf{A}^\dagger$  denote the transpose and conjugate transpose of  $\mathbf{A}$ , respectively. We let  $\mathbf{v}_1 \geq \mathbf{v}_2$  denote entry-wise inequality between vectors. We let  $v_m$  represent the  $m$ -th entry of vector  $\mathbf{v}$ . We use  $\|\cdot\|$  and  $\|\cdot\|_1$  to denote  $L^2$ - and  $L^1$ -norms, respectively. We use  $\mathbb{R}$ ,  $\mathbb{C}$ , and  $\mathbb{Z}$  to denote real, complex, and integer spaces, respectively.

**1) Massive MIMO Downlink Channel:** As shown in Figure 1, we consider an M-MIMO cellular downlink system, where the BS has  $M$  antennas and serves  $N$  active single-antenna users. We assume that the system operates under a time-slotted fashion and time is indexed by  $t \in \{0, 1, 2, \dots\}$ . We let  $\mathbf{H}[t] \in \mathbb{C}^{N \times M}$  denote the channel gain matrix in time-slot  $t$  between the BS and the users. We assume independent quasi-static block fading, i.e., each entry in  $\mathbf{H}[t]$  is constant in one time-slot and independently varies in the next time-slot. In M-MIMO cellular networks, it is typical that  $M \gg N$  (e.g.,  $M$  is in hundreds or even thousands, while  $N$  could be well less than tens). Thanks to such excess degrees of freedom at the BS, it is possible for the BS to serve all  $N$  users by simultaneously forming  $N$  spatial beams. Moreover, one important property of M-MIMO channels is that, under favorable propagation conditions, the row vectors of  $\mathbf{H}[t]$  are asymptotically orthogonal as  $M \rightarrow \infty$  [2]. This property enables the use of simple matched-filter (MF) beamforming strategy to approach the MIMO broadcast channels [2]<sup>1</sup>. Thus, in what follows, we will briefly introduce some related preliminaries of MF beamforming for M-MIMO.

**2) Matched-Filter Beamforming:** For the M-MIMO cellular downlink in Figure 1, the received signal of user  $n$  in time-slot  $t$  can be written as:  $y_n[t] = x_n[t] \sqrt{p_n[t]} \mathbf{h}_n^\dagger[t] \mathbf{w}_n[t] + \sum_{j=1, j \neq n}^N x_j[t] \sqrt{p_j[t]} \mathbf{h}_n^\dagger[t] \mathbf{w}_j[t] + v_n[t]$ , where  $\mathbf{h}_n[t] \in \mathbb{C}^M$  is the channel gain vector seen at user  $n$  in time-slot  $t$ , i.e., the  $n$ -th row in  $\mathbf{H}[t]$ ;  $p_n[t]$  is the power allocated to user  $n$  in time-slot  $t$ ;  $x_n[t]$  represents a unit-power data symbol

<sup>1</sup>Information theoretic studies have shown that the dirty paper coding (DPC) is capacity-achieving for MIMO broadcast channels [13]. However, DPC is a nonlinear precoding scheme that is difficult to be implemented in practice due to the complex successive encoding and decoding procedures. Also, it is shown that MF's capacity loss compared to DPC is negligible in the high signal-to-noise ratio regime [2].

intended for user  $n$  in time-slot  $t$ ;  $\mathbf{w}_n[t] \in \mathbb{C}^N$  is a unit-norm linear precoding vector for user  $n$  in time-slot  $t$ ; and  $v_n[t]$  is the white complex Gaussian noise at user  $n$  in time-slot  $t$  with power  $N_0$ . Under MF beamforming, we simply let  $\mathbf{w}_n[t] = \mathbf{h}_n[t]$ , i.e., the  $n$ -th row in  $\mathbf{H}[t]$ . In this setting, the achievable rate under MF beamforming can be computed as:

$$r_n[t] = \log_2 \left( 1 + \frac{p_n[t] |\mathbf{h}_n^\dagger[t] \mathbf{w}_n[t]|^2}{N_0 + \sum_{j=1, j \neq n}^N p_j[t] |\mathbf{h}_n^\dagger[t] \mathbf{w}_j[t]|^2} \right) \stackrel{(a)}{\approx} \log_2 \left( 1 + \frac{p_n[t]}{N_0} \|\mathbf{h}_n[t]\|^2 \right), \quad (1)$$

where (a) follows from the fact that the rows of  $\mathbf{H}[t]$  in M-MIMO channels are nearly orthogonal, i.e.,  $\mathbf{h}_n^\dagger[t] \mathbf{w}_j[t] = \mathbf{h}_n^\dagger[t] \mathbf{h}_j[t] \approx 0$ ,  $\forall n \neq j$ , when  $M$  is sufficiently large.

We assume that the channel fading can be characterized by a total of  $F$  states  $\mathbf{H}^{(1)}, \dots, \mathbf{H}^{(F)}$ , where each  $\mathbf{H}^{(f)} \in \mathbb{C}^{N \times M}$ ,  $f = 1, \dots, F$ , corresponds to the channel qualities between the  $M$  antennas and  $N$  users in state  $f$ . For each  $\mathbf{H}^{(f)}$ , we let  $\mathcal{C}_{\mathbf{H}^{(f)}}$  denote the achievable MF rate region, which is the convex hull of all achievable MF rate vectors in state  $f$  under all feasible power allocations:

$$\mathcal{C}_{\mathbf{H}^{(f)}} \triangleq \text{CH} \left\{ \mathbf{r}^{(f)}, 1 \leq n \leq N \left| \begin{array}{l} r_n^{(f)} = \log_2 \left( 1 + \frac{p_n}{N_0} \|\mathbf{h}_n^{(f)}\|^2 \right) \\ p_n \geq 0, \forall n, \sum_{n=1}^N p_n \leq P_{\max} \end{array} \right. \right\},$$

where  $\text{CH}\{\cdot\}$  represents the convex hull operation,  $p_n$  denotes the power allocated of user  $n$ , and  $P_{\max}$  denotes the maximum transmission power at the BS. Clearly, due to the maximum power constraint, there exists a  $r_{\max} < \infty$  such that  $r_n^{(f)} \leq r_{\max}$ ,  $\forall n, f$ . We let  $\mathbf{r}^{(f)} = [r_1^{(f)}, \dots, r_N^{(f)}]^\top \in \mathbb{R}^N$  denote the feasible MF rates in state  $f$ . We let  $\pi_f \triangleq \Pr\{\mathbf{H}[t] = \mathbf{H}^{(f)}\}$  be the stationary distribution of the channel state process being in state  $f$ . We let  $\bar{\mathcal{C}}$  denote the mean MF achievable rate region, which can be written as:

$$\bar{\mathcal{C}} \triangleq \left\{ \mathbf{r} \left| \mathbf{r} = \sum_{f=1}^F \pi_f \mathbf{r}^{(f)}, \forall \mathbf{r}^{(f)} \in \mathcal{C}_{\mathbf{H}^{(f)}} \right. \right\}. \quad (2)$$

We note that, in this paper, neither the channel state statistics nor  $\bar{\mathcal{C}}$  is assumed to be known at the BS under the QCS algorithm, which will be introduced in Section 3.

**3) B-Bit Limited CSI:** The MF beamforming precoder  $\mathbf{w}_n[t] = \mathbf{h}_n[t]$ ,  $\forall n, t$ , means that the BS requires full CSI  $\mathbf{H}[t]$ ,  $\forall t$ . However, as mentioned in Section 1, it becomes increasingly difficult and expensive to acquire full CSI as  $M$  gets large in M-MIMO systems. One way to address this challenge is to use limited CSI by quantizing the channel (e.g., [14–18]). As shown in Figure 1, such limited CSI can be obtained from a small amount of feedback by each user using a limited number of bits  $B$  to represent a quantized version of the channel state. Alternatively, in time-division multiplexing mode, Figure 1 represents that the BS uses  $B$  bits to quantize the uplink CSI, which will be used for downlink transmissions. In both cases, the value of  $B$  can be viewed as a means to balance the trade-off between CSI accuracy and CSI acquisition time/complexity costs. The  $B$ -bit limited CSI for each user  $n$  is based on a vector quantization codebook  $\mathcal{B}_n \triangleq \{\mathbf{c}_n^1, \dots, \mathbf{c}_n^{2^B}\}$ , where  $\mathbf{c}_n^i \in \mathbb{C}^M$ ,  $i = 1, \dots, 2^B$ , denotes a codeword. With the CSI  $\mathbf{h}_n[t]$  in time  $t$ , a codeword for each user  $n$  is chosen by:

$$i_n^*[t] = \arg \max_{j \in \{1, \dots, 2^B\}} |\mathbf{h}_n^\dagger[t] \mathbf{c}_n^j| = \arg \min_{j \in \{1, \dots, 2^B\}} \sin^2(\angle(\mathbf{h}_n[t], \mathbf{c}_n^j)), \quad (3)$$

where  $i_n^*[t]$  denotes the index of the chosen codeword. We let  $\hat{\mathbf{H}}[t] \in \mathbb{C}^{N \times M}$  denote the corresponding channel gain matrix in time-slot  $t$  by aggregating all codewords  $i_n^*[t]$ ,  $\forall n$ . Then, by treating  $\hat{\mathbf{H}}[t]$  as if it is the accurate CSI, the BS performs MF beamforming to construct  $N$  spatial channels. However, due to the inaccuracy of  $\hat{\mathbf{H}}[t]$ , multi-user interference may become non-negligible. Clearly, the impact of multi-user interference depends heavily on the codebook size  $2^B$  and the design of the quantization codebook.

Let  $\mathcal{C}_{\mathbf{H}[t]|\hat{\mathbf{H}}[t]}$  denote the actual MF rate region achieved under  $\mathbf{H}[t]$  based on the belief that  $\hat{\mathbf{H}}[t]$  is the accurate CSI. Also, let  $\hat{\mathbf{H}}_1[t]$  and  $\hat{\mathbf{H}}_2[t]$  represent two estimated CSI values obtained by using  $B_1$  and  $B_2$  bits, respectively. Further, we let  $\mathcal{C}_{\mathbf{H}[t]}$  denote the original MF achievable rate region under full CSI  $\mathbf{H}[t]$ . Then, one can show the following inclusion result of the MF achievable rate regions under limited CSI in M-MIMO networks (the proof is relegated to Appendix A):

**LEMMA 1 (MF RATE REGION INCLUSION).** *If  $B_1 \leq B_2$ , then there exists a CSI quantization scheme under which  $\mathcal{C}_{\mathbf{H}[t]|\hat{\mathbf{H}}_1[t]} \subseteq \mathcal{C}_{\mathbf{H}[t]|\hat{\mathbf{H}}_2[t]}$ . Further,  $\mathcal{C}_{\mathbf{H}[t]|\hat{\mathbf{H}}[t]} \rightarrow \mathcal{C}_{\mathbf{H}[t]}$  as  $B \rightarrow \infty$ .*

**4) Queueing Model:** As illustrated in Figure 1, the BS maintains a separate queue for each user. Let  $a_n[t]$  denote the number of packets injected into queue  $n$  in time-slot  $t$ . As shown in Figure 1, the arrival processes  $\{a_n[t]\}$ ,  $\forall n$ , are controlled by a congestion controller. Also, we assume that there exists a finite constant  $A^{\max}$  such that  $a_n[t] \leq A^{\max}$ ,  $\forall n, t$ . Let  $\mathbf{s}_B[t] \triangleq [s_{B,1}[t], \dots, s_{B,N}[t]]^\top$  denote the scheduled service rate vector in time-slot  $t$  based on the belief that the current  $B$ -bit limited CSI is accurate (the scheduling algorithm that determines  $\mathbf{s}_B[t]$  will be presented in Section 3). Then, the queue-length of each user evolves as follows:  $q_n[t+1] = (q_n[t] - s_{B,n}[t] + a_n[t])^+$ ,  $\forall n$ , where  $(\cdot)^+ \triangleq \max(0, \cdot)$ . Let  $\mathbf{q}[t] = [q_1[t], \dots, q_N[t]]^\top$ . In this paper, we adopt the following notion of queue-stability (same as in [7, 8]): we say that a network is *stable* if the steady-state total queue-length is finite, i.e.,

$$\limsup_{t \rightarrow \infty} \mathbb{E} \{ \|\mathbf{q}[t]\|_1 \} < \infty. \quad (4)$$

**5) Problem Statement:** Let  $\bar{a}_n \triangleq \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} a_n[t]$  denote the average controlled arrival rate of user  $n$ . Each user  $n$  is associated with a utility function  $U_n(\bar{a}_n)$ , which represents the utility gained by user  $n$  when data is injected at rate  $\bar{a}_n$ . We assume that  $U_n(\cdot)$ ,  $\forall n$ , is strictly concave, monotonically increasing, and twice continuously differentiable. We assume that  $U_n(\cdot)$  satisfies the following strong concavity condition: there exist constants  $0 < \phi < \Phi < \infty$  such that  $\phi \leq -U_n''(a_n) \leq \Phi$ ,  $\forall a_n \in [0, A^{\max}]$ , where  $A^{\max}$  is the maximum arrival rates for burst control. For example,  $\log(a_n + \epsilon)$  is strongly concave for  $\epsilon > 0$ . Our goal is to maximize  $\sum_{n=1}^N U_n(\bar{a}_n)$ , subject to the MF beamforming rate region  $\mathcal{C}_{\mathbf{H}[t]|\hat{\mathbf{H}}[t]}$  due to limited CSI in each time-slot and the queue-stability constraint. Putting together the models presented above yields the following joint congestion control and scheduling (JCCS) optimization problem:

$$\begin{aligned} \text{JCCS: Maximize} \quad & \sum_{n=1}^N U_n(\bar{a}_n) \\ \text{subject to} \quad & \text{Queue-length stability constraint in (4),} \\ & s_{B,n}[t] \in \mathcal{C}_{\mathbf{H}[t]|\hat{\mathbf{H}}[t]}, a_n[t] \in [0, A^{\max}] \forall n, t. \end{aligned}$$

Note that, when perfect CSI is available ( $B \rightarrow \infty$ ), the well-known QCS algorithmic framework [6–9] optimally solves Problem JCCR in the following sense: The  $\bar{\mathbf{a}} \triangleq [\bar{a}_1, \dots, \bar{a}_N]^\top$  obtained from the QCS algorithm achieves a utility optimality gap  $O(1/K)$  at the expense of queue-length scaling as  $O(K)$ , where  $K > 0$  is a system parameter. Hence, the utility optimality gap can be made arbitrarily small by increasing  $K$  asymptotically (implying an asymptotically large queueing delay). However, in M-MIMO cellular networks, it is not clear whether or not the QCS algorithmic framework will be optimal based on  $B$ -bit limited CSI. This constitutes the main discussions in the next section.

### 3. PERFORMANCE ANALYSIS OF THE QCS ALGORITHM WITH LIMITED CSI

In this section, we first present a variant of the QCS algorithm adapted for M-MIMO with  $B$ -bit limited CSI in Section 3.1. Then, we will examine a deterministic problem related to Problem JCCS in Section 3.2 to facilitate our discussions. The main theoretical results and their proofs will be presented in Sections 3.3 and 3.4, respectively.

#### 3.1 The QCS Algorithm with Limited CSI

**Algorithm 1:** Queue-Length-Based Congestion Control and Scheduling for M-MIMO Cellular Networks with  $B$ -Bit CSI.

**Initialization:**

1. Select an appropriate  $K > 0$ . Let  $q_n[0] = 0, \forall n$ .

**Main Loop:**

2. *Queue-Length-Based MaxWeight Scheduler:* In time-slot  $t \geq 1$ , given the queue-length vector  $\mathbf{q}[t] \triangleq [q_1[t], \dots, q_N[t]]^\top$  and the  $B$ -bit estimated CSI  $\widehat{\mathbf{H}}[t]$ , the scheduler chooses a power allocation  $\mathbf{p}[t] = [p_1[t], \dots, p_N[t]]^\top$  such that the (believed) MF achievable rates  $\mathbf{r}[t] = \arg \max_{\mathbf{x} \in \mathcal{C}_{\widehat{\mathbf{H}}[t]}} (\mathbf{q}[t])^\top \mathbf{x}$ . As a result, the actual MF achievable service rates  $s_{B,n}[t], \forall n$ , under  $\mathbf{p}[t]$  are:

$$s_{B,n}[t] = \log_2 \left( 1 + \frac{p_n[t] |\mathbf{h}_n^\dagger[t] \widehat{\mathbf{h}}_n[t]|^2}{N_0 + \sum_{j=1, j \neq n}^N p_j[t] |\mathbf{h}_n^\dagger[t] \widehat{\mathbf{h}}_j[t]|^2} \right). \quad (5)$$

3. *Congestion Controller:* Given the queue-length vector  $\mathbf{q}[t] \triangleq [q_1[t], \dots, q_N[t]]^\top$ , the congestion controller chooses data inject rates  $a_n[t], \forall n$ , which are integer-valued random variables satisfying:

$$\mathbb{E}\{a_n[t] | q_n[t]\} = \min \left\{ U_n'^{-1} \left( \frac{q_n[t]}{K} \right), A^{\max} \right\}, \quad (6)$$

$$\mathbb{E}\{a_n^2[t] | q_n[t]\} \leq A_2^{\max} < \infty, \quad \forall q_n[t], \quad (7)$$

where  $U_n'^{-1}(\cdot)$  represents the inverse function of first-order derivative of  $U_n(\cdot)$ . In (6) and (7),  $A^{\max}$  and  $A_2^{\max}$  are positive constants.

4. *Queue-Length Updates:* Update the queue-lengths as:

$$q_n[t+1] = (q_n[t] - s_{B,n}[t] + a_n[t])^+, \quad \forall n. \quad (8)$$

Let  $t = t + 1$ . Go to Step 2 and repeat the scheduling and congestion control processes.

Some remarks on Algorithm 1 are in order: Different from traditional cellular networks, the M-MIMO BS has sufficient spatial degrees of freedom to serve all users simultaneously thanks to the large number of antennas and

transceiver chains. Hence, the user selection challenge (potentially NP-hard) in the scheduling component in traditional QCS algorithms does not exist in Algorithm 1. Instead, now the main challenge is reflected in the limited and inaccurate CSI, which leads to suboptimal service rates in (5). This incurs service rate losses compared to the full CSI case, where the MaxWeight scheduler is of the form  $\mathbf{s}[t] = \arg \max_{\mathbf{x} \in \mathcal{C}_{\mathbf{H}[t]}} (\mathbf{q}[t])^\top \mathbf{x}$ . In what follows, we will focus on the impact of this inaccurate MaxWeight scheduling solution due to the  $B$ -bit limited CSI.

#### 3.2 A Deterministic Problem

To facilitate the presentation of our theoretical results in Section 3.3, we first introduce a  $K$ -parameterized *deterministic* problem, where we assume that the channel state process is not random and fixed at its mean level. That is, the mean achievable rate region  $\bar{\mathcal{C}}^B \triangleq \{\mathbf{r} | \mathbf{r} = \sum_{f=1}^F \pi_f \hat{\mathbf{r}}^{(f)}, \forall \mathbf{r}^{(f)} \in \mathcal{C}_{\mathbf{H}^{(f)} | \widehat{\mathbf{H}}_B^{(f)}}\}$ , where  $\mathcal{C}_{\mathbf{H}^{(f)} | \widehat{\mathbf{H}}_B^{(f)}}$  represents the actual MF rate region achieved under  $\mathbf{H}^{(f)}$  based on the belief that the CSI is  $\widehat{\mathbf{H}}_B^{(f)}$ , i.e., the  $B$ -bit quantized CSI for state  $f$ . Also, the congestion control and scheduling variables are time-invariant, which are denoted as  $a_n$  and  $s_{B,n}, \forall n$ , respectively. Then, the deterministic congestion control and scheduling problem ( $K$ -DJCCS) can be written as:

$$\begin{aligned} K\text{-DJCCS: Maximize} \quad & K \sum_{n=1}^N U_n(a_n) \\ \text{subject to} \quad & a_n - s_{B,n} \leq 0, \quad \forall n, \\ & s_{B,n} \in \bar{\mathcal{C}}^B, \quad a_n \in [0, a^{\max}], \quad \forall n. \end{aligned}$$

Since Problem  $K$ -DJCCS is strictly convex, an optimal solution exists and is unique. Further, we associate dual variables  $q_{B,n} \geq 0, \forall n$  with the constraints  $a_n - s_{B,n} \leq 0, \forall n$ , to obtain the Lagrangian as follows:

$$\Theta_K(\mathbf{q}_B) \triangleq \max_{\mathbf{a}, \mathbf{s}_B \in \bar{\mathcal{C}}^B} \left\{ K \sum_{n=1}^N U_n(a_n) + \sum_{n=1}^N q_{B,n} (s_{B,n} - a_n) \right\}, \quad (9)$$

where the vector  $\mathbf{q}_B \triangleq [q_{B,1}, \dots, q_{B,N}]^\top \in \mathbb{R}_+^N$  contains all dual variables. Then, the Lagrangian dual problem of Problem  $K$ -DJCCS can be written as:

$$\begin{aligned} K\text{-LD-JCCS: Minimize} \quad & \Theta_K(\mathbf{q}_B) \\ \text{subject to} \quad & \mathbf{q}_B \in \mathbb{R}_+^N. \end{aligned}$$

It can be verified that Problem  $K$ -DJCCS satisfies the Slater condition [19]. Hence, the optimal value of Problem  $K$ -LD-JCCS is equal to that of Problem  $K$ -DJCCS. Let  $(\mathbf{a}_B^*, \mathbf{s}_B^*)$  and  $\mathbf{q}_{B,(K)}^*$  be a pair of optimal primal and dual solutions. Then,  $\mathbf{q}_{B,(K)}^*$  can be shown to have the following properties:

**LEMMA 2 (OPTIMAL DUAL SOLUTION SCALING).** *For a given  $K$ ,  $\mathbf{q}_{B,(K)}^* = K \mathbf{q}_{B,(1)}^*$ , or equivalently,  $\mathbf{q}_{B,(K)}^*$  scales linearly as  $O(K)$  and the slope is determined by the entries in  $\mathbf{q}_{B,(1)}^*$ . Further,  $\mathbf{q}_{B_1,(1)}^* \geq \mathbf{q}_{B_2,(1)}^*$  if  $B_1 \leq B_2$ .*

Lemma 2 can be proved by examining the Karush-Kuhn-Tucker (KKT) conditions [19] of Problem  $K$ -DJCCS and the proof is relegated to Appendix B. Also, by noting the fact that  $K$  is just a scaling factor in the objective function and  $\mathbf{a}_B^* = \mathbf{s}_B^*$  at optimality (by KKT conditions), we immediately have the following result for  $\mathbf{a}_B^*$ :

**LEMMA 3 (OPTIMAL PRIMAL SOLUTION).** *The optimal congestion control rate  $\mathbf{a}_B^*$  is independent of  $K$  and equal to the optimal service rate  $\mathbf{s}_B^*$  over  $\bar{\mathcal{C}}^B$ .*

### 3.3 Main Results

In this section, we present the main performance analysis results of Algorithm 1. Our first result says that the steady-state queue-lengths  $\mathbf{q}^\infty$  stay in a neighborhood of  $\mathbf{q}_{B,(K)}^*$  (the existence of steady-state will also be proved later). Further, the scaling of the steady-state queue-length deviation from  $\mathbf{q}_{B,(K)}^*$  exhibits a phase-transitioning phenomenon:

**THEOREM 1 (QUEUE-LENGTH PHASE TRANSITIONING).** *For any  $B$ -bit CSI quantization scheme in Algorithm 1 with parameter  $K$ , there exists a critical value  $B_{\text{cr}}$  that is independent of  $K$ , such that the following hold:*

- For all  $0 < B < B_{\text{cr}}$ ,  $\mathbb{E}\{\|\mathbf{q}^\infty - \mathbf{q}_{B,(K)}^*\|\} = O(D_{(B)}K)$ , where the parameter  $D_{(B)} \geq 0$  depends on the quantization codebook design and shrinks as  $B$  increases;
- For all  $B \geq B_{\text{cr}}$ ,  $\mathbb{E}\{\|\mathbf{q}^\infty - \mathbf{q}_{B,(K)}^*\|\} = O(\sqrt{K})$ .

Collectively, Theorem 1 and Lemma 2 describe the steady-state queue-length behaviors. In particular, they show that if  $B \geq B_{\text{cr}}$ , the steady state queue-length deviation is upper bounded by  $O(\sqrt{K})$ , which is small compared to the magnitude of  $\mathbf{q}_{B,(K)}^*$ , which grows linearly as  $O(K)$  and the slope is affected by  $B$ : the larger the value of  $B$ , the more gradual the slope. Note that the scaling of the queue-length deviation for  $B \geq B_{\text{cr}}$  is the same as the classical result under full CSI [12]. This implies an interesting and surprising insight that full CSI is not necessary to induce certain desirable queuing behaviors in M-MIMO cellular networks.

Now, let  $a_{B,n}^\infty \triangleq \mathbb{E}\{\min\{U_n'^{-1}(q_n^\infty/K), a^{\max}\}\}$ ,  $\forall n$ , be the steady-state congestion control rates under some  $B$ -bit CSI quantization and let  $\mathbf{a}_B^\infty \triangleq [a_{B,1}^\infty, \dots, a_{B,N}^\infty]^\top$ . Our second main result is on the scaling of  $\mathbf{a}_B^\infty$ 's deviation from  $\mathbf{a}_B^*$ :

**THEOREM 2 (RATE-CONTROL PHASE TRANSITIONING).** *For any  $B$ -bit CSI quantization scheme in Algorithm 1 with parameter  $K$ , there exists a critical value  $B_{\text{cr}}$  (same as in Theorem 1) such that the following hold:*

- For all  $0 < B < B_{\text{cr}}$ ,  $\|\mathbf{a}_B^\infty - \mathbf{a}_B^*\| = O(D_{(B)})$ , where the parameter  $D_{(B)} \geq 0$  is the same as in Theorem 1;
- For all  $B \geq B_{\text{cr}}$ ,  $\|\mathbf{a}_B^\infty - \mathbf{a}_B^*\| = O(1/\sqrt{K})$ .

Similar to the results in Theorem 1, Theorem 2 combined with Lemma 3 suggest that a phase transitioning phenomenon also exists in  $\mathbf{a}_B^\infty$ : When  $B < B_{\text{cr}}$ , parameter  $K$  becomes ineffective in the control of  $\mathbf{a}_B^\infty$ 's deviation from  $\mathbf{a}_B^*$ . On the other hand, when  $B \geq B_{\text{cr}}$ ,  $\mathbf{a}_B^\infty$ 's deviation from  $\mathbf{a}_B^*$  scales as  $O(1/\sqrt{K})$  and can be made arbitrarily small by increasing  $K$ . Since this  $O(1/\sqrt{K})$  scaling is the same as that under full CSI [7, 8],  $B_{\text{cr}}$  represents the smallest codebook size of the given CSI quantization scheme that recovers the performance control functionality of parameter  $K$ .

### 3.4 Proofs of the Main Theorems

In this subsection, we provide proofs for Theorems 1 and 2. To this end, we first show a positive Harris-recurrence result of the queue-length process, which implies the existence of steady-state and will be useful for proving Theorems 1 and 2 later. Let  $\mathbb{1}_{\mathcal{A}}(\mathbf{x})$  be the indicator function that takes value 1 if  $\mathbf{x} \in \mathcal{A}$  and 0 otherwise. Then, the queue-length positive Harris-recurrence result can be stated as follows:

**THEOREM 3 (QUEUE-LENGTH POSITIVE RECURRENCE).** *Consider a Lyapunov function  $V(\mathbf{q}[t]) \triangleq \frac{1}{2K}\|\mathbf{q}[t] - \mathbf{q}_{B,(K)}^*\|^2$*

*for a given  $K$ . For the scheduler (5) and congestion controller (6)–(7), there exist constants  $\delta, \eta > 0$ , both independent of  $K$ , such that the queue-length process  $\{\mathbf{q}[t]\}_{t=0}^\infty$  satisfies the following conditional mean drift condition:*

$$\begin{aligned} \mathbb{E}\{\Delta V(\mathbf{q}[t]|\mathbf{q}[t])\} &\triangleq \mathbb{E}\{V(\mathbf{q}[t+1]) - V(\mathbf{q}[t])|\mathbf{q}[t]\} \\ &\leq -\frac{\delta}{\Phi K}\|\mathbf{q}[t] - \mathbf{q}_{B,(K)}^*\|\mathbb{1}_{\mathcal{B}^c}(\mathbf{q}[t]) + \eta\mathbb{1}_{\mathcal{B}}(\mathbf{q}[t]), \end{aligned} \quad (10)$$

where  $\mathcal{B} \triangleq \{\mathbf{q} \in \mathbb{Z}_+^N \mid \|\mathbf{q} - \mathbf{q}_{B,(K)}^*\| \leq \beta K\}$  for some constant  $\beta > 0$  and  $\mathcal{B}^c$  denotes the complement of  $\mathcal{B}$  in  $\mathbb{Z}_+^N$ .

We relegate the proof details of Theorem 3 to Appendix C. The inequality in (10) shows that the conditional mean drift is negative when the deviation of the queue-length vector  $\mathbf{q}[t]$  away from  $\mathbf{q}_{B,(K)}^*$  is sufficiently large. Since (10) is exactly the Foster-Lyapunov criterion [20, Proposition 1.5.3],  $\{\mathbf{q}[t]\}_{t=0}^\infty$  is positive recurrent. Thus, a steady-state distribution of queue-lengths exists. We denote the queue-length vector in steady-state as  $\mathbf{q}^\infty$ . With Theorem 3, we are now in a position to prove Theorem 1.

**PROOF OF THEOREM 1.** To prove Theorem 1, we use the following  $\alpha$ -parameterized quadratic Lyapunov function:

$$V_\alpha(\mathbf{q}[t]) = \frac{1}{2K^\alpha}\|\mathbf{q}[t] - \mathbf{q}_{B,(K)}^*\|^2, \quad (11)$$

where the parameter  $\alpha \in \{0, 1\}$  and its value will be specified later. Following similar steps in the proof of Theorem 3 (see Appendix C), we can bound the conditional mean Lyapunov drift as follows:

$$\begin{aligned} &\mathbb{E}\{V_\alpha(\mathbf{q}[t+1]) - V_\alpha(\mathbf{q}[t])|\mathbf{q}[t]\} \\ &\stackrel{(a)}{\leq} \frac{1}{K^\alpha}(\mathbf{q}[t] - \mathbf{q}_{B,(K)}^*)^\top (\mathbb{E}\{\mathbf{a}[t]|\mathbf{q}[t]\} - \mathbf{s}_B^*) + \\ &\quad \frac{1}{K^\alpha}\mathbb{E}\{(\mathbf{q}[t] - \mathbf{q}_{B,(K)}^*)^\top (\mathbf{s}^* - \mathbf{s}_B[t])|\mathbf{q}[t]\} + \frac{D_0}{K^\alpha}, \\ &\stackrel{(b)}{\leq} \frac{1}{K^\alpha}\left[-\frac{1}{\Phi K}\|\mathbf{q}[t] - \mathbf{q}_{B,(K)}^*\|^2 + D_0\right] + \\ &\quad \frac{1}{K^\alpha}\mathbb{E}\{(\mathbf{q}[t] - \mathbf{q}_{B,(K)}^*)^\top (\mathbf{s}_B^* - \mathbf{s}_B[t])|\mathbf{q}[t]\} \\ &\stackrel{(c)}{\leq} \frac{1}{K^\alpha}\left[-\frac{1}{\Phi K}\|\mathbf{q}[t] - \mathbf{q}_{B,(K)}^*\|^2 + D_0\right] + \\ &\quad \frac{1}{K^\alpha}\mathbb{E}\{(\mathbf{q}[t])^\top (\mathbf{s}^* - \mathbf{s}_B[t])|\mathbf{q}[t]\}, \end{aligned} \quad (12)$$

where  $D_0 \triangleq \frac{N}{2}(A_2^{\max} + (s^{\max})^2)$  and  $\mathbf{s}^* \triangleq \lim_{B \rightarrow \infty} \mathbf{s}_B^*$ . In (12), (a) follows from adding and subtracting  $\mathbf{s}_B^*$ ; (b) follows from (37); and (c) follows from  $\mathbf{s}_B^* \leq \mathbf{s}^*$  (by Lemma 1) and the scheduler design, which implies  $(\mathbf{q}_{B,(K)}^*)^\top \mathbf{s}_B[t] \leq (\mathbf{q}_{B,(K)}^*)^\top \mathbf{s}_B^*$ . Next, consider the  $T$ -step conditional mean Lyapunov drift. For any  $\mathbf{q}[0] \geq \mathbf{0}$ , we have that

$$\begin{aligned} \mathbb{E}\{V_\alpha(\mathbf{q}[T])|\mathbf{q}[0]\} - V_\alpha(\mathbf{q}[0]) &= \sum_{t=0}^{T-1} \mathbb{E}\{V(\mathbf{q}[t+1]) - V(\mathbf{q}[t])|\mathbf{q}[0]\} \\ &\stackrel{(a)}{=} \sum_{t=0}^{T-1} \sum_{\mathbf{q} \in \mathbb{Z}_+^N} [\Pr(\mathbf{q}[t] = \mathbf{q}|\mathbf{q}[0])\mathbb{E}\{V_\alpha(\mathbf{q}[t+1]) - V_\alpha(\mathbf{q}[t])|\mathbf{q}[t] = \mathbf{q}\}] \\ &\stackrel{(b)}{\leq} \sum_{t=0}^{T-1} \sum_{\mathbf{q} \in \mathbb{Z}_+^N} \Pr(\mathbf{q}[t] = \mathbf{q}|\mathbf{q}[0]) \left\{ \frac{1}{K^\alpha} \left[ -\frac{1}{\Phi K} \|\mathbf{q}[t] - \mathbf{q}_{B,(K)}^*\|^2 + D_0 \right] \right\} \\ &+ \sum_{t=0}^{T-1} \sum_{\mathbf{q} \in \mathbb{Z}_+^N} \Pr(\mathbf{q}[t] = \mathbf{q}|\mathbf{q}[0]) \left\{ \frac{1}{K^\alpha} \mathbb{E}\{\mathbf{q}^\top (\mathbf{s}^* - \mathbf{s}_B[t])\} \right\}, \end{aligned} \quad (13)$$

where (a) follows from the fact that  $\mathbf{q}[t]$  is a discrete state Markov chain in  $\mathbb{Z}_+^N$  and (b) follows from (12). Note that for any  $\mathbf{q}[t] \in \mathbb{Z}_+^N$ ,  $\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \Pr(\mathbf{q}[t] = \mathbf{q} | \mathbf{q}[0]) = \pi_{\mathbf{q}}^\infty$ , where  $\pi_{\mathbf{q}}^\infty$  denotes the stationary distribution of the Markov chain  $\mathbf{q}[t]$ . Moving  $V(\mathbf{q}[0])$  to the right hand side, dividing both sides by  $T$ , and letting  $T \rightarrow \infty$  yields:

$$0 \leq J + \sum_{\mathbf{q} \in \mathbb{Z}_+^N} \pi_{\mathbf{q}}^\infty (\mathbf{q})^\top (\mathbf{s}^* - \mathbf{s}_B^\infty) = J + \mathbb{E}\{(\mathbf{q}^\infty)^\top (\mathbf{s}^* - \mathbf{s}_B^\infty)\}, \quad (14)$$

where  $J \triangleq \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \sum_{\mathbf{q} \in \mathbb{Z}_+^N} \Pr(\mathbf{q}[t] = \mathbf{q} | \mathbf{q}[0]) \left\{ \frac{1}{K^\alpha} \left[ \frac{1}{\Phi K} \|\mathbf{q}[t] - \mathbf{q}_{B,(K)}^*\|^2 + D_0 \right] \right\}$ ,  $\mathbf{s}_B^\infty \triangleq \arg \max_{\mathbf{x} \in \mathcal{C}_{\mathbf{H}[\infty]} \bar{\mathbf{H}}[\infty]} (\mathbf{q}^\infty)^\top \mathbf{x}$  represents the steady-state service rates with  $B$ -bit CSI.

Next, consider the term  $\mathbb{E}\{(\mathbf{q}^\infty)^\top (\mathbf{s}^* - \mathbf{s}_B^\infty)\}$  in (14). For any given realization of  $\mathbf{q}^\infty$  in the steady-state, from the design of the MaxWeight scheduler in (5), we have that

$$(\mathbf{q}^\infty)^\top \mathbf{s}^* \leq \max_{\mathbf{x} \in \mathcal{C}_{\mathbf{H}[\infty]}} (\mathbf{q}^\infty)^\top \mathbf{x} = (\mathbf{q}^\infty)^\top \mathbf{s}^\infty. \quad (15)$$

where  $\mathbf{s}^\infty \triangleq \lim_{B \rightarrow \infty} \mathbf{s}_B^\infty$  and  $\mathbf{H}[\infty]$  represent the full CSI in the steady state. Hence, for any realization of  $\mathbf{q}^\infty$  such that  $\mathbf{q}^\infty \neq \rho \mathbf{s}^*$  for some  $\rho \in \mathbb{R}$ , if  $B$  is sufficiently large, we must have  $(\mathbf{q}^\infty)^\top \mathbf{s}^* - (\mathbf{q}^\infty)^\top \mathbf{s}_B^\infty \leq 0$ . Hence, there exists a critical value  $B_{cr}$  such that for all  $B > B_{cr}$ , the average value of  $(\mathbf{q}^\infty)^\top \mathbf{s}^* - (\mathbf{q}^\infty)^\top \mathbf{s}_B^\infty$  can be made non-positive, i.e.,  $\mathbb{E}\{(\mathbf{q}^\infty)^\top (\mathbf{s}^* - \mathbf{s}_B^\infty)\} \leq 0$ . Hence, we consider two cases based on the positivity of  $\mathbb{E}\{(\mathbf{q}^\infty)^\top (\mathbf{s}^* - \mathbf{s}_B^\infty)\}$  as follows:

*Case I):*  $B \geq B_{cr}$  such that  $\mathbb{E}\{(\mathbf{q}^\infty)^\top (\mathbf{s}^* - \mathbf{s}_B^\infty)\} \leq 0$ : In this case, it follows from (14) that

$$0 \leq \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \sum_{\mathbf{q} \in \mathbb{Z}_+^N} \Pr(\mathbf{q}[t] = \mathbf{q} | \mathbf{q}[0]) \times \left\{ \frac{1}{K^\alpha} \left[ -\frac{1}{\Phi K} \|\mathbf{q}[t] - \mathbf{q}_{B,(K)}^*\|^2 + D_0 \right] \right\}. \quad (16)$$

We now consider the term in the second line in (16) by setting  $\alpha = 0$ . Similar to the proof of Theorem 3, suppose that  $\|\mathbf{q}[t] - \mathbf{q}_{B,(K)}^*\| \geq \beta \sqrt{K}$ , where  $\beta$  will be specified shortly. This implies that  $\frac{1}{\|\mathbf{q}[t] - \mathbf{q}_{B,(K)}^*\|} \leq \frac{1}{\beta}$ . Then, the second line in (16) can be upper bounded as:

$$\begin{aligned} -\frac{1}{\Phi K} \|\mathbf{q}[t] - \mathbf{q}_{B,(K)}^*\|^2 + D_0 &= -\frac{1}{\Phi \sqrt{K}} \|\mathbf{q}[t] - \mathbf{q}_{B,(K)}^*\| \times \\ &\quad \left( \frac{\|\mathbf{q}[t] - \mathbf{q}_{B,(K)}^*\|}{\sqrt{K}} + \frac{D_0 \Phi \sqrt{K}}{\|\mathbf{q}[t] - \mathbf{q}_{B,(K)}^*\|} \right) \\ &\leq -\frac{1}{\Phi \sqrt{K}} \|\mathbf{q}[t] - \mathbf{q}_{B,(K)}^*\| \left( \beta - \frac{D_0 \Phi}{\beta} \right). \end{aligned} \quad (17)$$

Hence, by choosing  $\beta > \sqrt{D_0 \Phi}$ , we have

$$-\frac{1}{\Phi K} \|\mathbf{q}[t] - \mathbf{q}_{B,(K)}^*\|^2 + D_0 \leq -\frac{\hat{\delta}}{\Phi \sqrt{K}} \|\mathbf{q}[t] - \mathbf{q}_{B,(K)}^*\|, \quad (18)$$

where  $\hat{\delta} = \beta - \frac{D_0 \Phi}{\beta} > 0$ . Plugging in  $\beta > \sqrt{D_0 \Phi}$  to define a ball  $\mathcal{B} \triangleq \{\mathbf{q} : \|\mathbf{q} - \mathbf{q}_{B,(K)}^*\| \leq \sqrt{D_0 \Phi K}\}$ , we have

$$-\frac{1}{\Phi K} \|\mathbf{q}[t] - \mathbf{q}_{B,(K)}^*\|^2 + D_0 \leq -\frac{\delta}{\sqrt{K}} \|\mathbf{q}[t] - \mathbf{q}_{B,(K)}^*\|, \text{ if } \mathbf{q}[t] \in \mathcal{B}^c.$$

On the other hand, when  $\|\mathbf{q}[t] - \mathbf{q}_{B,(K)}^*\| \leq \sqrt{D_0 \Phi K}$ , it is clear that  $-(1/\Phi K) \|\mathbf{q}[t] - \mathbf{q}_{B,(K)}^*\|^2 + D_0 \leq \eta$  for some  $\eta > 0$ . Combining these facts, we have

$$\begin{aligned} -\frac{1}{\Phi K} \|\mathbf{q}[t] - \mathbf{q}_{B,(K)}^*\|^2 + D_0 \\ \leq -\frac{\delta}{K} \|\mathbf{q}[t] - \mathbf{q}_{B,(K)}^*\| \mathbb{1}_{\mathcal{B}^c}(\mathbf{q}[t]) + \eta \mathbb{1}_{\mathcal{B}}(\mathbf{q}[t]). \end{aligned} \quad (19)$$

Substituting (19) into (16) yields:

$$\begin{aligned} 0 \leq \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \sum_{\mathbf{q} \in \mathbb{Z}_+^N} \Pr(\mathbf{q}[t] = \mathbf{q} | \mathbf{q}[0]) \times \\ \left( -\frac{\delta}{K} \|\mathbf{q}[t] - \mathbf{q}_{B,(K)}^*\| \mathbb{1}_{\mathcal{B}^c}(\mathbf{q}) + \eta \mathbb{1}_{\mathcal{B}}(\mathbf{q}) \right) \\ = \eta \sum_{\mathbf{q} \in \mathcal{B}} \pi_{\mathbf{q}}^\infty - \frac{\delta}{\sqrt{K}} \sum_{\mathbf{q} \in \mathcal{B}^c} \|\mathbf{q} - \mathbf{q}_{B,(K)}^*\| \pi_{\mathbf{q}}^\infty. \end{aligned} \quad (20)$$

where we use the fact that,  $\forall \mathbf{q} \in \mathbb{Z}_+^N$ ,  $\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \Pr\{\mathbf{q}[t] = \mathbf{q} | \mathbf{q}[0]\} = \pi_{\mathbf{q}}^\infty$ . Re-arranging the terms and with some manipulations, the above inequality can be written as:

$$\begin{aligned} \frac{\delta}{\sqrt{K}} \sum_{\mathbf{q} \in \mathbb{Z}_+^N} \|\mathbf{q} - \mathbf{q}_{B,(K)}^*\| \pi_{\mathbf{q}}^\infty \leq \sum_{\mathbf{q} \in \mathcal{B}} \left( \eta + \frac{\delta}{\sqrt{K}} \|\mathbf{q} - \mathbf{q}_{B,(K)}^*\| \right) \pi_{\mathbf{q}}^\infty \\ \leq (\eta + \delta \beta) \sum_{\mathbf{q} \in \mathcal{B}} \pi_{\mathbf{q}}^\infty \leq (\eta + \delta \beta), \end{aligned} \quad (21)$$

where the second inequality follows from the definition of  $\mathcal{B}$ . Note here that the left-hand-side is precisely  $\frac{\delta}{\sqrt{K}} \mathbb{E}\{\|\mathbf{q}^\infty - \mathbf{q}_{B,(K)}^*\|\}$ . Thus, multiplying both sides by  $\sqrt{K}/\delta$ , we have:

$$\mathbb{E}\{\|\mathbf{q}^\infty - \mathbf{q}_{B,(K)}^*\|\} \leq \left( \beta + \frac{\eta}{\delta} \right) \sqrt{K} = O(\sqrt{K}). \quad (22)$$

*Case II):*  $B \leq B_{cr}$  such that  $\mathbb{E}\{(\mathbf{q}^\infty)^\top (\mathbf{s}^* - \mathbf{s}_B^\infty)\} > 0$ : In this case, we set  $\alpha = 1$ . It thus follows from (12) that:

$$\begin{aligned} \mathbb{E}\{\Delta V_1(\mathbf{q}[t]) | \mathbf{q}[t]\} \leq -\frac{1}{\Phi K^2} \|\mathbf{q}[t] - \mathbf{q}_{B,(K)}^*\|^2 + \\ \frac{1}{K} \|\mathbf{q}[t] - \mathbf{q}_{B,(K)}^*\|^\top D_{(B)} + \frac{D_0}{K}, \end{aligned} \quad (23)$$

where  $D_{(B)}$  is defined in the proof of Theorem 3 (cf. Eq. (34)). Note that (23) is identical to (38). Then, following exactly the same steps as in the proof of Theorem 3, we have:

$$\mathbb{E}\{\Delta V_1(\mathbf{q}[t]) | \mathbf{q}[t] = \mathbf{q}\} \leq -\frac{\delta_1}{K} \|\mathbf{q} - \mathbf{q}_{B,(K)}^*\| \mathbb{1}_{\mathcal{B}_1^c}(\mathbf{q}) + \eta_1 \mathbb{1}_{\mathcal{B}_1}(\mathbf{q}),$$

where  $\delta_1$ ,  $\eta_1$ , and  $\mathcal{B}_1$  are the same as in the proof of Theorem 3. Then, it follows from (13) that

$$\begin{aligned} \mathbb{E}\{V_1(\mathbf{q}[T]) | \mathbf{q}[0]\} - V_1(\mathbf{q}[0]) \leq \eta_1 \sum_{\mathbf{q} \in \mathcal{B}_1} \sum_{t=0}^{T-1} \Pr\{\mathbf{q}[t] = \mathbf{q} | \mathbf{q}[0]\} \\ - \frac{\delta_1}{K} \sum_{\mathbf{q} \in \mathcal{B}_1^c} \|\mathbf{q} - \mathbf{q}_{B,(K)}^*\| \sum_{t=0}^{T-1} \Pr\{\mathbf{q}[t] = \mathbf{q} | \mathbf{q}[0]\}. \end{aligned} \quad (24)$$

Following similar steps as in Case I to divide  $T$  on both sides on (24) and let  $T \rightarrow \infty$ , we have  $0 \leq \eta_1 \sum_{\mathbf{q} \in \mathcal{B}_1} \pi_{\mathbf{q}}^\infty - \frac{\delta_1}{K} \sum_{\mathbf{q} \in \mathcal{B}_1^c} \|\mathbf{q} - \mathbf{q}_{B,(K)}^*\| \pi_{\mathbf{q}}^\infty$ . Re-arranging the terms and with some manipulations, the above inequality can be written as:

$$\begin{aligned} \frac{\delta_1}{K} \sum_{\mathbf{q} \in \mathbb{Z}_+^N} \|\mathbf{q} - \mathbf{q}_{B,(K)}^*\| \pi_{\mathbf{q}}^\infty \leq \sum_{\mathbf{q} \in \mathcal{B}_1} \left( \eta_1 + \frac{\delta_1}{K} \|\mathbf{q} - \mathbf{q}_{B,(K)}^*\| \right) \pi_{\mathbf{q}}^\infty \\ \leq (\eta_1 + \delta_1 \beta_1) \sum_{\mathbf{q} \in \mathcal{B}} \pi_{\mathbf{q}}^\infty \leq (\eta_1 + \delta_1 \beta_1), \end{aligned}$$

where  $\beta_1$  is the same as in the proof of Theorem 3. Note that the left-hand-side is  $\frac{\delta_1}{K} \mathbb{E}\{\|\mathbf{q}^\infty - \mathbf{q}_{B,(K)}^*\|\}$ . Multiplying both sides by  $\frac{K}{\delta_1}$ , we have:

$$\mathbb{E}\{\|\mathbf{q}^\infty - \mathbf{q}_{B,(K)}^*\|\} \leq \left( \beta_1 + \frac{\eta_1}{\delta_1} \right) K$$

$$= \left( \left[ (D_{(B)}\Phi) + \sqrt{(D_{(B)}\Phi)^2 + 4D_0\Phi} \right] + \frac{\eta}{\delta} \right) K = O(D_{(B)}K).$$

This completes the proof of Theorem 1.  $\square$

**PROOF OF THEOREM 2.** To show the results in Theorem 2, we first note that  $\mathbb{E}\{a_n[t]|q_n[t]\} = \min\{U_n'^{-1}(\frac{q_n[t]}{K}, A^{\max})\}$  and  $a_n^* = U_n'^{-1}(\frac{q_n^*}{K})$ ,  $\forall n$ . Thus, we have:

$$\begin{aligned} \|\mathbf{a}_B^\infty - \mathbf{a}_B^*\| &\leq \|\mathbf{a}_B^\infty - \mathbf{a}_B^*\|_1 \\ &= \sum_{n=1}^N \left| \mathbb{E}\left\{ \min\left\{ U_n'^{-1}\left(\frac{q_n^\infty}{K}, A^{\max}\right) \right\} \right\} - U_n'^{-1}\left(\frac{q_{B,(K),n}^*}{K}\right) \right| \\ &\stackrel{(a)}{\leq} \sum_{n=1}^N \mathbb{E}\left\{ \left| \min\left\{ U_n'^{-1}\left(\frac{q_n^\infty}{K}, A^{\max}\right) \right\} - U_n'^{-1}\left(\frac{q_{B,(K),n}^*}{K}\right) \right| \right\} \\ &\stackrel{(b)}{\leq} \sum_{n=1}^N \mathbb{E}\left\{ \left| U_n'^{-1}\left(\frac{q_n^\infty}{K}\right) - U_n'^{-1}\left(\frac{q_{B,(K),n}^*}{K}\right) \right| \right\} \\ &\stackrel{(c)}{=} \sum_{n=1}^N \mathbb{E}\left\{ \left| \left[ U_n'^{-1}\left(\frac{\tilde{q}_n}{K}\right) \right]' \left( \frac{q_n^\infty}{K} - \frac{q_{B,(K),n}^*}{K} \right) \right| \right\} \\ &\stackrel{(d)}{\leq} \sum_{n=1}^N \mathbb{E}\left\{ \left| \frac{1}{U_n''\left(\frac{\tilde{q}_n}{K}\right)} \left| \frac{q_n^\infty}{K} - \frac{q_{B,(K),n}^*}{K} \right| \right\} \\ &\leq \sum_{n=1}^N \mathbb{E}\left\{ \frac{1}{\phi K} |q_n^\infty - q_{B,(K),n}^*| \right\} = \frac{1}{\phi K} \mathbb{E}\{\|\mathbf{q}^\infty - \mathbf{q}_{B,(K)}^*\|_1\} \\ &\leq \frac{\sqrt{N}}{\phi K} \mathbb{E}\{\|\mathbf{q}^\infty - \mathbf{q}_{B,(K)}^*\|\}, \end{aligned} \quad (25)$$

where (a) follows from Jensen's inequality and the convexity of the  $L^1$ -norm; (b) follows from relaxing the projection onto  $[0, A^{\max}]$ ; (c) follows from the mean value theorem; and (d) follows from the inverse function lemma. Recall in the proof of Theorem 1 (cf. (14)), we have  $0 \leq J + \sum_{\mathbf{q} \in \mathbb{Z}_+^N} \pi_{\mathbf{q}}^\infty(\mathbf{q})^\top (\mathbf{s}^* - \mathbf{s}_B^\infty) = J + \mathbb{E}\{(\mathbf{q}^\infty)^\top (\mathbf{s}^* - \mathbf{s}_B^\infty)\}$ . Again, based on the positivity of the term  $\mathbb{E}\{(\mathbf{q}^\infty)^\top (\mathbf{s}^* - \mathbf{s}_B^\infty)\}$ , we consider two cases:

*Case I):*  $B > B_{\text{cr}}$  such that  $\mathbb{E}\{(\mathbf{q}^\infty)^\top (\mathbf{s}^* - \mathbf{s}_B^\infty)\} \leq 0$ : In this case, we can again discard  $\mathbb{E}\{(\mathbf{q}^\infty)^\top (\mathbf{s}^* - \mathbf{s}_B^\infty)\}$  in (14) and let  $\alpha = 0$  to obtain:

$$0 \leq \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \sum_{\mathbf{q} \in \mathbb{Z}_+^N} \Pr(\mathbf{q}[t] = \mathbf{q}|\mathbf{q}[0]) \times \left\{ -\frac{1}{\Phi K} \|\mathbf{q}[t] - \mathbf{q}_{B,(K)}^*\|^2 \right\} + D_0.$$

By re-arranging, multiplying both sides by  $\Phi K$ , and noting that  $\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \Pr\{\mathbf{q}[t] = \mathbf{q}|\mathbf{q}[0]\} = \pi_{\mathbf{q}}^\infty$ , we have

$$\mathbb{E}\{\|\mathbf{q}^\infty - \mathbf{q}_{B,(K)}^*\|^2\} \leq D_0 \Phi K. \quad (26)$$

It then follows from (25) that

$$\begin{aligned} \|\mathbf{a}_B^\infty - \mathbf{a}_B^*\|^2 &\leq \left( \frac{\sqrt{N}}{\phi K} \mathbb{E}\{\|\mathbf{q}^\infty - \mathbf{q}_{B,(K)}^*\|\} \right)^2 \\ &\stackrel{(a)}{\leq} \frac{N}{\phi^2 K^2} \mathbb{E}\{\|\mathbf{q}^\infty - \mathbf{q}_{B,(K)}^*\|^2\} \stackrel{(b)}{\leq} \frac{N}{\phi^2 K^2} D_0 \Phi K = \frac{N D_0}{\phi^2 K}, \end{aligned} \quad (27)$$

where (a) follows from Jensen's inequality; and (b) follows from (26). Taking square root on both sides of (27) yields  $\|\mathbf{a}_B^\infty - \mathbf{a}_B^*\| = O(1/\sqrt{K})$ .

*Case II):*  $B \leq B_{\text{cr}}$  such that  $\mathbb{E}\{(\mathbf{q}^\infty)^\top (\mathbf{s}^* - \mathbf{s}_B^\infty)\} > 0$ : In this case, we set  $\alpha = 1$  and it follows from (12) that:

$$\begin{aligned} \mathbb{E}\{\Delta V_1(\mathbf{q}[t])|\mathbf{q}[t]\} &\leq -\frac{1}{\Phi K^2} \|\mathbf{q}[t] - \mathbf{q}_{B,(K)}^*\|^2 + \\ &\quad \frac{D_{(B)}}{K} \|\mathbf{q}[t] - \mathbf{q}_{B,(K)}^*\| + \frac{D_0}{K} \\ &= -\frac{1}{\Phi K^2} \left( \|\mathbf{q}[t] - \mathbf{q}_{B,(K)}^*\| - \frac{D_{(B)}\Phi K}{2} \right)^2 + D, \end{aligned} \quad (28)$$

where  $D_{(B)}$  is defined in the proof of Theorem 3 (cf. Eq. (34)) and  $D \triangleq \frac{D_{(B)}}{4} + \frac{D_0}{\Phi K}$ . Telescoping the inequality in (28) from  $t = 0$  to  $T - 1$  yields:

$$\begin{aligned} \mathbb{E}\{V_1(\mathbf{q}[T])|\mathbf{q}[0]\} - V_1(\mathbf{q}[0]) &\leq -\frac{1}{\Phi K^2} \sum_{t=0}^{T-1} \sum_{\mathbf{q} \in \mathbb{Z}_+^N} \Pr\{\mathbf{q}[t] = \mathbf{q}|\mathbf{q}[0]\} \\ &\quad \times \left( \|\mathbf{q}[t] - \mathbf{q}_{B,(K)}^*\| - \frac{D_{(B)}\Phi K}{2} \right)^2 + DT. \end{aligned} \quad (29)$$

Dividing both sides of (29) by  $\frac{T}{K^2}$ , letting  $T \rightarrow \infty$ , and noting that  $\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \Pr\{\mathbf{q}[t] = \mathbf{q}|\mathbf{q}[0]\} = \pi_{\mathbf{q}}^\infty$ ,  $\forall \mathbf{q} \in \mathbb{Z}_+^N$ , we have that:

$$\mathbb{E}\left\{ \left( \|\mathbf{q}^\infty - \mathbf{q}_{B,(K)}^*\| - \frac{D_{(B)}\Phi K}{2} \right)^2 \right\} \leq D\Phi K^2.$$

Taking square root on both sides yields:

$$\left[ \mathbb{E}\left\{ \left( \|\mathbf{q}^\infty - \mathbf{q}_{B,(K)}^*\| - \frac{D_{(B)}\Phi K}{2} \right)^2 \right\} \right]^{\frac{1}{2}} \leq K\sqrt{D\Phi}. \quad (30)$$

Moreover, examining the left-hand-side of (30), we have

$$\begin{aligned} &\left[ \mathbb{E}\left\{ \left( \|\mathbf{q}^\infty - \mathbf{q}_{B,(K)}^*\| - \frac{D_{(B)}\Phi K}{2} \right)^2 \right\} \right]^{\frac{1}{2}} \\ &\stackrel{(a)}{\geq} \mathbb{E}\left\{ \left[ \left( \|\mathbf{q}^\infty - \mathbf{q}_{B,(K)}^*\| - \frac{D_{(B)}\Phi K}{2} \right)^2 \right]^{\frac{1}{2}} \right\} \\ &= \mathbb{E}\left\{ \left| \|\mathbf{q}^\infty - \mathbf{q}_{B,(K)}^*\| - \frac{D_{(B)}\Phi K}{2} \right| \right\} \\ &\geq \mathbb{E}\left\{ \|\mathbf{q}^\infty - \mathbf{q}_{B,(K)}^*\| - \frac{D_{(B)}\Phi K}{2} \right\} \\ &= \mathbb{E}\{\|\mathbf{q}^\infty - \mathbf{q}_{B,(K)}^*\|\} - \frac{D_{(B)}\Phi K}{2}, \end{aligned} \quad (31)$$

where (a) follows from Jensen's inequality. Combining (25), (30), and (31) yields:

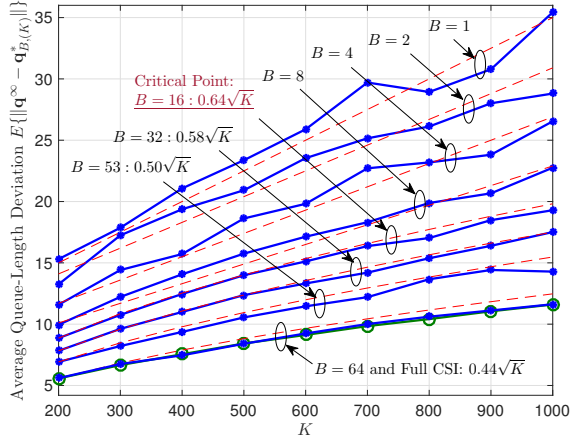
$$\begin{aligned} \|\mathbf{a}_B^\infty - \mathbf{a}_B^*\| &\leq \frac{\sqrt{N}}{\phi K} \mathbb{E}\{\|\mathbf{q}^\infty - \mathbf{q}_{B,(K)}^*\|\} \\ &= \frac{\sqrt{N}}{\phi K} \left( \frac{D_{(B)}\Phi K}{2} + K\sqrt{D\Phi} \right) = O(D_{(B)}). \end{aligned}$$

Note that Cases I and II are exactly the same results as stated in Theorem 2. This completes the proof.  $\square$

## 4. NUMERICAL RESULTS

In this section, we conduct numerical experiments to verify the theoretical results presented in Section 3. In our simulations, we use a 128-antenna M-MIMO base station with MF precoding to serve four users. Each user's channel is i.i.d. Rayleigh faded. The maximum total signal-to-noise ratio (SNR) of the BS is set to 30dB. We use  $\log(\cdot + 0.001)$  as the utility function for each user, i.e., the proportional fairness metric [10]. We adopt the random vector quantization (RVQ) scheme, which has been widely used in the MIMO limited CSI feedback literature [14, 15, 15, 18]. The value of  $B$  is set to be 1, 2, 4, 8, 16, 32, 53, and 64, covering cases





**Figure 2:** Average queue-length deviation  $\mathbb{E}\{\|\mathbf{q}^\infty - \mathbf{q}_{B,K}^*\|\}$  with respect to  $K$  for  $B = 1, 2, 4, 8, 16, 32, 53, 64$ .

from the simplest two-state channel quantization to channel quantizations with high granularity.

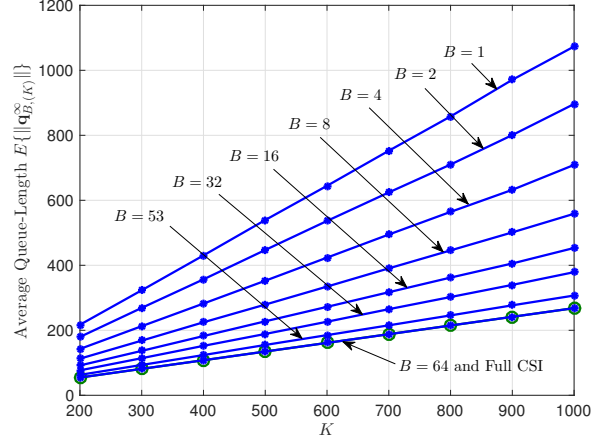
We first study the impacts of  $B$  on the delay performance. The results of average queue-length deviations with respect to the changes of  $B$  are illustrated in Figure 2. For each scaling curve, we also plot an accompanying line (the red dash lines) to highlight its growth trend. We can observe that when  $B$  is small, the average queue-length deviation grows faster than the square root law and approximately exhibits a linear growth with respect to  $K$ . This confirms the first part of Theorem 1. In this example,  $B = 16$  bits turns out to be the critical point of phase transitioning, beyond which the queue-length deviations are bounded by the  $O(\sqrt{K})$  scaling curves. This confirms the second part of Theorem 1. Also, when  $B = 64$ , we can see that the queue-length deviations almost coincide with that in the full CSI case, showing that the 64-bit RVQ scheme is almost as accurate as full CSI.

The results of average queue-lengths' growth with respect to  $K$  under different values of  $B$  are illustrated in Figure 3. We can see from Figure 3 that the average queue-lengths increase linearly with respect to  $K$  under all  $B$  values, agreeing with Lemma 2. Also, the value of  $B$  plays an important role in the slope of the linear scaling: the large the  $B$  value, the more gradual the slope, again confirming Theorem 1. Also, the slope of  $B = 64$  is almost the same as that of full CSI.

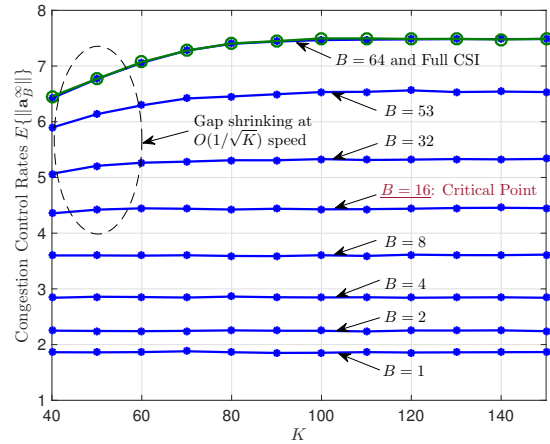
Next, we study the impacts of  $B$  on the congestion control performance and the results are illustrated in Figure 4. When  $B$  is small, we can observe in Figure 4 that  $\mathbf{a}_B^\infty$  are independent of  $K$  and only affected by  $B$ . The congestion control rates approach that under full CSI as  $B$  increases. This confirms the first part of Theorem 2 and Lemma 3. Similar to the growths of queue-length deviations, we can also observe that  $B = 16$  is the critical point, beyond which the congestion control rates start to exhibit an  $O(1/\sqrt{K})$  shrinking gap to  $\mathbf{a}_B^*$ . All of these observations agree with the phase transitioning results in Theorem 2.

## 5. CONCLUSION

In this paper, we conducted an in-depth theoretical study on the impact of limited CSI on the performances of the queue-length-based joint congestion control and scheduling algorithm in M-MIMO cellular networks. We have theoretically characterized the queueing delay and congestion con-



**Figure 3:** The growths of average queue-lengths with respect to  $K$  for  $B = 1, 2, 4, 8, 16, 32, 53$ , and  $64$ .



**Figure 4:** The steady-state congestion control rates with respect to  $K$  for  $B = 1, 2, 4, 8, 16, 32, 53$ , and  $64$ .

trol scalings under limited CSI. We showed that there exist phase transitioning phenomena in the steady-state queue-length and congestion control rate deviations with respect to CSI quality. Collectively, our theoretical results in this paper advance the understanding of the interactions and trade-offs between delay, throughput, and the accuracy/complexity of CSI acquisition in M-MIMO networks. Our work also establishes a unifying theoretical framework as well as practical design guidelines to enable the development of effective channel quantization schemes for M-MIMO networks. Moreover, our work serves as a first step toward an exciting M-MIMO networking research paradigm that explores various new congestion control and scheduling algorithmic designs, which could potentially offer better throughput and delay performances under limited CSI.

## 6. REFERENCES

- [1] E. G. Larsson, O. Edfors, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 186–195, Feb. 2014.
- [2] F. Rusek, D. Persson, B. K. Lau, E. G. Larsson, T. L. Marzetta, O. Edfors, and F. Tufvesson, "Scaling up MIMO: Opportunities and challenges with very large



- arrays,” *IEEE Signal Process. Mag.*, vol. 30, no. 1, pp. 40–46, Jan. 2013.
- [3] L. Lu, G. Y. Li, A. L. Swindlehurst, A. Ashikhmin, and R. Zhang, “An overview of massive MIMO: Benefits and challenges,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 5, pp. 742–758, Oct. 2014.
- [4] Argos: Practical many-antenna base stations. [Online]. Available: argos.rice.edu
- [5] E. G. Larsson and F. Tufvesson, “Massive MIMO systems tutorial,” 2013. [Online]. Available: http://www.commsys.isy.liu.se/vlm/icc\_tutorial\_LP2.pdf
- [6] X. Lin and N. B. Shroff, “The impact of imperfect scheduling on cross-layer congestion control in wireless networks,” *IEEE/ACM Trans. Netw.*, vol. 14, no. 2, pp. 302–315, Apr. 2006.
- [7] A. Eryilmaz and R. Srikant, “Joint congestion control, routing, and MAC for stability and fairness in wireless networks,” *IEEE J. Sel. Areas Commun.*, vol. 24, no. 8, pp. 1514–1524, Aug. 2006.
- [8] M. J. Neely, E. Modiano, and C.-P. Li, “Fairness and optimal stochastic control for heterogeneous networks,” *IEEE/ACM Trans. Netw.*, vol. 16, no. 2, pp. 396–409, Apr. 2008.
- [9] A. Eryilmaz and R. Srikant, “Fair resource allocation in wireless networks using queue-length-based scheduling and congestion control,” *IEEE/ACM Trans. Netw.*, vol. 15, no. 6, pp. 1333–1344, Dec. 2007.
- [10] X. Lin, N. B. Shroff, and R. Srikant, “A tutorial on cross-layer optimization in wireless networks,” *IEEE J. Sel. Areas Commun.*, vol. 24, no. 8, pp. 1452–1463, Aug. 2006.
- [11] C. Shepard, H. Yu, N. Anand, L. E. Li, T. L. Marzetta, R. Yang, and L. Zhong, “Argos: Practical many-antenna base stations,” in *Proc. ACM MobiCom*, Istanbul, Turkey, August 2012, pp. 53–64.
- [12] A. Eryilmaz and R. Srikant, “Fair resource allocation in wireless networks using queue-length-based scheduling and congestion control,” in *Proc. IEEE INFOCOM*, Miami, FL, Mar. 2005, pp. 1804–1814.
- [13] H. Weingarten, Y. Steinberg, and S. Shamai (Shitz), “The capacity region of the Gaussian multiple-input multiple-output broadcast channel,” *IEEE Trans. Inf. Theory*, vol. 52, no. 9, pp. 3936–3964, Sep. 2006.
- [14] N. Jindal, “MIMO broadcast channels with finite rate feedback,” *IEEE Trans. Inf. Theory*, vol. 52, no. 11, pp. 5045–5059, Nov. 2006.
- [15] W. Santipach and M. Honig, “Signature optimization for CDMA with limited feedback,” *IEEE Trans. Inf. Theory*, vol. 51, no. 10, pp. 3475–3492, Oct. 2005.
- [16] W. Santipach and M. L. Honig, “Asymptotic performance of MIMO wireless channels with limited feedback,” in *Proc. IEEE Mil. Commun. Conf.*, vol. 1, Oct. 2003, pp. 141–146.
- [17] —, “Asymptotic capacity of beamforming with limited feedback,” in *Proc. IEEE ISIT*, Jul. 2004, p. 290.
- [18] C. K. Au-Yeung and D. J. Love, “On the performance of random vector quantization limited feedback beamforming in a MISO system,” *IEEE Trans. Wireless Commun.*, vol. 6, no. 2, pp. 458–462, Feb.

2007.

- [19] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty, *Nonlinear Programming: Theory and Algorithms*, 3rd ed. New York, NY: John Wiley & Sons Inc., 2006.
- [20] S. P. Meyn and R. L. Tweedie, *Markov Chains and Stochastic Stability*, 2nd ed. Cambridge, UK: Cambridge University Press, 2009.

## APPENDIX

### A. PROOF OF LEMMA 1

For ease of exposition, we first show the second part of Lemma 1. Let  $\{p_n[t], n = 1, \dots, N\}$  be an arbitrary feasible power allocation. Since the BS performs MF beamforming by treating  $\widehat{\mathbf{H}}[t]$  as if it is the accurate CSI, the received signal can be written as  $y_n[t] = s_n[t]p_n[t]\mathbf{h}_n^\top[t]\widehat{\mathbf{w}}_n[t] + \sum_{j=1, j \neq n}^N s_j[t]p_j[t]\mathbf{h}_n^\top[t]\widehat{\mathbf{w}}_j[t] + v_n[t]$ , where  $\widehat{\mathbf{w}}_j[t] = \widehat{\mathbf{h}}_j[t]$ ,  $1 \leq j \leq N$ , i.e., the  $j$ -th row of  $\widehat{\mathbf{H}}[t]$ . Hence, the MF rates  $s_{B,n}[t]$  achieved under  $\mathbf{H}[t]$  based on the belief that the CSI is  $B$ -bit CSI  $\widehat{\mathbf{H}}[t]$  can be computed as:

$$s_{B,n}[t] = \log_2 \left( 1 + \frac{p_n[t]|\mathbf{h}_n^\dagger[t]\widehat{\mathbf{h}}_n[t]|^2}{N_0 + \sum_{j=1, j \neq n}^N p_j[t]|\mathbf{h}_n^\dagger[t]\widehat{\mathbf{h}}_j[t]|^2} \right) < \log_2 \left( 1 + \frac{p_n[t]}{N_0} \|\mathbf{h}_n[t]\|^2 \right) = s_n[t], \quad \forall n, \quad (32)$$

where the inequality in (32) holds because  $|\mathbf{h}_n^\dagger[t]\widehat{\mathbf{h}}_n[t]|^2 \leq \|\mathbf{h}_n[t]\|^2$  and  $|\mathbf{h}_n^\dagger[t]\widehat{\mathbf{h}}_j[t]|^2 \geq 0$ . Thus, for every rate point  $\mathbf{s}_B[t] = [s_{B,1}[t], \dots, s_{B,N}[t]]^T \in \mathcal{C}_{\widehat{\mathbf{H}}[t]}$ , its corresponding power allocation  $\{p_1[t], \dots, p_N[t]\}$  achieves a rate point  $\mathbf{s}[t] = [s_1[t], \dots, s_N[t]]^T \in \mathcal{C}_{\mathbf{H}[t]}$  that dominates  $\mathbf{s}_B[t]$  in every coordinate. Hence,  $\mathcal{C}_{\widehat{\mathbf{H}}[t]} \subseteq \mathcal{C}_{\mathbf{H}[t]}$ . Also, as  $B \rightarrow \infty$ ,  $\widehat{\mathbf{H}}[t] \rightarrow \mathbf{H}[t]$ . It thus follows from (32) that  $\mathbf{s}_B[t] \uparrow \mathbf{s}[t]$ , which implies that  $\mathcal{C}_{\widehat{\mathbf{H}}[t]} \rightarrow \mathcal{C}_{\mathbf{H}[t]}$ .

Next, we argue why the first part of Lemma 1 is true. Let  $\mathcal{B}_n^1$  and  $\mathcal{B}_n^2$  denote the vector quantization codebooks corresponding to  $B_1$  and  $B_2$  bits, respectively. Since  $B_1 \leq B_2$ , it follows that the codebook sizes  $|\mathcal{B}_n^1| \leq |\mathcal{B}_n^2|$ . Hence, given codebook  $\mathcal{B}_n^1$ , one can construct  $\mathcal{B}_n^2$  by simply retaining all codewords in  $\mathcal{B}_n^1$  and adding new code words that are not in  $\mathcal{B}_n^1$ , which implies  $\mathcal{B}_n^1 \subset \mathcal{B}_n^2$ . As a result, for any given CSI  $\mathbf{h}_n[t]$ , one can always find a codeword in  $\mathcal{B}_n^2$  whose distance to  $\mathbf{h}_n[t]$  is not larger than that from  $\mathcal{B}_n^1$  in the sense of (3). Hence, the SINR term in (32) becomes larger under  $\mathcal{B}_n^2$ , implying  $s_{B_1,n}[t] \leq s_{B_2,n}[t]$ . Since this is true for arbitrary power allocation, we have  $\mathcal{C}_{\widehat{\mathbf{H}}[t]} \subseteq \mathcal{C}_{\mathbf{H}[t]}$ .

### B. PROOF OF LEMMA 2

Dividing  $K$  on both sides of (9), we have  $\frac{1}{K}\Theta_K(\mathbf{q}_B) = \max_{\mathbf{a}, s_B \in \bar{\mathcal{C}}^B} \left\{ \sum_{n=1}^N U_n(a_n) + \sum_{n=1}^N \widehat{q}_{B,n}(s_{B,n} - a_n) \right\}$ , where  $\widehat{q}_{B,n} = q_{B,n}/K$ . Note that the right hand side is precisely  $\Theta_1(\mathbf{q}_B)$ , for which the maximizer is  $\widehat{\mathbf{q}} = \mathbf{q}_{B,(1)}^*$ . Hence, we have  $\Theta_K(\mathbf{q})$  is maximized at  $K\mathbf{q}_{B,(1)}^*$ . This proves the first part of Lemma 2.

To show the second part of Lemma 2, we first note from the KKT complementary slackness condition and the monotonicity of  $U_n(\cdot)$  that, at optimality,  $a_n^* = s_{B,n}^*, \forall n$ . We let  $a_n^*(B_1)$  and  $a_n^*(B_2)$  denote the optimal congestion control rates under  $B_1$  and  $B_2$ , respectively. If  $B_1 \leq B_2$ , we have

from Lemma 1 that  $s_{B_1,n}^* \leq s_{B_2,n}^*$ , which further implies  $a_n^*(B_1) \leq a_n^*(B_2)$ . On the other hand, from the KKT stationarity condition, we have  $U'_n(a_n^*(B)) - q_{(B),n}^* = 0$ . Since  $a_n^*(B_1) \leq a_n^*(B_2)$ , it follows from the concavity of  $U_n(\cdot)$  that  $q_{(B_1),n}^* \geq q_{(B_2),n}^*$ . This completes the proof.

### C. PROOF OF THEOREM 3

Consider the quadratic Lyapunov function defined in Theorem 3:  $V(\mathbf{q}[t]) = \frac{1}{2K} \|\mathbf{q}[t] - \mathbf{q}_{B,(K)}^*\|^2$ , where  $\mathbf{q}[t]$  represents the queue-length vector in time-slot  $t$  under parameters  $K$  and  $B$ ; and  $\mathbf{q}_{B,(K)}^*$  denotes the optimal dual solution for the static version of Problem JCCR under parameter  $K$ . Then, the one-slot mean Lyapunov drift of  $V_K(\mathbf{q}[t])$ , which can be computed as:

$$\begin{aligned} & \mathbb{E}\{V(\mathbf{q}[t+1]) - V(\mathbf{q}[t]) | \mathbf{q}[t]\} \\ &= \frac{1}{2K} \mathbb{E} \left\{ (\mathbf{q}[t+1] - \mathbf{q}[t])^\top (\mathbf{q}[t+1] + \mathbf{q}[t] - 2\mathbf{q}_{B,(K)}^*) | \mathbf{q}[t] \right\} \\ &\stackrel{(a)}{\leq} \frac{1}{2K} \mathbb{E} \left\{ (-\mathbf{s}_B[t] + \mathbf{a}[t])^\top (2\mathbf{q}[t] - 2\mathbf{q}_{B,(K)}^* - \mathbf{s}_B[t] + \mathbf{a}[t]) | \mathbf{q}[t] \right\} \\ &= \frac{1}{K} (\mathbf{q}[t] - \mathbf{q}_{B,(K)}^*)^\top (-\mathbf{s}_B[t] + \mathbf{a}[t]) + \frac{1}{2K} \mathbb{E} \left\{ \|\mathbf{s}_B[t] + \mathbf{a}[t]\|^2 \right\}, \end{aligned}$$

where (a) follows from the non-expansive property of the  $\max\{0, \cdot\}$  operation. Note that, from the definition of Algorithm 1, we have  $\mathbb{E}\{\|\mathbf{a}[t]\|^2 | \mathbf{q}[t]\} < A_2^{\max} N$ . Also, since  $s_{B,n}[t]$  falls in a bounded instantaneous capacity region  $\mathcal{C}_{\hat{\mathbf{H}}[t]}$ ,  $\forall n$ , we must have  $s_{B,n}[t] \leq s^{\max}$  for some  $s^{\max} > 0$ . Hence, by defining  $D_0 \triangleq \frac{N}{2}(A_2^{\max} + (s^{\max})^2)$ , we have

$$\begin{aligned} & \mathbb{E} \{ \Delta V(\mathbf{q}[t]) | \mathbf{q}[t] \} \leq \frac{1}{K} (\mathbf{q}[t] - \mathbf{q}_{B,(K)}^*)^\top \mathbb{E} \{ \mathbf{a}[t] - \mathbf{s}_B[t] \} + \frac{D_0}{K} \\ &\stackrel{(a)}{=} \frac{1}{K} (\mathbf{q}[t] - \mathbf{q}_{B,(K)}^*)^\top (\mathbb{E} \{ \mathbf{a}[t] | \mathbf{q}[t] \} - \mathbf{s}_B^*) + \\ &\quad \frac{1}{K} \mathbb{E} \{ (\mathbf{q}[t] - \mathbf{q}_{B,(K)}^*)^\top (\mathbf{s}_B^* - \mathbf{s}_B[t]) | \mathbf{q}[t] \} + \frac{D_0}{K}, \\ &\stackrel{(b)}{\leq} \frac{1}{K} (\mathbf{q}[t] - \mathbf{q}_{B,(K)}^*)^\top (\mathbb{E} \{ \mathbf{a}[t] | \mathbf{q}[t] \} - \mathbf{s}_B^*) + \\ &\quad \frac{1}{K} \|\mathbf{q}[t] - \mathbf{q}_{B,(K)}^*\| \times \mathbb{E} \{ \|\mathbf{s}_B^* - \mathbf{s}_B[t]\| | \mathbf{q}[t] \} + \frac{D_0}{K}, \end{aligned} \quad (33)$$

where  $\mathbf{s}_B^*$  is such that  $(\mathbf{s}_B^*, \mathbf{q}_{B,(K)}^*)$  is a pair of optimal primal and dual solutions to Problem  $K$ -DJCCS under parameter  $K$ . In (33), (a) follows from adding and subtracting  $\mathbf{s}_B^*$  as well as the fact that  $\mathbf{a}[t]$  is independent of the channel state and determined solely by  $\mathbf{q}[t]$ ; and (b) follows from Cauchy-Schwarz inequality.

Note from Lemma 3 that  $\mathbf{s}_B^*$  is independent of  $K$  and  $s_{B,n}[t] \in \mathcal{C}_{\hat{\mathbf{H}}[t]}$  is upper-bounded. Thus, we have

$$\mathbb{E} \{ \|\mathbf{s}_B^* - \mathbf{s}_B[t]\| | \mathbf{q}[t] \} \leq D_{(B)} \triangleq \max_{\mathbf{q}: \|\mathbf{q}\|=1} \mathbb{E} \{ \|\mathbf{s}_B^* - \mathbf{s}_B\| | \mathbf{q} \}, \quad (34)$$

where  $D_{(B)}$  signifies that its value depends on  $B$ . Hence, we can further upper bound (33) as:

$$\begin{aligned} & \mathbb{E} \{ \Delta V(\mathbf{q}[t]) | \mathbf{q}[t] \} \leq \frac{1}{K} (\mathbf{q}[t] - \mathbf{q}_{B,(K)}^*)^\top (\mathbb{E} \{ \mathbf{a}[t] | \mathbf{q}[t] \} - \mathbf{s}_B^*) + \\ &\quad \frac{1}{K} \|\mathbf{q}[t] - \mathbf{q}_{B,(K)}^*\| D_{(B)} + \frac{D_0}{K}, \end{aligned} \quad (35)$$

Now, let us consider the first term on the right hand side in (35), i.e.,  $\frac{1}{K} (\mathbf{q}[t] - \mathbf{q}_{B,(K)}^*)^\top (\mathbb{E} \{ \mathbf{a}[t] | \mathbf{q}[t] \} - \mathbf{s}_B^*)$ . Since  $U_n(\cdot)$  is concave and increasing,  $\forall n$ , we have

$$(q_n[t] - q_{B,(K),n}^*)^\top \left[ U_n'^{-1} \left( \frac{q_n[t]}{K} \right) - U_n'^{-1} \left( \frac{q_{B,(K),n}^*}{K} \right) \right] \leq 0.$$

Thus, by Cauchy-Schwarz inequality, we have:

$$\begin{aligned} & (\mathbf{q}[t] - \mathbf{q}_{B,(K)}^*)^\top (\mathbb{E} \{ \mathbf{a}[t] | \mathbf{q}[t] \} - \mathbf{s}_B^*) = \sum_{n=1}^N (q_n[t] - q_{B,(K),n}^*)^\top \\ &\quad \times \left[ U_n'^{-1} \left( \frac{q_n[t]}{K} \right) - U_n'^{-1} \left( \frac{q_{B,(K),n}^*}{K} \right) \right] \leq - \sum_{n=1}^N |q_n[t] - \\ &\quad q_{B,(K),n}^*| \left| U_n'^{-1} \left( \frac{q_n[t]}{K} \right) - U_n'^{-1} \left( \frac{q_{B,(K),n}^*}{K} \right) \right|. \end{aligned} \quad (36)$$

By the strong convexity of  $-U_n(\cdot)$  and the Lipschitz continuity of  $U_n'(\cdot)$ , we have

$$|U_n'(a_{n,1}) - U_n'(a_{n,2})| \leq \Phi |a_{n,1} - a_{n,2}|.$$

Therefore, by the inverse function lemma, we have

$$\frac{1}{\Phi} \left| \frac{q_n[t]}{K} - \frac{q_{B,(K),n}^*}{K} \right| \leq \left| U_n'^{-1} \left( \frac{q_n[t]}{K} \right) - U_n'^{-1} \left( \frac{q_{B,(K),n}^*}{K} \right) \right|.$$

Hence, we can further upper-bound (36) as:

$$\begin{aligned} & (\mathbf{q}[t] - \mathbf{q}_{B,(K)}^*)^\top (\mathbb{E} \{ \mathbf{a}[t] | \mathbf{q}[t] \} - \mathbf{s}_B^*) \leq - \frac{1}{\Phi K} \sum_{n=1}^N (q_n[t] - \\ &\quad q_{B,(K),n}^*)^2 = - \frac{1}{\Phi K} \|\mathbf{q}[t] - \mathbf{q}_{B,(K)}^*\|^2. \end{aligned} \quad (37)$$

Substituting (37) into (35), we have

$$\begin{aligned} & \mathbb{E} \{ \Delta V(\mathbf{q}[t]) | \mathbf{q}[t] \} \leq - \frac{1}{\Phi K^2} \|\mathbf{q}[t] - \mathbf{q}_{B,(K)}^*\|^2 + \\ &\quad \frac{1}{K} \|\mathbf{q}[t] - \mathbf{q}_{B,(K)}^*\| D_{(B)} + \frac{D_0}{K}. \end{aligned} \quad (38)$$

Now, suppose that  $\|\mathbf{q}[t] - \mathbf{q}_{B,(K)}^*\| \geq \beta_1 K$ , where  $\beta_1$  will be specified shortly. Note also that  $K \geq 1$ , we have

$$\frac{1}{\|\mathbf{q}[t] - \mathbf{q}_{B,(K)}^*\|} \leq \frac{1}{\beta_1 K} \leq \frac{1}{\beta_1}.$$

It then follows that (38) can be further upper bounded as:

$$\begin{aligned} & \mathbb{E} \{ \Delta V(\mathbf{q}[t]) | \mathbf{q}[t] \} = - \frac{1}{\Phi K} \|\mathbf{q}[t] - \mathbf{q}_{B,(K)}^*\| \cdot \frac{\|\mathbf{q}[t] - \mathbf{q}_{B,(K)}^*\|}{K} \\ &\quad + \frac{1}{K} \|\mathbf{q}[t] - \mathbf{q}_{B,(K)}^*\| D_1 + \frac{D_0}{K} \frac{\|\mathbf{q}[t] - \mathbf{q}_{B,(K)}^*\|}{\|\mathbf{q}[t] - \mathbf{q}_{B,(K)}^*\| K} \\ &\leq - \frac{1}{\Phi K} \|\mathbf{q}[t] - \mathbf{q}_{B,(K)}^*\| \left( \beta_1 - D_{(B)} \Phi - \frac{D_0 \Phi}{\beta_1} \right). \end{aligned} \quad (39)$$

By choosing  $\beta_1$  such that  $\beta_1 - D_1 \Phi - \frac{D_0 \Phi}{\beta_1} > 0$ , we have

$$\mathbb{E} \{ \Delta V(\mathbf{q}[t]) | \mathbf{q}[t] \} \leq - \frac{\hat{\delta}_1}{\Phi K} \|\mathbf{q}[t] - \mathbf{q}_{B,(K)}^*\| \quad (40)$$

where  $\hat{\delta}_1 = \beta_1 - D_{(B)} \Phi - \frac{D_0 \Phi}{\beta_1}$ . Solving  $\beta_1 - D_{(B)} \Phi - \frac{D_0 \Phi}{\beta_1} = 0$  and plugging in the obtained  $\beta_1$  to define a ball  $\mathcal{B}_1 \triangleq \{ \mathbf{q} : \|\mathbf{q} - \mathbf{q}_{B,(K)}^*\| \leq \frac{K}{2} [(D_{(B)} \Phi) + \sqrt{(D_{(B)} \Phi)^2 + 4D_0 \Phi}] \}$ , we have

$$\mathbb{E} \{ \Delta V(\mathbf{q}[t]) | \mathbf{q}[t] \} \leq - \frac{\delta_1}{K} \|\mathbf{q}[t] - \mathbf{q}_{B,(K)}^*\|, \text{ if } \mathbf{q}[t] \in \mathcal{B}_1^c, \quad (41)$$

where  $\delta_1 \triangleq \frac{\hat{\delta}_1}{\Phi}$ . On the other hand, when  $\mathbf{q}[t] \in \mathcal{B}_1$ , it is clearly true that  $\mathbb{E} \{ \Delta V(\mathbf{q}[t]) | \mathbf{q}[t] \} \leq \eta_1$  for some  $\eta_1 > 0$ . Combining these facts yields the following:

$$\mathbb{E} \{ \Delta V(\mathbf{q}[t]) | \mathbf{q}[t] = \mathbf{q} \} \leq - \frac{\delta_1}{K} \|\mathbf{q} - \mathbf{q}_{B,(K)}^*\| \mathbb{1}_{\mathcal{B}_1^c}(\mathbf{q}) + \eta_1 \mathbb{1}_{\mathcal{B}_1}(\mathbf{q}).$$

This completes the proof of Theorem 3.