Training and Feedback Optimization for Multiuser MIMO Downlink

Mari Kobayashi[1], Nihar Jindal[2], Giuseppe Caire[3]

[1] SUPELEC, Gif-sur-Yvette, 91192, France

[2] University of Minnesota, Minneapolis MN, 55455 USA

[3] University of Southern California, Los Angeles CA, 90089 USA

Abstract

We consider a MIMO fading broadcast channel where the fading channel coefficients are constant over time-frequency blocks that span a coherent time $\times$ a coherence bandwidth. In closed-loop systems, channel state information at transmitter (CSIT) is acquired by the downlink training sent by the base station and an explicit feedback from each user terminal. In open-loop systems, CSIT is obtained by exploiting uplink training and channel reciprocity. We use a tight closed-form lower bound on the ergodic achievable rate in the presence of CSIT errors in order to optimize the overall system throughput, by taking explicitly into account the overhead due to channel estimation and channel state feedback. Based on three time-frequency block models inspired by actual systems, we provide some useful guidelines for the overall system optimization. In particular, digital (quantized) feedback is found to offer a substantial advantage over analog (unquantized) feedback.

Keywords: MIMO broadcast channel, Multiuser MIMO Downlink, Channel State Information Feedback, Channel Estimation.

I. Introduction

The downlink of a wireless system with one Base Station (BS) with N_t antennas and K User Terminals (UTs) with a single antenna each is modeled by a MIMO Gaussian broadcast channel [1], defined by

$$y_k[i] = \mathbf{h}_k^{\mathsf{H}}\mathbf{x}[i] + z_k[i], \quad k = 1, \ldots, K \tag{1}$$

for $i = 1, \ldots, T$, where $y_k[i]$ is the channel output at UT k, $z_k[i] \sim \mathcal{CN}(0, N_0)$ is the corresponding Additive White Gaussian Noise (AWGN) process, $\mathbf{h}_k \in \mathbb{C}^{N_t}$ is the vector of channel

coefficients from the BS antenna array to the k-th UT antenna and $\mathbf{x}[i]$ is the vector of channel input symbols transmitted by the BS, subject to the average power constraint $\mathbb{E}[|\mathbf{x}[i]|^2] \leq P$ (enforced for each channel use i). We denote the downlink signal-to-noise ratio (SNR) at each UT by $\rho \triangleq \frac{P}{N_0}$. We assume a block fading model where the channel vectors $\{\mathbf{h}_k\}$ remain constant over a coherence block of T channel uses. The block length T is related to two physical channel parameters, the coherence time T_c and the coherence bandwidth W_c by $T = W_c T_c$. For example, taking as typical values $W_c = 500$ kHz and $T_c = 2.5$ ms (from [2]), we obtain $T = 1250$ channel uses.

Albeit suboptimal, zero-forcing (ZF) beamforming with $K = N_t$ users captures the fundamental trend in terms of degrees of freedom (or "multiplexing gain") [2]. Therefore, we focus on this case for its analytical tractability. In order to perform ZF beamforming (or any other multiuser MIMO precoding), the BS must have an accurate estimate of the downlink channel. Such information, referred to as the *Channel State Information at the Transmitter* (CSIT) is acquired by using downlink training and channel state feedback. On the one hand, in TDD systems with self-calibrating devices, owing to the fact that uplink and downlink take place in the same channel coherence bandwidth, CSIT can be acquired directly from the uplink pilot symbols. On the other hand, the uplink-downlink channel reciprocity does not hold in Frequency-Division Duplexing (FDD) systems where uplink and downlink take places in different widely separated frequency bands. This is also the case in Time-Division Duplexing (TDD) systems where uplink and downlink may time-share the same band but the non-linear devices are not self-calibrated and therefore induce non-reciprocal effects. In the latter case, an explicit CSIT feedback must be used. In any case, the rates achievable with ZF beamforming depend critically on the quality of the CSIT, however, high quality CSIT can be achieved by dedicating a significant amount of time resource to downlink training and (for FDD) to channel state feedback. It follows that there is a non-trivial tradeoff between the benefits of improving the CSIT and the overhead in channel estimation and feedback.

In this work, we determine the optimum fraction of resources that should be dedicated to training/feedback in several cases of interest. In particular, we consider three time-frequency block models depicted in Fig. 1. These models can be viewed as an idealization of the actual systems such as LTE [3] and aim at capturing the essential features. In Section III, we consider the optimization of the net spectral efficiency based on model 1 where both training and feedback

consume "downlink" channel uses. This analysis applies naturally to TDD with or without reciprocity and FDD where downlink training and (uplink) feedback are performed in the same fading coherence block, via some hand-shaking protocol. In Section IV, we consider a different viewpoint based on models 2 and 3, in which the CSIT feedback consumes "uplink" channel uses. These models are more relevant to FDD systems. The question that we address is "how much uplink resource should one pay in order to achieve a certain downlink spectral efficiency?". By solving the corresponding optimization problem, we characterize the uplink/downlink spectral efficiency region. At which point of this tradeoff region the system should operate is a function of the specific system requirements such as uplink/downlink traffic demands. For a fixed demand, the optimal operation point can be found adaptively. Further, we study the effect of temporally correlated fading channels and feedback delay where CSIT is obtained through a one-step prediction model (model 3). This corresponds to the case when the downlink block bandwidth W_f and the block length T_f are significantly shorter than the coherence bandwidth W_c and the coherence interval T_c, respectively. Finally, Section V presents some considerations for the case of $K > N_t$ users with some downlink scheduling and user selection [4]. This case is very relevant in practice, but its analysis has escaped so far a full closed-form characterization. Therefore, we provide results by combining Monte Carlo simulation and closed-form analysis.

The optimization of training has been studied in the context of point-to-point MIMO channels in the literature, e.g., [5], [6], [7], [8], [9]. In [5], the point-to-point MIMO communication is considered and only downlink training is addressed for the case of no CSIT and imperfect Channel State Information at the Receiver (CSIR). On the other hand, in [7], perfect CSIR is assumed and the resources to be used for channel feedback are investigated. In [8], [9] the model of [5] is extended to also incorporate quantized channel feedback and transmitter beamforming. Although the setup is quite similar to ours, the emphasis of [8], [9] on the asymptotic regime, where the number of antennas and T are simultaneously taken to infinity, leads to rather different conclusions as compared to the present work. In [6], a MIMO broadcast with downlink training and perfect channel feedback (i.e., the BS is also able to view the received training symbols) is considered. It is shown that the sum rate achievable with a dirty paper coding-based strategy has a very similar form to the achievable rate expressions in [5], and thus many of the conclusions from [5] directly carry over. On the other hand, we consider the more practical case where there is imperfect feedback from each UT to the BS and also study achievable rates with ZF beamforming,

which has lower complexity than dirty-paper coding. The present work is an extension of [10], [11], where the same optimization was investigated assuming that both downlink training and uplink feedback are performed within the same block (model 1). In this paper, we provide more complete guidelines on the overall system optimization for the various scenarios of interest.

II. Channel State Estimation and Feedback

When the multiuser MIMO downlink is operated in a closed-loop mode, the CSIT is obtained through the following phases:

1) Common downlink training: $T_{\rm tr}$ shared pilot symbols (i.e., $\frac{T_{\rm tr}}{N_t}$ pilots per BS antenna) are transmitted on each channel coherence block to allow all UTs to estimate their downlink channel vectors $\{\mathbf{h}_k\}$ based on the observation

$$\mathbf{s}_k = \sqrt{\frac{T_{\rm tr}P}{N_t}}\ \mathbf{h}_k + \mathbf{z}_k. \tag{2}$$

Using linear MMSE estimation, the per-coefficient estimation error variance is given by

$$\frac{1}{1+\left(\frac{T_{\rm tr}}{N_t}\right)\rho} \tag{3}$$

2) Channel feedback: Each UT feeds back its channel estimation immediately after the training phase. We focus on the scenario where the feedback channel is modeled as an AWGN channel with the SNR ρ, identical to the nominal downlink SNR. Because UT's are assumed to access the feedback channel orthogonally, a total of $T_{\rm fb}$ channel symbols translates into $\frac{T_{\rm fb}}{N_t}$ feedback channel uses per UT. Different feedback strategies are described in Section III.

The BS obtains the channel state matrix $\widehat{\mathbf{H}} = [\widehat{\mathbf{h}}_1, \ldots, \widehat{\mathbf{h}}_{N_t}]$ based on the training/feedback information. Errors in the CSIT available to the BS stems from two sources: the channel estimation error during the common training phase, and the distortion incurred during the feedback phase. Then, the BS computes the ZF beamforming vector $\widehat{\mathbf{v}}_k$ to be a unit-norm vector orthogonal to the subspace $\mathcal{S}_k = \mathrm{span}\{\widehat{\mathbf{h}}_j^{\mathsf{H}} : j \neq k\}$ for all k. In this case, the ergodic rate achievable by UT k with equal-power allocation across UT's and Gaussian random coding is given by:

$$R_k = \mathbb{E}\left[\log\left(1+\frac{|\mathbf{h}_k^{\mathsf{H}}\widehat{\mathbf{v}}_k|^2\frac{\rho}{N_t}}{1+\frac{\rho}{N_t}\sum_{j\neq k}|\mathbf{h}_k^{\mathsf{H}}\widehat{\mathbf{v}}_j|^2}\right)\right], \tag{4}$$

assuming each UT is aware of its received signal-to-interference plus noise ratio (SINR).[1] The residual interference due to non-zero "leakage" coefficients $\{|\mathbf{h}_k^{\mathsf{H}}\hat{\mathbf{v}}_j|\}$ decreases the achievable rate. In [12], it is shown that the rate in (4) is tightly lower-bounded by

$$R_k \geq R_k^{\mathrm{ZF}} - \overline{\Delta R}_k \tag{5}$$

where R_k^{ZF} is the rate achievable with perfect CSIT and $\overline{\Delta R}_k$ denotes the *rate gap*, given in closed form by

$$\overline{\Delta R}_k \triangleq \log\left(1 + \frac{\rho}{N_t}\sum_{j\neq k}\mathbb{E}\left[|\mathbf{h}_k^{\mathsf{H}}\hat{\mathbf{v}}_j|^2\right]\right). \tag{6}$$

Assuming that the channel statistics are symmetric over users and space, R_k, R_k^{ZF} and $\overline{\Delta R}_k$ do not depend on k, therefore the subscript k will be omitted in the following. The rate gap depends on T_{tr}, T_{fb} and the training/feedback strategy and will be generally denoted by the function $\overline{\Delta R}(T_{\mathrm{tr}}, T_{\mathrm{fb}})$. Explicit expressions are found in [12] for the cases addressed in this paper.

III. Joint Optimization of Training and Feedback

In this section, we focus on model 1 of Fig. 1 where training and CSIT feedback consume downlink channel uses. Model 1 (a) refers to the TDD system exploiting the channel reciprocity, while model 1 (b) refers to either the TDD without reciprocity or the FDD system in which the downlink training and the feedback are performed in the same fading coherence block. In both cases, the maximization of the *net* downlink spectral efficiency is formulated as

$$\max_{T_{\mathrm{tr}},T_{\mathrm{fb}}:T_{\mathrm{tr}}+T_{\mathrm{fb}}\leq T}\left(1-\frac{T_{\mathrm{tr}}+T_{\mathrm{fb}}}{T}\right)\left(R^{\mathrm{ZF}} - \overline{\Delta R}(T_{\mathrm{tr}}, T_{\mathrm{fb}})\right). \tag{7}$$

It is convenient to consider the maximization in two steps, by writing:

$$\max_{T_t\leq T}\ \max_{T_{\mathrm{tr}}+T_{\mathrm{fb}}=T_t}\left(1-\frac{T_{\mathrm{tr}}+T_{\mathrm{fb}}}{T}\right)\left(R^{\mathrm{ZF}} - \overline{\Delta R}(T_{\mathrm{tr}}, T_{\mathrm{fb}})\right). \tag{8}$$

Furthermore, the rate gap can be put in the general form (see [12])

$$\overline{\Delta R}(T_{\mathrm{tr}}, T_{\mathrm{fb}}) = \log\left(1 + g(T_{\mathrm{tr}}, T_{\mathrm{fb}})\right) \tag{9}$$

[1]Such knowledge can be acquired through an additional dedicated training round as discussed in [12]. This training round does not significantly affect the present work, and thus is ignored for the sake of simplicity.

where the function $g(\cdot,\cdot)$ depends on the feedback strategy and shall be specified later. Because the first multiplicative term is constant when $T_{\rm tr}+T_{\rm fb}=T_t$, the inner maximization corresponds to minimization of the function $g(\cdot,\cdot)$, subject to the constraint $T_{\rm tr}+T_{\rm fb}\leq T_t$. Letting $g(T_t)\triangleq \min_{T_{\rm tr}+T_{\rm fb}\leq T_t} g(T_{\rm tr},T_{\rm fb})$ denote the solution of the inner maximization in (8), we can solve the outer maximization by searching for the optimal value $0<T_t\leq T$.

A. TDD with channel reciprocity

When channel reciprocity holds, open-loop CSIT estimation can be obtained from the uplink pilot symbols. In this case, the amount of uplink training can be optimized as a special case of (8) where no CSIT feedback is used.[2] In [12, Remark 4.2], the rate gap for a TDD system that uses $T_{\rm tr}$ *uplink* training symbols is given by:

$$\overline{\Delta R}=\log\left(1+\frac{N_t-1}{T_{\rm tr}}\right) \tag{10}$$

which corresponds to $g^{\rm tdd}(T_{\rm tr})=\frac{\theta_{\rm tr}}{T_{\rm tr}}$ with $\theta_{\rm tr}=N_t-1$. Plugging this into (7), we maximize the net spectral efficiency given by

$$f(T_{\rm tr})=\left(1-\frac{T_{\rm tr}}{T}\right)\left[R^{\rm ZF}-\log\left(1+\frac{\theta_{\rm tr}}{T_{\rm tr}}\right)\right]. \tag{11}$$

Because $f(\cdot)$ is concave in $T_{\rm tr}$, the optimal $T_{\rm tr}^\star$ can be found by numerically solving for $\frac{\partial f}{\partial T_{\rm tr}}=0$ where

$$\frac{\partial f}{\partial T_{\rm tr}} \;=\; \frac{\theta_{\rm tr}\left(1-\frac{T_{\rm tr}}{T}\right)}{T_{\rm tr}^2\left(1+\frac{\theta_{\rm tr}}{T_{\rm tr}}\right)}-\frac{1}{T}\left[R^{\rm ZF}-\log\left(1+\frac{\theta_{\rm tr}}{T_{\rm tr}}\right)\right]. \tag{12}$$

Although a closed-form solution for $T_{\rm tr}^\star$ cannot be found, we can study the scaling of the optimal $T_{\rm tr}^\star$ with the system parameters. It is not difficult to see that the derivative in (12) is upperbounded by $\frac{1}{T}\widetilde{f}(T_{\rm tr})$, where

$$\widetilde{f}(T_{\rm tr})=\frac{\theta_{\rm tr}\left(T-T_{\rm tr}\right)}{T_{\rm tr}^2}-\left[R^{\rm ZF}-\frac{\theta_{\rm tr}}{T_{\rm tr}}\right] \tag{13}$$

The concavity of $f(\cdot)$ implies that the solution $\widetilde{T}_t$ of the equation $\widetilde{f}(T_{\rm tr})=0$ is an upper bound to the optimal value $T_t^\star$. Solving for $\widetilde{f}(T_{\rm tr})=0$, we find

$$T_{\rm tr}^\star\leq\widetilde{T}_{\rm tr}=\sqrt{\frac{\theta_{\rm tr}T}{R^{\rm ZF}}}. \tag{14}$$

[2]Note that a similar optimization is considered in [13], although in that work analysis of this optimization is not performed.

Furthermore, when the rate gap is small such that $\log\left(1+\frac{\theta_{\rm tr}}{T_{\rm tr}}\right) \approx \frac{\theta_{\rm tr}}{T_{\rm tr}}$ (which becomes accurate for large T), the upperbound also becomes a very good approximation.

Two interesting behaviors are obtained from (21): 1) for a fixed SNR (i.e., constant $R^{\rm ZF}$) $T_{\rm tr}^{\star}$ increases as $O(\sqrt{T})$ as $T \to \infty$; 2) for a fixed block length T, $T_{\rm tr}^{\star}$ decreases as $O(1/\sqrt{R^{\rm ZF}})$ for large SNR, or equivalently, it decreases as $O(1/\sqrt{\log(\rho)})$ since $R^{\rm ZF} = \log(\rho) + O(1)$ for large SNR.

Next, we examine the impact of $T_{\rm tr}^{\star}$ on the net achievable rate. By the definition of $T_{\rm tr}^{\star}$ we have:

$$f\left(T_{\rm tr}^{\star}\right) \geq f\left(\widetilde{T}_{\rm tr}\right) = \left(1-\sqrt{\frac{\theta_{\rm tr}}{R^{\rm ZF}T}}\right)\left[R^{\rm ZF} - \log\left(1+\sqrt{\frac{\theta_{\rm tr}R^{\rm ZF}}{T}}\right)\right] \tag{15}$$

The rate gap with respect to $R^{\rm ZF}$ can therefore be upper bounded as:

$$R^{\rm ZF} - f\left(T_{\rm tr}^{\star}\right) \leq R^{\rm ZF} - f\left(\widetilde{T}_t\right) \tag{16}$$

$$= \sqrt{\frac{\theta_{\rm tr}R^{\rm ZF}}{T}} + \log\left(1+\sqrt{\frac{\theta_{\rm tr}R^{\rm ZF}}{T}}\right) - \sqrt{\frac{\theta_{\rm tr}}{R^{\rm ZF}T}}\log\left(1+\sqrt{\frac{\theta_{\rm tr}R^{\rm ZF}}{T}}\right) \tag{17}$$

$$\leq 2\sqrt{\frac{\theta_{\rm tr}R^{\rm ZF}}{T}} \tag{18}$$

where the final inequality is reached by dropping the last term in (17) and using $\log(1+x) \leq x$. Thus, the gap to a perfect CSIT system decreases roughly as $O(1/\sqrt{T})$ as T increases.

For a future reference, it is worthwhile to notice that model 1 (a) corresponds to model 1 (b) with perfect feedback such that the BS knows the UT channel estimates. As a result, the net rate achievable with TDD, channel reciprocity and open-loop CSIT estimation serves as an upper bound to the rate achievable with *any* form of CSIT feedback considered in the following.

B. Analog Feedback

An option for the CSIT feedback scheme consists of sending the channel coefficients as QAM unquantized modulation symbols. This is usually referred to as "analog feedback" in the literature, since the scheme is indeed akin to analog amplitude/phase modulation. Because each UT is allowed $\frac{T_{\rm fb}}{N_t}$ feedback channel uses, this scheme transmits each channel coefficient over $\frac{T_{\rm fb}}{N_t^2}$ feedback channel uses (if $T_{\rm fb} > N_t^2$, each coefficient is effectively repeated $\frac{T_{\rm fb}}{N_t^2}$ times on

the feedback channel). At the BS receiver, MMSE estimation is used. The resulting rate gap is described as [12, Section IV] and results in the $g(\cdot,\cdot)$ function

$$g^{\text{analog}}(T_{\text{tr}}, T_{\text{fb}}) = \frac{N_t - 1}{T_{\text{tr}}} + \frac{N_t(N_t-1)}{T_{\text{fb}}}. \tag{19}$$

For the sake of generality, we consider a generalized form of (19) as $g^{\text{analog}}(T_{\text{tr}}, T_{\text{fb}}) = \frac{\theta_{\text{tr}}}{T_{\text{tr}}} + \frac{\theta_{\text{fb}}}{T_{\text{fb}}}$, for two non-negative weights θ_{tr} and θ_{fb}. Comparing (19) with (10), we notice that the previous TDD open-loop case corresponds to letting $\theta_{\text{fb}} = 0$, consistently with the fact that in this case no CSIT feedback is used.

It is immediate to check that the minimization of $g^{\text{analog}}(T_{\text{tr}}, T_{\text{fb}})$ subject to $T_{\text{tr}} + T_{\text{fb}} = T_t$, and to $T_{\text{tr}}, T_{\text{fb}} \geq 0$ is a convex problem. The corresponding Lagrangian [14] is given by

$$\mathcal{L}(T_{\text{tr}}, T_{\text{fb}}, \mu) = g(T_{\text{tr}}, T_{\text{fb}}) + \frac{1}{\mu^2}(T_{\text{tr}} + T_{\text{fb}})$$

where $\mu > 0$ is the Lagrangian multiplier for the equality constraint. The KKT conditions [14] yield the solution $T_{\text{tr}}^\star = \sqrt{\theta_{\text{tr}}}\mu$ and $T_{\text{fb}}^\star = \sqrt{\theta_{\text{fb}}}\mu$. Imposing the equality constraint and eliminating μ, we obtain:

$$T_{\text{tr}}^\star = \sqrt{\frac{\theta_{\text{tr}}}{\mathcal{K}}}T_t, \quad T_{\text{fb}}^\star = \sqrt{\frac{\theta_{\text{fb}}}{\mathcal{K}}}T_t \tag{20}$$

where we let $\mathcal{K} = (\sqrt{\theta_{\text{tr}}} + \sqrt{\theta_{\text{fb}}})^2$, and the resulting objective value is given by $g^{\text{analog}}(T_t) = \frac{\mathcal{K}}{T_t}$.

The outer optimization (step 2) is now characterized in terms of a single variable T_t and reduces to the maximization of (11) where we replace T_{tr} and θ_{tr} by T_t and $\mathcal{K}$, respectively. As a result, we find the optimal scaling for T_t as

$$T_t^\star \leq \widetilde{T}_t = \sqrt{\frac{\mathcal{K}T}{R^{\text{ZF}}}}. \tag{21}$$

Hence, the same analysis holds for the total length $T_t^\star$ of training and feedback. In addition, the following upper bound on $T_{\text{tr}}^\star$ can be obtained by combining (21) with (20)

$$T_{\text{tr}}^\star \leq \sqrt{\frac{\theta_{\text{tr}}}{\mathcal{K}}}\widetilde{T}_t = \sqrt{\frac{\theta_{\text{tr}}T}{R^{\text{ZF}}}} = \sqrt{\frac{(N_t-1)T}{R^{\text{ZF}}}}. \tag{22}$$

According to this upperbound, the optimal downlink training is independent of θ_{fb}, and thus of the efficiency of the feedback channel.

Similarly, we obtain the effective rate gap with respect to R^{ZF} as

$$R^{\text{ZF}} - f\left(T_{\text{tr}}^\star\right) \leq 2\sqrt{\frac{\mathcal{K}R^{\text{ZF}}}{T}} \tag{23}$$

Comparing this and the corresponding expression (18) for the open-loop TDD, we see that the analog feedback incurs a rate gap increase by a factor $1 + \sqrt{N_t}$.

C. Error-Free Digital Feedback

We now analyze a digital feedback technique where each UT quantizes its estimated channel vector into a B-bits message and then maps these bits into $\frac{T_{\rm fb}}{N_t}$ transmit symbols. For the quantization step we consider an ensemble of random vector quantizers (RVQ) with directional quantization as described in [15]. Assuming the feedback messages are received error-free, in [12, Section V] it is shown that the rate gap is given by

$$\overline{\Delta R} = \log\left(1 + \frac{N_t - 1}{T_{\rm tr}} + \rho\ 2^{-\frac{B}{N_t - 1}}\right). \tag{24}$$

For the time being, we assume unrealistically that error-free communication is possible over the feedback channel at a rate equal to its capacity of $\log_2(1+\rho)$ bits per channel use. Letting $B = \frac{T_{\rm fb}}{N_t}\log_2(1+\rho)$, we obtain

$$g^{\rm digital}(T_{\rm tr}, T_{\rm fb}) = \frac{N_t - 1}{T_{\rm tr}} + \rho\,(1+\rho)^{-\frac{T_{\rm fb}}{N_t(N_t-1)}}. \tag{25}$$

Following the two-step approach, we minimize the above function subject to $T_{\rm tr} + T_{\rm fb} = T_t$. Since $g^{\rm digital}(\cdot,\cdot)$ is convex in $T_{\rm tr}, T_{\rm fb}$, we form the Lagrangian and readily obtain

$$T_{\rm tr} \;=\; \mu\sqrt{N_t - 1},\;\; T_{\rm fb} = N_t(N_t-1)\frac{2\ln(\mu) + \ln\left(\frac{\rho\ln(1+\rho)}{N_t(N_t-1)}\right)}{\ln(1+\rho)} \tag{26}$$

where $\mu > 0$ is chosen so that the equality constraint is fulfilled. Note that $T_{\rm fb}$ grows as $O(\ln\mu)$, much slower than the linear increase (in μ) for $T_{\rm tr}$.

Contrary to the earlier analog feedback case, we cannot express $g^{\rm digital}(T_t)$ in a simple closed form. However, using (26) we can eliminate μ and express $T_{\rm fb}$ as a function of $T_{\rm tr}$:

$$T_{\rm fb} = N_t(N_t-1)\frac{2\ln(T_{\rm tr}) + \ln\left(\frac{\rho\ln(1+\rho)}{N_t(N_t-1)^2}\right)}{\ln(1+\rho)}, \tag{27}$$

and thus the net spectral efficiency can be written as:

$$\left(1 - \frac{T_{\rm tr} + N_t(N_t-1)\frac{2\ln(T_{\rm tr}) + \ln\left(\frac{\rho\ln(1+\rho)}{N_t(N_t-1)^2}\right)}{\ln(1+\rho)}}{T}\right) \times \left[R^{\rm ZF} - \log\left(1 + \tfrac{N_t-1}{T_{\rm tr}} + \tfrac{N_t(N_t-1)^2}{(T_{\rm tr})^2\ln(1+\rho)}\right)\right].$$

Because $T_{\rm fb}$ increases logarithmically in $T_{\rm tr}$, and decreases with the SNR ρ, its effect on the maximization is rather negligible. As a result, the maximization of $T_{\rm tr}$ is very similar to the case of TDD with channel reciprocity. In other words, the error-free digital feedback performs almost as good as the TDD open-loop upper bound.

D. Digital Feedback with Errors

We consider a practical digital feedback scheme with a very low complexity. In particular, we assume that the B feedback bits are transmitted on the uplink by using uncoded QAM. Each UT makes use of $\frac{T_{\rm fb}}{N_t}$ feedback channel uses for its CSIT feedback. Assuming that quantization bits are arbitrarily mapped to the QAM constellation symbols, the error of any symbol renders the feedback from a particular UT effectively useless and thus leads to a zero rate.[3] Under this assumption, the achievable net spectral efficiency is given as a solution to

$$\max_{T_{\rm tr},T_{\rm fb}:T_{\rm tr}+T_{\rm fb}\leq T} \left(1-\frac{T_{\rm tr}+T_{\rm fb}}{T}\right)(1-P_{e,\rm fb})\left[R_k^{\rm ZF}-\overline{\Delta R}\right] \tag{28}$$

where $\overline{\Delta R}$ is defined in (24) and where $P_{e,\rm fb}$ is the feedback message error probability. The size of the QAM constellation is given by $M=2^{\frac{BN_t}{T_{\rm fb}}}$ and yields a symbol error probability [16]

$$P_s = 1-\left(1-2\left(1-\frac{1}{\sqrt{M}}\right)Q\left(\sqrt{\frac{3\rho}{M-1}}\right)\right)^2, \tag{29}$$

and a corresponding feedback message error probability

$$P_{e,\rm fb} = 1-(1-P_s)^{\frac{T_{\rm fb}}{N_t}}. \tag{30}$$

Following the two-step optimization approach, we rewrite the outer optimization as

$$\max_{T_{\rm t}\leq T}\left(1-\frac{T_t}{T}\right)\left[R^{\rm ZF}-\widetilde{\Delta R}(T_t)\right] \tag{31}$$

where the effective rate-loss $\widetilde{\Delta R}(T_t)$, incorporating the loss due to erroneous feedback, is given by

$$\widetilde{\Delta R}(T_t) \;=\; \min_{T_{\rm tr}+T_{\rm fb}=T_t}\; \left\{(1-P_{e,\rm fb})\log\left(1+\frac{N_t-1}{T_{\rm tr}}+\rho M^{-\frac{T_{\rm fb}}{N_t(N_t-1)}}\right)+P_{e,\rm fb}R_k^{\rm ZF}\right\}. \tag{32}$$

If the QAM constellation size is suitably optimized, the probability of feedback error can be made sufficiently small when the number of feedback bits $\frac{BN_t}{T_{\rm fb}}$ per user is large. For example, for $N_t=4$ at 10 dB with $B=25$ bits and 4-QAM, we have $P_{e,\rm fb}=0.0194$. As a result, the minimization in (32) is very similar to the minimization of $g^{\rm digital}(T_{\rm tr},T_{\rm fb})$ for error-free feedback in (25).

We conclude this section by providing some numerical examples to compare the performance of different feedback strategies. In Fig. 2 the optimal values of $T_{\rm tr}$ and $T_{\rm fb}$ are plotted versus block

[3]This point can be made rigorous, but we limit ourselves to the present intuitive argument for the sake of space limitation.

length T for analog feedback, error-free digital feedback, and QAM-based digital feedback along with the uplink training length $T_{\rm tr}$ for the TDD system. Most striking is the fact that the optimal values of $T_{\rm tr}$ are essentially identical for the three feedback techniques as well as for TDD. Furthermore, although not shown here, the optimal values of $T_{\rm tr}$ are very well approximated by $\sqrt{\frac{(N_t-1)T}{R^{\rm ZF}}}$ as in (22). The number of feedback symbols, however, depends critically on the feedback method. Because analog feedback is so inefficient, a large number of feedback symbols are used so that the rate gap due to feedback is minimized. On the other hand, digital feedback is very efficient and a relatively small number of feedback symbols is required.

In Fig. 3, the sum spectral efficiency is plotted versus block length T. Although not shown here, the rate approximations based upon (18) are seen to become increasingly accurate as T increases for analog and TDD. Analog feedback is outperformed by digital feedback with or without errors, for any T. This is because digital feedback offers a significantly smaller distortion as compared to analog whenever $T_{\rm fb}$ is larger than (approximately) N_t^2 (i.e., one symbol per channel coefficient) [12, Section VI], and for reasonable block lengths it is optimal to use $T_{\rm fb}$ considerably larger than N_t^2 (see Fig. 2).

IV. Separate Uplink and Downlink Bandwidths

In FDD systems, the uplink and downlink bandwidths are generally separated and the amount of channel uses per block length dedicated to the CSIT feedback impacts the uplink spectral efficiency as an overhead, rather than the downlink as in the previous section. In this section we focus on models 2 and 3 of Fig. 1 assuming that the downlink and uplink bandwidths are a priori fixed. The challenge here consists of determining the tradeoff region of downlink spectral efficiency versus uplink CSIT feedback overhead.

For this purpose, we consider the net downlink spectral efficiency, accounting for the training overhead, as a function of $T_{\rm fb}$. For each value of $T_{\rm fb}$, the optimal number of downlink training symbols is found, and the corresponding net downlink spectral efficiency is given by:

$$w\left(T_{\rm fb}\right) \triangleq \max_{T_{\rm tr}\leq T} \left(1-\frac{T_{\rm tr}}{T}\right)\left(R^{\rm ZF}-\log\left(1+\frac{N_t-1}{T_{\rm tr}}+\Delta(T_{\rm fb})\right)\right) \tag{33}$$

where $\Delta(T_{\rm fb})$ denotes the loss term due to CSIT feedback. By solving for the maximization with respect to $T_{\rm tr}$, we obtain a tight lower bound on the optimal downlink spectral efficiency achievable with ZF beamforming as a function of the parameter $T_{\rm fb}$, that quantifies the number of channel uses per block spent for the CSIT feedback over the uplink.

In the following, we first characterize such a tradeoff for the cases of the AWGN feedback channel based on model 2. Then, we address the case of a temporally correlated channel with feedback delay and channel prediction by considering model 3.

A. AWGN feedback link

For the orthogonal access over the AWGN feedback channel, we have $\Delta(T_{\rm fb}) = \frac{N_t(N_t-1)}{T_{\rm fb}}$ for analog feedback, or $\Delta(T_{\rm fb}) = \rho\,(1+\rho)^{-\frac{T_{\rm fb}}{N_t(N_t-1)}}$ for error-free digital feedback (see (19) and (25)). As seen previously, the effect of feedback errors can be made sufficiently small even by very simple schemes based on uncoded QAM modulation. Hence, due to the space limitation, we provide only the analysis for the case of error-free digital feedback operating at the uplink AWGN capacity, which captures the essential behavior of digital feedback while allowing for much simpler analytical expressions. Nevertheless, in the numerical results we provide also the results for a 4QAM-based digital feedback for the sake of comparison.

By simple manipulation, the objective function can be rewritten as:

$$\left(1-\frac{T_{\rm tr}}{T}\right)\left(R^{\rm ZF}-\log\left(1+\Delta(T_{\rm fb})\right)-\log\left(1+\frac{N_t-1}{T_{\rm tr}(1+\Delta(T_{\rm fb}))}\right)\right). \tag{34}$$

Hence, the optimization has the same form as in Section III-A, with $R^{\rm ZF}$ replaced by $R^{\rm ZF}-\log\left(1+\Delta(T_{\rm fb})\right)$ and N_t-1 replaced by $\frac{N_t-1}{1+\Delta(T_{\rm fb})}$. It follows that we can immediately write the bound on the optimal training length as

$$T_{\rm tr}^{\star}(T_{\rm fb}) \le \widetilde{T}_{\rm tr}(T_{\rm fb}) = \sqrt{\frac{(N_t-1)T}{\left(R^{\rm ZF}-\log\left(1+\Delta(T_{\rm fb})\right)\right)\left(1+\Delta(T_{\rm fb})\right)}}. \tag{35}$$

Although $T_{\rm tr}^{\star}(T_{\rm fb})$ does depend on $T_{\rm fb}$, this dependency is very weak whenever $T_{\rm fb}$ is not too small. Thus, very little is lost by simply choosing $T_{\rm tr} = \sqrt{\frac{(N_t-1)T}{R^{\rm ZF}}}$.

Using the same arguments as in Section III-A, the downlink spectral efficiency can be lower bounded by

$$w\left(T_{\rm fb}\right) \ge \left(1-\sqrt{\frac{N_t-1}{TR^{\rm ZF}}}\right)\left(R^{\rm ZF}-\log\left(1+\sqrt{\frac{R^{\rm ZF}(N_t-1)}{T}}+\Delta(T_{\rm fb})\right)\right) \tag{36}$$

$$\ge R^{\rm ZF}-2\sqrt{\frac{R^{\rm ZF}(N_t-1)}{T}}-\left(\frac{1-\sqrt{\frac{N_t-1}{TR^{\rm ZF}}}}{1+\sqrt{\frac{R^{\rm ZF}(N_t-1)}{T}}}\right)\Delta(T_{\rm fb}). \tag{37}$$

Using the expressions for $\Delta(T_{\rm fb})$ we have:

$$w^{\text{analog}}\left(T_{\rm fb}\right) \geq R^{\rm ZF} - 2\sqrt{\frac{R^{\rm ZF}(N_t-1)}{T}} - \left(\frac{1-\sqrt{\frac{(N_t-1)}{TR^{\rm ZF}}}}{1+\sqrt{\frac{R^{\rm ZF}(N_t-1)}{T}}}\right)\frac{N_t(N_t-1)}{T_{\rm fb}} \tag{38}$$

$$w^{\text{digital}}\left(T_{\rm fb}\right) \geq R^{\rm ZF} - 2\sqrt{\frac{R^{\rm ZF}(N_t-1)}{T}} - \left(\frac{1-\sqrt{\frac{(N_t-1)}{TR^{\rm ZF}}}}{1+\sqrt{\frac{R^{\rm ZF}(N_t-1)}{T}}}\right)\rho\left(1+\rho\right)^{-\frac{T_{\rm fb}}{N_t(N_t-1)}}. \tag{39}$$

Notice that the spectral efficiency penalties due to training and feedback are separable in these lower bounds. Based upon these expressions, we expect that the downlink spectral efficiency $w^{\text{digital}}(T_{\rm fb})$ with digital feedback converges very quickly to the rate accounting for the optimized training overhead, which is approximately $R^{\rm ZF} - 2\sqrt{\frac{R^{\rm ZF}(N_t-1)}{T}}$, whereas convergence is much slower with analog feedback.

The above definitions of w^{analog} and w^{digital} characterize the net downlink spectral efficiency as a function of the number of uplink symbols per block length used for CSIT feedback. In terms of system design, it is more meaningful to characterize the downlink *rate* as a function of the uplink *bandwidth* used for channel feedback. Under the block-fading model adopted in this paper, the channel is constant for T_c seconds over the bandwidth of W_c. Since $T_{\rm fb}$ uplink symbols are used for channel feedback for every block, the uplink bandwidth used for channel feedback is given by $\frac{T_{\rm fb}}{T_c}$ Hz, and the downlink rate is given by $W_c w(T_{\rm fb})$ in bit/sec (bps).

We can take advantage of the above analysis to understand the fundamental tradeoff between downlink and uplink rate. To this end, we employ a simplistic model of the uplink in which we assume the uplink bandwidth of $W_{\rm up}$ Hz and the uplink spectral efficiency of $C_{\rm up}$ bps/Hz. Since feedback consumes $\frac{T_{\rm fb}}{T_c}$ Hz of uplink bandwidth, the remaining bandwidth of $W_{\rm up} - \frac{T_{\rm fb}}{T_c}$ Hz is available for uplink data transmission. Thus the uplink data *rate* is

$$R_{\rm up}(T_{\rm fb}) = \left(W_{\rm up} - \frac{T_{\rm fb}}{T_c}\right)C_{\rm up}. \tag{40}$$

while the downlink *rate* is

$$R_{\rm down}(T_{\rm fb}) = W_c w\left(T_{\rm fb}\right). \tag{41}$$

As $T_{\rm fb}$ increases, the downlink rate $R_{\rm down}$ increases at the expense of decreasing uplink rate $R_{\rm up}$. In order to determine the operating point on the $(R_{\rm down}, R_{\rm up})$ Pareto-optimal boundary, a common method consists of maximizing the weighted sum of rates:

$$\max_{T_{\rm fb}} \quad \lambda R_{\rm down}(T_{\rm fb}) + \overline{\lambda} R_{\rm up}(T_{\rm fb}) \tag{42}$$

where $0 < \lambda < 1$ and $\overline{\lambda} = 1 - \lambda$. This optimization is equivalent to

$$\max_{T_{\rm fb}} \quad \lambda W_c w\left(T_{\rm fb}\right) - \overline{\lambda}\left(\frac{T_{\rm fb}}{T_c}\right) C_{\rm up}. \tag{43}$$

After multiplying both sides by T_c and taking the derivative with respect to $T_{\rm fb}$, we see that the optimal solution satisfies:

$$\lambda T w'\left(T_{\rm fb}\right) = \overline{\lambda} C_{\rm up} \quad \rightarrow \quad w'\left(T_{\rm fb}\right) = \frac{1}{T}\frac{\overline{\lambda}}{\lambda} C_{\rm up}. \tag{44}$$

More precisely, we obtain the optimal $T_{\rm fb}$ as a function of λ as

$$T_{\rm fb}^{\rm analog}(\lambda) = \sqrt{\frac{r N_t(N_t-1)T\lambda}{C_{\rm up}\overline{\lambda}}} \tag{45}$$

$$T_{\rm fb}^{\rm digital}(\lambda) = \frac{N_t(N_t-1)}{\log(1+\rho)} \log\left(\frac{r\rho\log(1+\rho)T\lambda}{N_t(N_t-1)C_{\rm up}\overline{\lambda}}\right) \tag{46}$$

where we let $r = \frac{1-\sqrt{\frac{(N_t-1)}{T R^{\rm ZF}}}}{1+\sqrt{\frac{R^{\rm ZF}(N_t-1)}{T}}}$. Clearly the feedback length is non-negative and upper bounded by T. Compared to analog feedback, the feedback length $T_{\rm fb}^{\rm digital}(\lambda)$ with digital feedback is almost insensitive to λ except the corner points ($\lambda = 0, 1$). By plugging the above expressions into (38), (39), the achievable rate can be parameterized by λ such that

$$w^{\rm analog}(\lambda) \geq R^{\rm ZF} - 2\sqrt{\frac{R^{\rm ZF}(N_t-1)}{T}} - \sqrt{\frac{r N_t(N_t-1)C_{\rm up}\overline{\lambda}}{T\lambda}} \tag{47}$$

$$w^{\rm digital}(\lambda) \geq R^{\rm ZF} - 2\sqrt{\frac{R^{\rm ZF}(N_t-1)}{T}} - \frac{N_t(N_t-1)C_{\rm up}\overline{\lambda}}{T\lambda\log(1+\rho)}. \tag{48}$$

The third term, representing the rate loss due to the imperfect feedback, is rather marginal both for analog and digital feedback schemes for a large T in the range $0 < \lambda < 1$. From these expressions, it can be expected that the tradeoff curve with digital feedback is sharper and dominates the curve with analog feedback.

To make this discussion more concrete, consider a single resource block in LTE, with bandwidth 200 kHz and duration 1 ms, corresponding to $T = 200$ in our model. We assume $C_{\rm up} = 1.512$ bps/Hz (per user) and an uplink bandwidth also equal to 200 kHz, for the sake of symmetry. The uplink-downlink sum rate boundary (expressed in kbps) and the corresponding feedback lengths are shown in Figs. 4 and 5. A well-designed system will typically operate near the sharp "knee" of the curves of Fig. 4, where the downlink rate is very close to its maximum value. Fortunately, because of the relatively low cost of channel feedback, the uplink rate is also

reasonably close to its maximum. From Fig. 5 we remark also that analog feedback requires a longer $T_{\rm fb}$ for a larger weight λ while the feedback length with digital feedback is almost constant. The training length was found to be 24 symbols for any scheme except for $\lambda \approx 0$. Note that the choice $T = 200$ is quite conservative. As argued in Section I, typical physical channel parameters yield a significantly larger T for low mobility users.

The takeaway message of section is that, unless uplink data rate is very strongly preferred over downlink data rate, it is efficient to operate the system at a point where the downlink spectral efficiency is very close to the perfect-feedback case.

B. Delayed feedback channel

In this section we study the uplink/downlink tradeoff by taking into account the effect of the feedback delay and the temporally correlated channel based on model 3. This model is motivated by the following scenario. In practice, the downlink resource allocation blocks, i.e. the block bandwidth W_f and block length T_f, might be defined a priori independently of W_c and T_c, while these coherence parameters depend on the propagation environment as well as the users mobility and may even vary from user to user. For the case of a fixed block length $T = W_f T_f$ much shorter than $W_c T_c$, the channel coefficients in subsequent blocks are correlated.

In order to model such situation, we assume that the channel fading coefficients are constant within each block of T symbols and changes from block to block according to a stationary Gaussian random process with power spectral density (Doppler spectrum) $S_h(\xi)$, strictly band-limited in $[-F, F]$, where $F < 1/2$ is the maximum normalized Doppler frequency shift, given by $F = \frac{v f_c}{c} T_f$, where v is the mobile terminal speed (m/s), f_c is the carrier frequency (Hz), c is the light speed (m/s). Furthermore, such a "Doppler process" satisfies $\int_{-F}^{F} \log S_h(\xi) d\xi > -\infty$. This condition holds for most (if not all) channel models usually adopted in the wireless mobile communication literature (see [17] and references therein), where the Doppler spectrum has no spectral nulls within the support $[-F, F]$. Because of symmetry and spatial independence, we can neglect the antenna index and consider scalar rather than vector processes.

Contrary to the block-by-block estimation previously considered, each UT k estimates $\mathbf{h}_k(t)$ based on the observation $\{s_k(t-\tau) : \tau = d, d+1, \ldots, \infty\}$ available at UT k up to block $t-d$ where d denotes the feedback delay in blocks of length $W_f T_f$ and $s_k(t) = \sqrt{\frac{T_{\rm tr} P}{M}} h_k(t) + z_k(t)$ is the received signal at UT k at block t. We focus on the case of $d = 0$ (filtering) and $d = 1$

(prediction) in the following. The equivalent model for both cases is given by

$$h_k(t) = \tilde{h}_k(t) + n_k(t) \tag{49}$$

where $\tilde{h}_k(t) = \mathbb{E}[h_k(t)|\{s_k(t-\tau)\}]$ denotes the estimated channel, independent of the estimation error $n_k(t) \sim \mathcal{CN}(0, \sigma_{\text{tr}}^2)$. The one-step prediction MMSE $(d = 1)$ is given by [18], [12]

$$\epsilon_1(\delta) \;\; = \;\; \delta^{1-2F} \exp\left(\int_{-F}^{F} \log(\delta + S_h(\xi))d\xi\right) - \delta \tag{50}$$

where we assume a unit-power process, $\int_{-F}^{F} S_h(\xi)d\xi = 1$, observed in background white noise with per-component variance $\delta = \frac{N_t}{T_{\text{tr}}\rho}$. The filtering MMSE $(d = 0)$ is related to $\epsilon_1(\delta)$ through the well-known maximal ratio combining formula

$$\epsilon_0(\delta) \;\; = \;\; \frac{\delta\epsilon_1(\delta)}{\delta + \epsilon_1(\delta)}. \tag{51}$$

Since $\tilde{h}_k(t)$ and $n_k(t)$ are independent, we have $\mathbb{E}[|\tilde{h}_k(t)|^2] = 1 - \sigma_{\text{tr}}^2$ for any k.

In [12, Section VI. B], it is shown that the rate gap is upper bounded by

$$\Delta R^d \leq \log\left(1 + \frac{N_t - 1}{T_{\text{tr}}}\frac{\epsilon_d(\delta)}{\delta} + \Delta(T_{\text{fb}})\right). \tag{52}$$

For simplicity, we focus on the case of a uniform Doppler spectrum $S_h(\xi) = \frac{1}{2F}$ for $-F \leq \xi \leq F$. This yields

$$\frac{\epsilon_1(\delta)}{\delta} \;\; = \;\; \left(1 + \frac{1}{2F\delta}\right)^{2F} - 1 \leq \left(\frac{1}{2F\delta}\right)^{2F} \tag{53}$$

where the last inequality can be easily shown. Using (51) and (53) we obtain

$$\frac{\epsilon_0(\delta)}{\delta} \;\; \leq \;\; \frac{1}{1 + (2F\delta)^{2F}}. \tag{54}$$

Plugging these expressions into (52), we obtain the rate gap upper bounds as

$$\overline{\Delta R}^{d=0} \;\; = \;\; \log\left(1 + \Delta(T_{\text{fb}}) + \frac{N_t - 1}{T_{\text{tr}}}\frac{1}{1 + \left(\frac{2FN_t}{\rho T_{\text{tr}}}\right)^{2F}}\right) \leq \log\left(1 + \Delta(T_{\text{fb}}) + \frac{N_t - 1}{T_{\text{tr}}}\right) \tag{55}$$

$$\overline{\Delta R}^{d=1} \;\; = \;\; \log\left(1 + \Delta(T_{\text{fb}}) + \frac{N_t - 1}{T_{\text{tr}}}\left(\frac{\rho T_{\text{tr}}}{2FN_t}\right)^{2F}\right). \tag{56}$$

We observe that that for the case of filtering $(d = 0)$, the rate gap upper bound reduces to that of the AWGN feedback link for sufficiently large ρ. In what follows, we consider the more interesting case of one-step prediction.

We can again maximize the net downlink achievable spectral efficiency for the one-step prediction case by solving

$$w(T_{\rm fb}) \;\stackrel{\Delta}{=}\; \max_{T_{\rm tr}\geq N_t} \left(1-\frac{T_{\rm tr}}{T}\right)\left[R^{\rm ZF}(P)-\log\left(1+\kappa T_{\rm tr}^{2F-1}+\Delta(T_{\rm fb})\right)\right] \tag{57}$$

where we defined the constant $\kappa = (N_t-1)\left(\frac{\rho}{2FN_t}\right)^{2F}$. By letting the RHS of (57) denote $f(T_{\rm tr},T_{\rm fb})$, we remark that the objective function $f(\cdot,\cdot)$ is concave in $T_{\rm tr}$. The the optimal $T_{\rm tr}$ in (57) satisfies

$$\frac{\kappa(1-F)(T-T_{\rm tr})}{T_{\rm tr}^{2-F}(1+\kappa T_{\rm tr}^{1-F}+\Delta(T_{\rm fb}))} = R^{\rm ZF}-\log(1+\Delta(T_{\rm fb}))-\log\left(1+\frac{\kappa T_{\rm tr}^{-(1-F)}}{1+\Delta(T_{\rm fb})}\right). \tag{58}$$

Following the same arguments as before, it follows that the solution $\widetilde{T}_{\rm tr}$ to the equation $\tilde{f}(T_{\rm tr})=0$ is an upper bound to the optimal $T_{\rm tr}^{\star}$, where

$$\tilde{f}(T_{\rm tr}) = \frac{\kappa(T-T_{\rm tr})}{T_{\rm tr}^{2-F}(1+\Delta(T_{\rm fb}))} - \left[R^{\rm ZF}-\log(1+\Delta(T_{\rm fb}))-\frac{\kappa T_{\rm tr}^{-(1-F)}}{1+\Delta(T_{\rm fb})}\right] \tag{59}$$

Explicitly, we find

$$T_{\rm tr}^{\star}(T_{\rm fb}) \leq \widetilde{T}_{\rm tr}(T_{\rm fb}) \;=\; \left(\frac{(N_t-1)T}{(1+\Delta(T_{\rm fb}))\{R^{\rm ZF}-\log(1+\Delta(T_{\rm fb}))\}}\right)^{\frac{1}{2-F}}\left(\frac{\rho}{2FM}\right)^{\frac{2F}{2-F}}. \tag{60}$$

As T increases, the training length $T_{\rm tr}$ scales as $O(T^{\frac{1}{2-F}})$ depending on the Doppler frequency shift $0<F<\frac{1}{2}$. For a fixed T, the training length is increasing in F. When the fading is quasi-static (i.e., very low mobility users with $v\approx 0$) such that the channel becomes perfectly predictable, the training length coincides with the expression (35) for the block-by-block estimation. Since the term $\Delta(T_{\rm fb})$ is negligible for a sufficiently large $T_{\rm fb}$, we can choose with little loss of optimality

$$T_{\rm tr} = \left(\frac{\kappa T}{R^{\rm ZF}}\right)^{\frac{1}{2-F}} = \left(\frac{(N_t-1)T}{R^{\rm ZF}}\right)^{\frac{1}{2-F}}\left(\frac{\rho}{2FN_t}\right)^{\frac{2F}{2-F}}. \tag{61}$$

Following in the footsteps of what has been done before, we can obtain the lower bound of the downlink spectral efficiency as

$$\begin{aligned} w(T_{\rm fb}) &\geq \left(1-\frac{T_{\rm tr}}{T}\right)\left[R^{\rm ZF}-\log\left(1+\kappa T_{\rm tr}^{2F-1}+\Delta(T_{\rm fb})\right)\right] \\ &\geq R^{\rm ZF}-\left[\frac{T_{\rm tr}}{T}R^{\rm ZF}+(N_t-1)\left(\frac{\rho}{2FN_t}\right)^{2F}T_{\rm tr}^{2F-1}\right]-\frac{(1-\frac{T_{\rm tr}}{T})\Delta(T_{\rm fb})}{1+(N_t-1)\left(\frac{\rho}{2FN_t}\right)^{2F}T_{\rm tr}^{2F-1}} \end{aligned} \tag{62}$$

where we can replace $T_{\rm tr}$ by (61). Solving the weighted sum rate maximization, we obtain the optimal $T_{\rm fb}$ in the same form of (45) and (46), for analog feedback and error-free digital feedback, respectively, where the term r is now replaced by $\frac{1-\frac{T_{\rm tr}}{T}}{1+(N_t-1)\left(\frac{\rho}{2FN_t}\right)^{2F}T_{\rm tr}^{2F-1}}$.

In order to quantify the impact of the delay on the uplink-downlink tradeoff, Fig. 6 shows the uplink-downlink sum rate Pareto boundary for different mobile speeds $v = 6, 50, 80$ km/h yielding the Doppler shift of $F = 0.011, 0.093, 0.148$, respectively, with the same parameters as Fig. 4. The corresponding feedback length as a function of λ is shown in Fig. 7, where we only plotted for $v = 6, 80$ km/h for the sake of clarity. We recall that $\lambda = 1$ corresponds to the corner point $(R_{\rm down}, 0)$ while $\lambda = 0$ corresponds to the other corner point $(0, R_{\rm up})$. As expected from (61), the training length increases for a higher mobile speed and is found to be $25, 36, 43$ symbols for $v = 6, 50, 80$ km/h, respectively. On the contrary, the feedback length is rather indifferent to the mobile speed v, although it tends to decrease for a larger v. On the uplink-downlink tradeoff curve, the higher mobile speed decreases significantly the downlink rate since the larger training length incurs a significant rate loss.

Fig. 8 shows the achievable downlink sum rate in kbps versus the mobile speed v km/h when the uplink feedback length is set to $T_{\rm fb} = 30$ over a block length of $T = 200$ symbols. We compare analog feedback, error-free digital feedback as well as 4QAM-based digital feedback. It is observed that by dedicating 15% of the uplink resource to the feedback, the uncoded 4QAM outperforms the analog feedback.

V. Allowing for Many Users

We conclude this paper by providing a discussion on the relevant case of $K > N_t$. Until now we have assumed that the number of users is fixed equal to the number of BS antennas N_t. In a real system there are often more than N_t users (with data awaiting at the BS). If more users feedback their channel information, the BS can utilize user selection and generally obtain a non-negligible increase in downlink spectral efficiency. Of course, allowing additional users to feed back will incur a larger uplink bandwidth cost. Indeed, a well designed system should optimize not only the total number of feedback symbols used on the uplink, but also the number of users who feed back their channel state. When the number of users enters into the picture, we see that the uplink-downlink tradeoff, which appeared rather trivial for a fixed number of users, becomes indeed interesting and non-trivial.

Although the lower bound of [12] does not hold when user selection is performed, it can be numerically verified that it is nonetheless a reasonable approximation of the rate with user selection and imperfect CSIT. For the sake of the space limitation, we focus on the separate uplink/downlink bands (model 2) although the other models can be adapted to the case of $K > N_t$ in a same manner. The corresponding downlink spectral efficiency is the solution to

$$w\left(T_{\mathrm{fb}}, K\right) \triangleq \max_{T_{\mathrm{tr}}: T_{\mathrm{tr}} \leq T} \left(1 - \frac{T_{\mathrm{tr}}}{T}\right) \left(R_K^{\mathrm{ZF}} - \log\left(1 + \frac{N_t - 1}{T_{\mathrm{tr}}} + \Delta(T_{\mathrm{fb}})\right)\right) \tag{63}$$

where now R_K^{ZF} denotes the perfect CSIT rate with ZF beamforming and user selection [19], [20] and K users. This is computed via Monte Carlo simulation due to the lack of an analytical expression. Since the T_{fb} feedback symbols are now split between K users, we now have $\Delta(T_{\mathrm{fb}}) = \rho\left(1+\rho\right)^{-\frac{T_{\mathrm{fb}}}{K(N_t-1)}}$ for the case of error-free digital feedback.

In Fig. 9, the downlink sum spectral efficiency $w\left(T_{\mathrm{fb}}, K\right)$ is plotted versus T_{fb} for $K = 4, \ldots, 8$. The spectral efficiency is maximized by letting $K = 4, 5, 6, 7, 8$ users feedback for $T_{\mathrm{fb}} \leq 24$, $25 \leq T_{\mathrm{fb}} \leq 29$, $30 \leq T_{\mathrm{fb}} \leq 36$, $37 \leq T_{\mathrm{fb}} \leq 41$, $T_{\mathrm{fb}} \geq 42$, respectively. Thus, the sum spectral efficiency is maximized by having approximately $\frac{T_{\mathrm{fb}}}{6}$ users feedback; this is very consistent with the findings of [21]. If the number of users is fixed to $K = 4$ there is virtually no benefit in increasing T_{fb} beyond 35 or 40 because at that point the feedback channel is essentially perfect. However, a larger T_{fb} enables more users to feed back and yields a non-negligible gain in the achievable rate. For $T_{\mathrm{fb}} \leq 200$, it turns out that no more than 31 users are needed. In Fig. 10 the same plot is given for $K = 4, \ldots, 31$, for ideal digital and QAM feedback. As T_{fb} increases the marginal benefit of feedback (i.e., the slope) decreases, but adding users does provide a reasonable benefit even up to the 31-st user.

We can also consider the tradeoff between uplink and downlink rate as done before. Plotted in Fig. 11 are the uplink and downlink sum rates, using precisely the same parameters as Fig. 4 (i.e., $T_c = 1$ msec and $W_c = 200$ kHz). We now see a non-trivial tradeoff for downlink rates larger than 1750 kbps (as before, it does not make sense to choose a smaller downlink rate than this unless uplink data rate is much more strongly preferred than downlink data rate). If uplink and downlink data rates are equally weighted, the optimal operating point corresponds to (approximately) $R_{\mathrm{up}} = 828$ Kbps and $R_{\mathrm{down}} = 1966$ Kbps, which is achieved with $K = 11$ and $T_{\mathrm{fb}} = 63$ symbols. Note that the substantial benefit of allowing more users to feed back means that roughly 30 $\%$ of the uplink bandwidth is used for channel feedback.

REFERENCES

[1] G. Caire and S. Shamai, "On the achievable throughput of a multiantenna Gaussian broadcast channel," *IEEE Trans. on Inform. Theory*, vol. 49, no. 7, pp. 1691–1706, 2003.

[2] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge University Press, 2005.

[3] S. Sesia, M. Baker, and I. Toufik, *LTE, The UMTS Long Term Evolution: From Theory to Practice*. Wiley-Blackwell, 2009.

[4] T. Yoo and A. Goldsmith, "On the optimality of multiantenna broadcast scheduling using zero-forcing beamforming," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 3, pp. 528–541, 2006.

[5] B. Hassibi and B. Hochwald, "How much training is needed in multiple-antenna wireless links?" *IEEE Trans. on Inform. Theory*, vol. 49, no. 4, pp. 951–963, 2003.

[6] A. Dana, M. Sharif, and B. Hassibi, "On the capacity region of multi-antenna Gaussian broadcast channels with estimation error," in *IEEE Int. Symp. on Inform. Theory*, July 2006.

[7] D. Love, "Duplex distortion models for limited feedback MIMO communication," *IEEE Trans. on Sig. Proc.*, vol. 54, pp. 766–774, Feb 2006.

[8] W. Santipach and M. Honig, "Capacity of beamforming with limited training and feedback," in *IEEE Int. Symp. on Inform. theory*, Seattle, Washington, 2006.

[9] ——, "Optimization of training and feedback for beamforming over a MIMO channel," in *IEEE Wireless Comm. and Networking Conf.*, 2007.

[10] M. Kobayashi, G. Caire, and N. Jindal, "How much training and feedback are needed in MIMO broadcast channels?" in *IEEE International Symposium on Information Theory, 2008. ISIT 2008*, 2008, pp. 2663–2667.

[11] ——, "Optimized training and feedback for MIMO dwnlink channels," in *Proc. IEEE Information Theory Workshop, Greece*, 2008.

[12] G. Caire, N. Jindal, M. Kobayashi, and N. Ravindran, "Multiuser MIMO Downlink Made Practical : Achievable Rates with Simple Channel State Estimation and Feedback Schemes," *Arxiv preprint cs.IT/0710.2642*.

[13] J. Jose, A. Ashikhmin, P. Whiting, and S. Vishwanath, "Scheduling and pre-conditioning in multi-user MIMO TDD systems," *Arxiv preprint cs.IT/0709.4513*, 2007.

[14] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[15] N. Jindal, "MIMO broadcast channels with finite rate feedback," *IEEE Trans. on Inform. Theory*, vol. 52, no. 11, pp. 5045–5059, November 2006.

[16] A. Goldsmith, *Wireless Communications*. Cambridge University Press, 2005.

[17] E. Biglieri, J. Proakis, S. Shamai, and D. di Elettronica, "Fading channels: information-theoretic and communications aspects," *IEEE Trans. on Inform. Theory*, vol. 44, no. 6, pp. 2619–2692, 1998.

[18] A. Lapidoth, "On the asymptotic capacity of stationary Gaussian fading channels," *IEEE Trans. on Inform. Theory*, vol. 51, no. 2, p. 437, 2005.

[19] T. Yoo and A. Goldsmith, "On the optimality of multiantenna broadcast scheduling using zero-forcing beamforming," *IEEE J. Select. Areas Commun.*, vol. 24, no. 3, pp. 528–541, 2006.

[20] G. Dimic and N. Sidiropoulos, "On downlink beamforming with greedy user selection: performance analysis and simple new algorithm," *IEEE Trans. on Sig. Proc.*, vol. 53, no. 10, pp. 3857–3868, October 2005.

[21] N. Ravindran and N. Jindal, "Multi-user diversity vs. accurate channel feedback for MIMO broadcast channels," *Arxiv preprint cs.IT/0710.1336*, 2007.

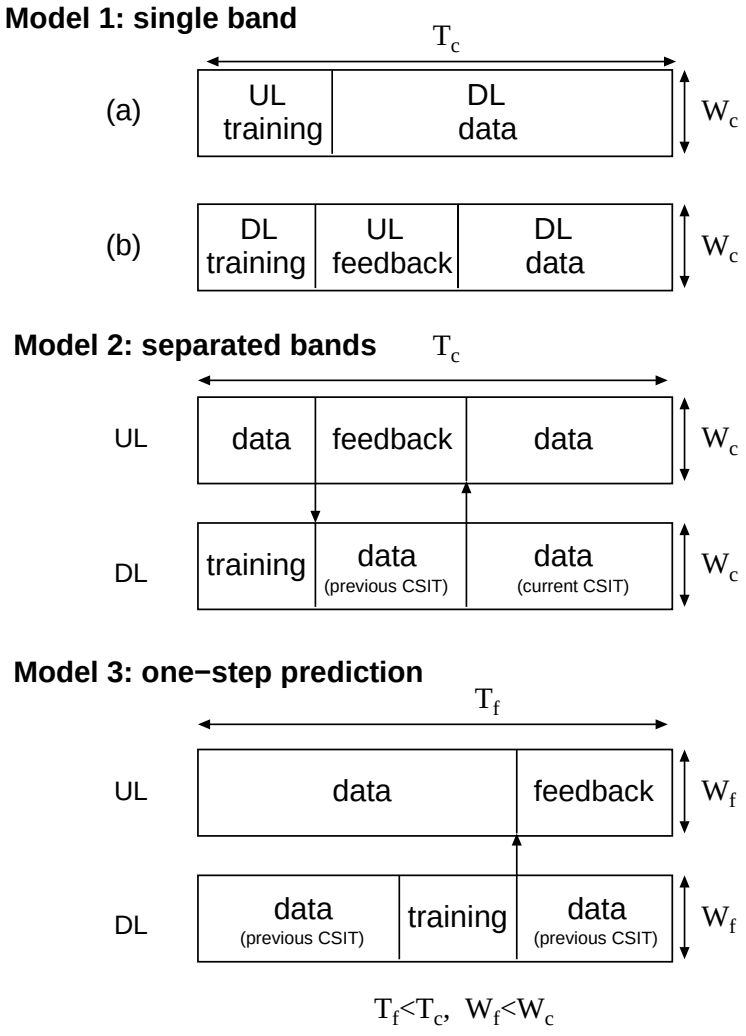

Fig. 1. Different time-frequency block models.

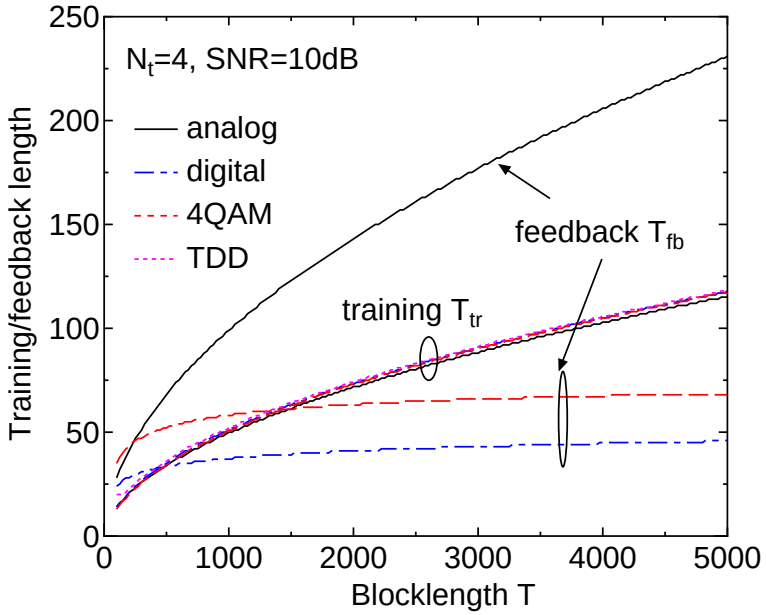

Fig. 2. Feedback/training length vs. block length for $N_t = 4, \rho = 10$ dB.

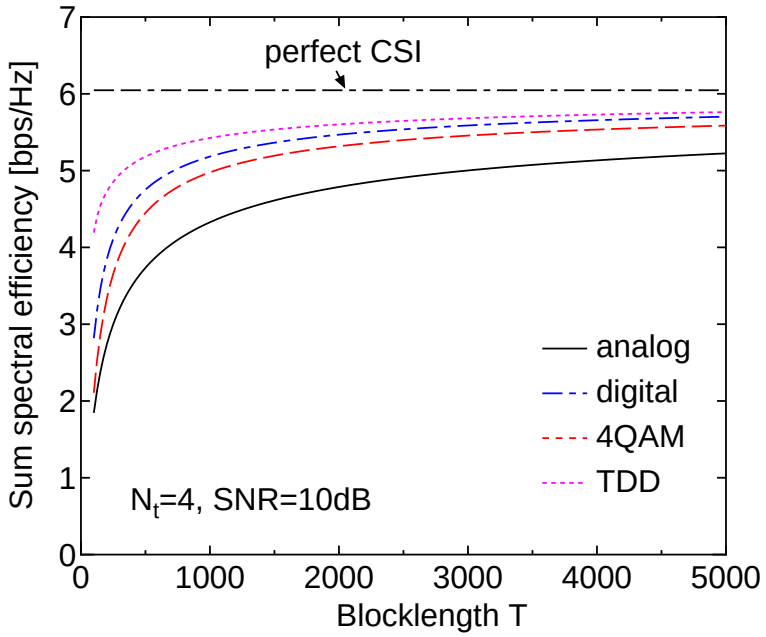

Fig. 3. Sum spectral efficiency vs. block length T.

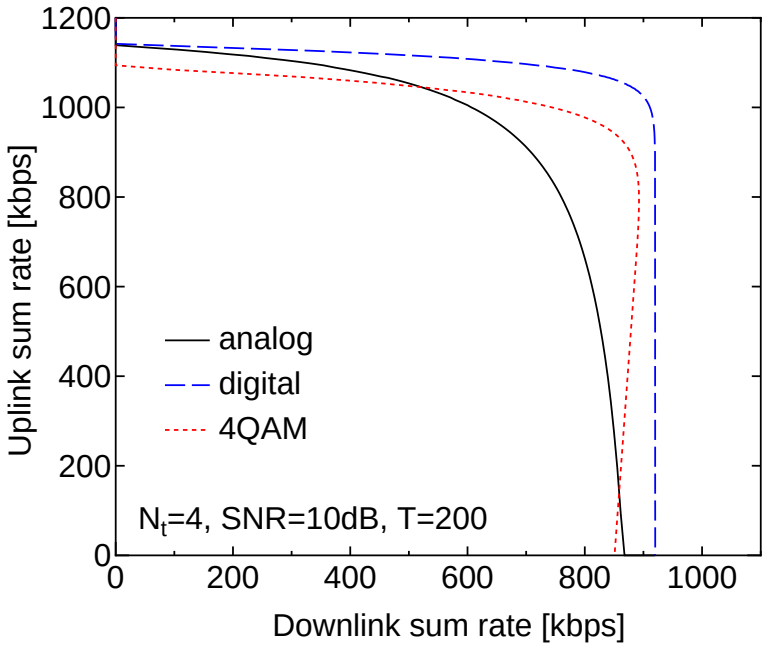

Fig. 4. Downlink vs. uplink tradeoff for $N_t = 4$, $\rho = 10$ dB, $T = 200$ symbols.

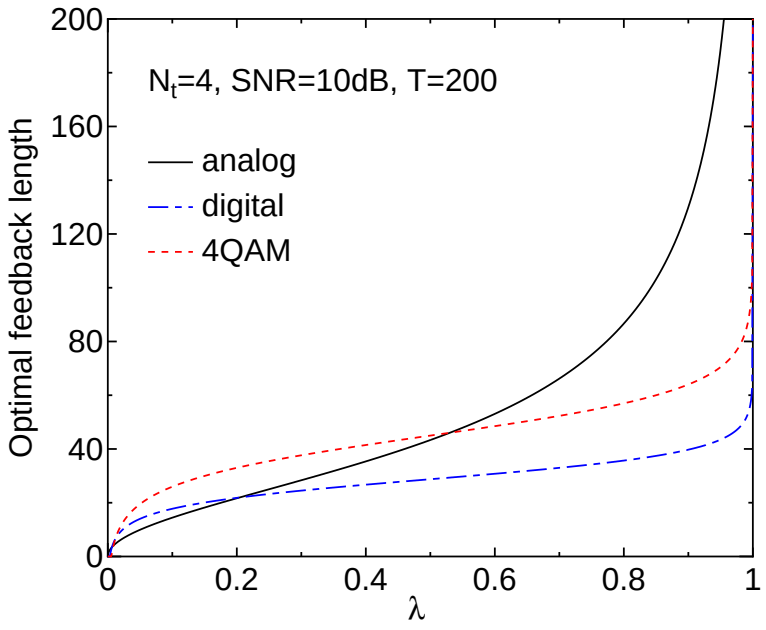

Fig. 5. Feedback length vs. λ for $N_t = 4$, $\rho = 10$ dB, $T = 200$ symbols.

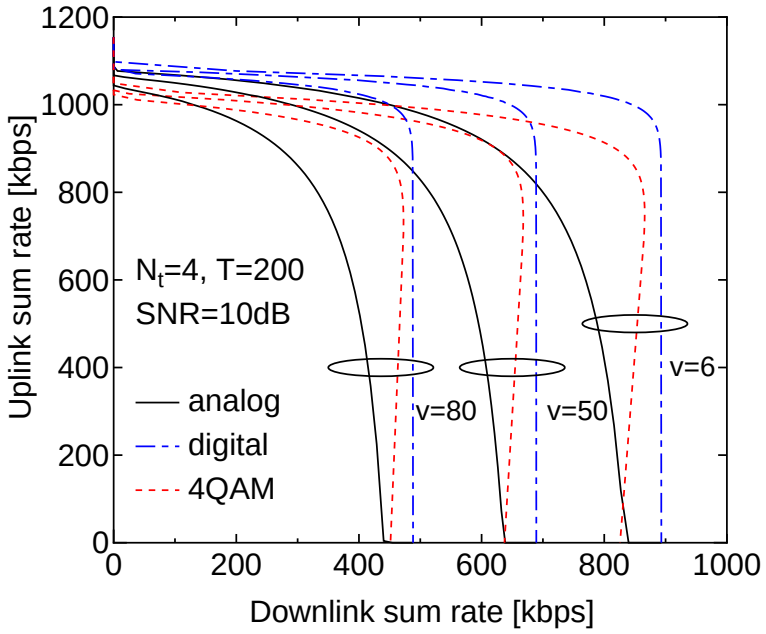

Fig. 6. Downlink vs. uplink tradeoff over the delayed feedback for $N_t = 4$, $\rho = 10$ dB, $T = 200$ symbols.

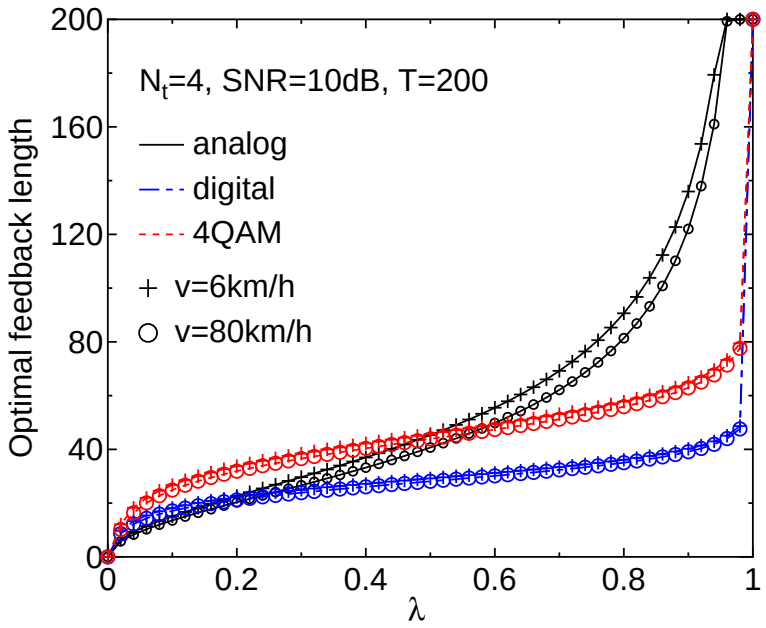

Fig. 7. Feedback length vs. λ for $N_t = 4$, $\rho = 10$ dB, $T = 200$ symbols.

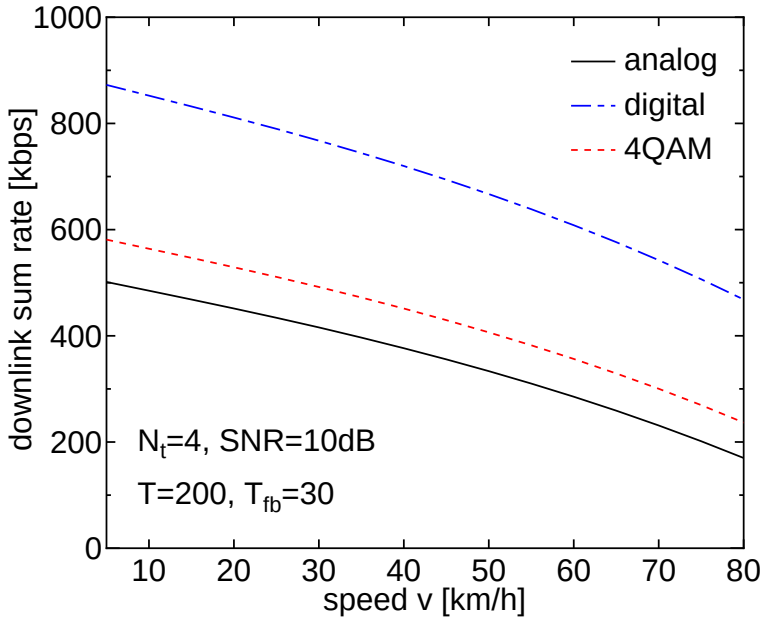

Fig. 8. Downlink rate vs. mobile speed for $T_{\rm fb} = 30$, $T = 200$.

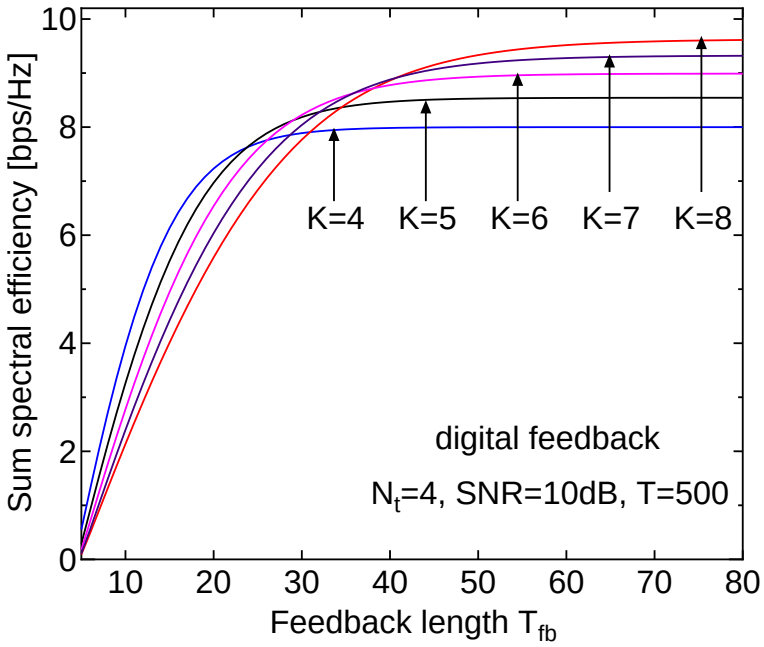

Fig. 9. Downlink sum spectral efficiency vs. feedback symbols ($T_{\rm fb}$) for $T = 500$, $N_t = 4$, $\rho = 10$ dB, for K from 4 to 8 users.

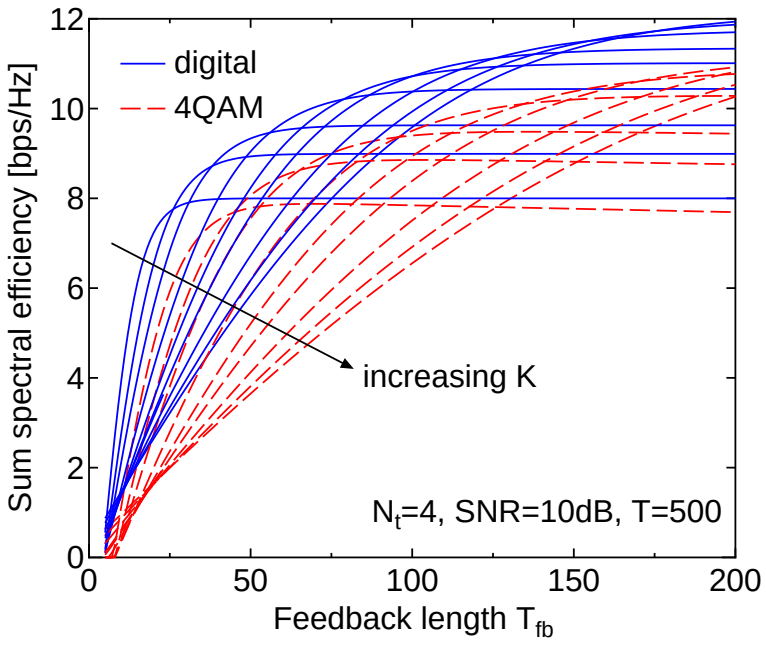

Fig. 10. Downlink sum spectral efficiency vs. feedback symbols (T_{fb}) for $T = 500$, $N_t = 4$, $\rho = 10$ dB, for K from 4 to 31 users.

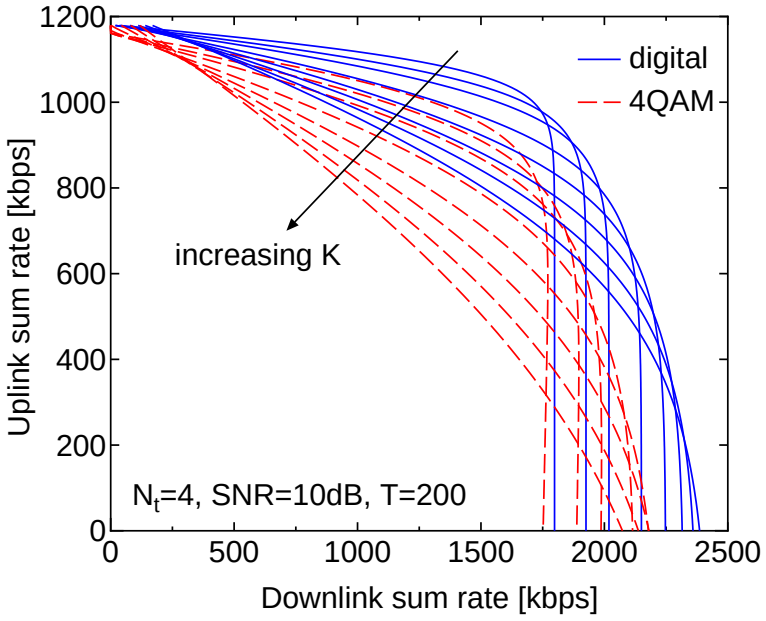

Fig. 11. Downlink vs. uplink tradeoff for $N_t = 4$, $\rho = 10$ dB, $T = 200$ symbols; allowing for up to 31 users to feed back.