Optimal Transmission of Short-Packet Communications in Multiple-Input Single-Output Systems

Chunhui Li, *Student Member, IEEE,* Nan Yang, *Member, IEEE,* and Shihao Yan, *Member, IEEE*

***Abstract*—We design the optimal transmission strategy which maximizes the average achievable data rate of the multiple-input single-output system which adopts short-packet communications. In this system, the N_A-antenna access point (AP) transmits to the single-antenna user with finite blocklength T after estimating the AP-user channel via downlink training and uplink feedback. For this system, we determine the optimal allocation of the finite resource (e.g., the total transmit power and a finite number of symbol periods) for downlink training, uplink feedback, and data transmission to maximize the average data rate. Specifically, we derive an approximate closed-form lower bound on the average data rate, an explicit result for the optimal number of symbol periods for downlink training, an easy-to-implement method to find the optimal number of symbol periods for uplink feedback, and a simple expression for the optimal power allocation between data transmission and downlink training. By using numerical results, we demonstrate the effectiveness of our analytical solutions and examine the impact of system parameters, e.g., N_A and T, on the optimal strategy.**

***Index Terms*—Channel training, feedback overhead, short-packet communications, power allocation.**

I. Introduction

Ultra-reliable and low-latency communication (URLLC) has been envisaged as the enabling paradigm to support the real-time communications with stringent requirements on latency and reliability. The realization of URLLC will bring transformational applications, e.g., smart manufacturing, autonomous networked vehicles, and remote surgery, to the human society. Notably, these applications typically require a target block error rate (BLER) less than 10^{-5} within a latency bound of 1 ms [1]. Such strictly low latency and low BLER impose an unprecedented restriction on the size of packets. Fortunately, short packets have been recognized as the typical forms of traffic generated in URLLC. For example, in industrial manufacturing and control systems, measurements and control commands are of small size (e.g., 10 to 20 bytes) [2]. Therefore, the theoretical investigation of the performance achieved by short-packet communications is of pivotal importance to realize URLLC in the near future.

There are unique challenges brought by short-packet communications. It has been pointed out that in the finite blocklength regime, the traditional performance metrics, e.g., Shannon capacity and outage capacity, provide *inaccurate* estimates on the maximum achievable rate [3], [4]. Indeed, the Shannon capacity is independent of BLERs while the outage capacity fails to capture the rate penalty caused by channel estimation overhead. By recognizing this, [3] investigated the maximum channel coding rate achievable at given finite blocklength and error probability. Specifically, the maximum achievable rate in the finite blocklength regime is tightly approximated as

$$R \approx C\left(\gamma\right) - \sqrt{V\left(\gamma\right)/T}Q^{-1}\left(\epsilon\right), \tag{1}$$

where T is the blocklength, ϵ is the BLER, γ is the signal-to-noise ratio (SNR), $C\left(\gamma\right) = \log_2\left(1+\gamma\right)$ is the Shannon capacity, $V\left(\gamma\right) = \left(\log_2 e\right)^2\left(1 - 1/(1+\gamma)^2\right)$ is the channel dispersion, and $Q^{-1}\left(\cdot\right)$ is the inverse Q-function.

Very recently, the benefits of short-packet communications have been examined for emerging wireless mechanisms, such as non-orthogonal multiple access [5], cooperative relaying [6], cooperative IoT networks [7], [8], wireless energy transfer [9], and radio resource management [10]. In practical communication systems, the wise resource allocation for channel estimation overhead plays a significant role in determining the transmission performance. Traditionally, the impact of channel estimation overhead has been studied in the asymptotic scenario with infinite blocklength (e.g., see [11], [12]). However, there have been only a few studies (e.g., [4]) that investigated the impact of channel estimation overhead for finite blocklength. While [4]–[6], [9]–[12] stand on their own merits, the design of short-packet communications with limited channel estimation overhead is still recognized as an open research issue.

In this paper, we design the optimal power and symbol period allocation to maximize the average data rate of the downlink in a multiple-input single-output (MISO) system which uses short-packet communications. In the system with finite blocklength T, the N_A-antenna access point (AP) estimates the downlink channel with the aid of downlink training and uplink feedback, and then performs data transmission. We derive an approximate closed-form expression for the lower bound on the average data rate taking into account T, based on which we determine the optimal symbol periods allocated to downlink training, uplink feedback, and data transmission, as well as the optimal power allocation between downlink training and data transmission.

II. System Model

We consider a MISO communication system where an N_A-antenna access point (AP) transmits small packets to a single-antenna user. We denote $\mathbf{h}_d$ as the $1 \times N_A$ channel vector from the AP to the user, the entries of which are subject to independent quasi-static Rayleigh fading. Therefore, the entries of $\mathbf{h}_d$ are independent and identically distributed (i.i.d.) circularly symmetric complex Gaussian random variables with zero mean and unit variance, i.e., $\mathbf{h}_d \sim \mathcal{CN}\left(0, \mathbf{I}_{N_A}\right)$. We assume that the entries of $\mathbf{h}_d$ remain constant during one

This work was supported by the Australian Research Council under Discovery Project Grant DP180104062.

C. Li and N. Yang are with the Research School of Electrical, Energy and Materials Engineering, The Australian National University, Canberra, ACT 2600, Australia (Emails: {chunhui.li, nan.yang}@anu.edu.au).

S. Yan is with the School of Engineering, Macquarie University, North Ryde, NSW 2109, Australia (Email: shihao.yan@mq.edu.au).

fading block. We also assume that the total duration of each fading block consists of T symbol periods (i.e., T channel uses), including T_t symbol periods used for downlink training, T_f symbol periods used for uplink feedback, and T_d symbol periods used for data transmission. Therefore, we have $T_t + T_f + T_d = T$. By considering the finite blocklength regime, we assume that T is relatively small such that the approximation in (1) is tight. According to [3] and [4], the approximation in (1) is tight even when T is as low as 100. We denote P_t and P_d as the transmit power per channel use at the AP for downlink training and data transmission, respectively. We further denote P as the average transmit power per channel use at the AP. Here, an average power constraint is considered over a fading block [12], i.e., $P_t T_t + P_d T_d \leq PT$. Additionally, we assume that the user and the AP have the knowledge about the statistical information of $\mathbf{h}_d$.

We assume that the channel estimation in the considered MISO system is performed as follows: First, the AP sends pilot sequences to the user for estimating $\mathbf{h}_d$, referred to as downlink training. Second, the user feeds back the estimate to the AP, referred to as uplink feedback. We next formulate downlink training and uplink feedback in the following.

1) Downlink Training: When the AP sends pilot sequences in T_t symbol periods, the received signal vector at the user is given by $\mathbf{y}_d = \sqrt{\Lambda}\mathbf{h}_d\mathbf{S}_d + \mathbf{n}_d$, where $\Lambda \triangleq P_t T_t / N_A$, $\mathbf{S}_d$ is the $N_A \times T_t$ pilot sequence matrix transmitted by the AP which satisfies $\mathbf{S}_d\mathbf{S}_d^\dagger = \mathbf{I}_{N_A}$, and $\mathbf{n}_d$ is the $1 \times T_t$ additive white Gaussian noise (AWGN) vector at the user with i.i.d entries, each of which follows the complex Gaussian distribution with zero mean and variance σ^2.

By adopting the MMSE estimator based on the known $\mathbf{S}_d$, the user obtains the estimate of $\mathbf{h}_d$ as $\hat{\mathbf{h}}_d = \frac{\sqrt{\Lambda}}{\Lambda+\sigma^2}\mathbf{y}_d\mathbf{S}_d^\dagger$. As per the property of MMSE, the channel estimation error, given by $\hat{\mathbf{e}}_d = \mathbf{h}_d - \hat{\mathbf{h}}_d$, is independent of the realizations of estimated channel [13]. We also note that $\hat{\mathbf{e}}_d$ and $\hat{\mathbf{h}}_d$ have i.i.d. entries. Specifically, each entry of $\hat{\mathbf{e}}_d$ follows the complex Gaussian distribution with zero mean and variance $\sigma^2_{\hat{e}_d}$ while each entry of $\hat{\mathbf{h}}_d$ follows the complex Gaussian distribution with zero mean and variance $\sigma^2_{\hat{h}_d}$, where $\sigma^2_{\hat{e}_d} = \sigma^2 / \left(\Lambda + \sigma^2\right)$ and $\sigma^2_{\hat{\mathrm{h}}_d} = \Lambda / \left(\Lambda + \sigma^2\right)$. We assume that $T_t \geq N_A$ is ensured in the MISO system to obtain a reliable estimate of $\mathbf{h}_d$.

2) Uplink Feedback: After downlink training, the user captures the channel direction information (CDI) given by $\tilde{\mathbf{h}} = \hat{\mathbf{h}}_d / \|\hat{\mathbf{h}}_d\|$. Then, the user quantizes the CDI by selecting the best quantization vector from the pre-shared codebook and conveys its index back to the AP over a feedback channel with zero propagation delay. Here, the propagation delay means the physical transmission duration from the AP to the user, which is formulated as the ratio between the transmission distance and the speed of light. Typically, the communication distance in URLLC is less than a few kilometers. Therefore, the propagation delay in the order of μs can be negligible [14]. Here, the codebook $\mathcal{C}$ is an $N_A \times 2^B$ matrix, i.e, $\mathcal{C} = \{\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_{2^B}\}$, where $\mathbf{w}_i$ refers to the $N_A \times 1$ channel vector and $i \in \left\{1, 2, \cdots, 2^B\right\}$. We clarify that the relationship between T_f and B is $B = T_f \log_2 M$, where M is the modulation order. The codebook is assumed to be designed offline and known to both the AP and the user. Given the codebook $\mathcal{C}$, the user chooses the quantization vector that maximizes the SNR as the best quantization vector, i.e., $\mathbf{w}_{\text{opt}} = \text{argmax}_{1 \leq i \leq 2^B} \left|\tilde{\mathbf{h}}\mathbf{w}_i\right|^2$. The user then feeds back the index of selected quantization vector to the AP. After obtaining the CDI, i.e., $\tilde{\mathbf{h}}$, through downlink training and uplink feedback, the AP sets the $N_A \times 1$ normalized beamforming vector as $\mathbf{w}_{\text{opt}}$ to transmit to the user. The transmitted signal $\mathbf{x}$ is written as $\mathbf{x} = \mathbf{w}_{\text{opt}} u$, where u is the information signal transmitted from the AP to the user. The received signal at the user in one symbol period is given by

$$y = \sqrt{P_d}\mathbf{h}_d\mathbf{x} + n = \sqrt{P_d}\hat{\mathbf{h}}_d\mathbf{w}_{\text{opt}} u + n_d. \tag{2}$$

We consider the worst-case scenario for the decoding process at the user where $n_d = \sqrt{P_d}\hat{\mathbf{e}}_d\mathbf{w}_{\text{opt}} u + n$ in (2) is approximated as a Gaussian random variable with the variance $\sigma^2_{n_d}$. Under this consideration, the SNR at the user is given by

$$\gamma = \left|\hat{\mathbf{h}}_d\mathbf{w}_{\text{opt}}\right|^2 P_d / \sigma^2_{n_d} = \rho_e \left\|\bar{\mathbf{h}}_d\right\|^2 \cos^2\left(\angle(\tilde{\mathbf{h}}, \mathbf{w}_{\text{opt}})\right), \tag{3}$$

where $\sigma^2_{n_d} = P_d\sigma^2_{\hat{e}_d} + \sigma^2$ with $\sigma^2_{\hat{e}_d} = \frac{N_A\sigma^2}{T_t P_t + N_A\sigma^2}$, $\rho_e = P_d\left(1 - \sigma^2_{\hat{e}_d}\right) / \sigma^2_{n_d}$, $\bar{\mathbf{h}}_d \triangleq \hat{\mathbf{h}}_d / \sigma_{\hat{h}_d}$ is the normalized channel vector with the standard deviation $\sigma_{\hat{h}_d}$, and $\cos^2\left(\angle(\tilde{\mathbf{h}}, \mathbf{w}_{\text{opt}})\right) = \left|\tilde{\mathbf{h}}\mathbf{w}_{\text{opt}}\right|^2$.

III. Performance Optimization

In this section, we perform the optimization of the symbol periods used for downlink training and uplink feedback, i.e., T_t and T_f, as well as the transmit power allocated to data transmission and channel training, aiming to maximize the average data rate under the average transmit power constraint. To this end, we denote η as the power allocation coefficient such that η and $1-\eta$ are the fraction of total transmit power allocated to data transmission and channel training, respectively. Thus, we have $\rho_d T_d = \eta\rho T$ and $\rho_t T_t = (1-\eta)\rho T$, where $\rho_d = P_d/\sigma^2$, $\rho_t = P_t/\sigma^2$, and $\rho = P/\sigma^2$.

A. Lower Bound on Average Data Rate

We first derive a lower bound on the average data rate in the context of short-packet communications. Considering n_d in (2) as a Gaussian random variable, a lower bound on the average data rate with limited channel estimation overhead for short-packet communications is

$$R = \tau\mathbb{E}\left[C(\gamma) - \sqrt{V(\gamma)/T}Q^{-1}(\epsilon)\right], \tag{4}$$

where $\tau = 1 - \left(T_t + T_f\right)/T$ and $\mathbb{E}\left[\cdot\right]$ denotes the expectation operation with respect to the channel gain. It is worth mentioning that (4) emphasizes the effects of the channel training and feedback overheads [15] for a given decoding error probability. We note that the average data rate is different from the average throughput $(1-\epsilon)R$ defined in [8], where the throughput is averaged over different decoding error probabilities. In this work, we set the decoding error probability as a constraint that can satisfy the reliability requirement. In addition, (4) only focuses on the rate which is used to transmit data. That is why it is named as the average data rate, but not the

$$\Phi\left(\tilde{\rho}_e, N_A\right) = \frac{e^{\frac{1}{\tilde{\rho}_e}}}{\ln 2\ \Gamma\left(N_A\right)\tilde{\rho}_e^{N_A}} \sum_{i=0}^{N_A-1} \binom{N_A-1}{i} (-1)^{N_A-1-i} \mathbf{G}_{2,3}^{3,0}\left(\begin{matrix} -i,-i \\ 0,-1-i,-1-i \end{matrix} \,\middle|\, \frac{1}{\tilde{\rho}_e} \right). \tag{6}$$

$$\Psi\left(\tilde{\rho}_e, N_A, T\right) = \sqrt{\frac{2\pi}{T}} \frac{Q^{-1}(\epsilon)}{\ln 2\ \Gamma\left(N_A\right)} e^{-(N_A-1)} (N_A-1)^{N_A-\frac{1}{2}} \sqrt{1-\left(1+\tilde{\rho}_e\left(N_A-1\right)\right)^{-2}}. \tag{7}$$

average throughput. In the following theorem, we derive an approximate closed-form expression for this lower bound.

Theorem 1: The approximate closed-form expression for the lower bound on the average data rate with limited channel estimation overhead is derived as

$$R \approx \tau\left[\Phi\left(\tilde{\rho}_e, N_A\right) - \Psi\left(\tilde{\rho}_e, N_A, T\right)\right], \tag{5}$$

where $\tilde{\rho}_e = \mu\rho_e$, $\mu = 1-(1-\frac{1}{N_A})2^{-\frac{B}{N_A-1}}$, and $\Phi\left(\tilde{\rho}_e, N_A\right)$ and $\Psi\left(\tilde{\rho}_e, N_A, T\right)$ are given by (6) and (7), respectively, on the top of the next page, with $\mathbf{G}_{2,3}^{3,0}\left(\cdot|\cdot\right)$ being the Meijer G-function [16, Eq. (9.301)]. It is worth mentioning that the results in [17] cannot be directly used in this work, since this work considers a different system model from [17]. Specifically, in this work the AP (i.e., the transmitter) obtains the channel state information (CSI) by performing downlink channel training and asking the user to feed back the index of the quantization vector in terms of the channel direction information. Differently, in [17] the transmitter obtains the CSI based on the channel reciprocity between uplink and downlink.

Proof: With the aid of the Jensen's inequality and the approximation of quantization errors given in [18], the average data rate in (4) can be approximated as

$$R \approx \tau\mathbb{E}_{\hbar}\left[C\left(\tilde{\rho}_e\hbar\right) - \sqrt{V\left(\tilde{\rho}_e\hbar\right)}Q^{-1}\left(\epsilon\right)/\sqrt{T}\right], \tag{8}$$

where, $\hbar = \left\|\bar{\mathbf{h}}_d\right\|^2$. Then we obtain (5) by following the procedure similar to [17, Appendix A]. ■

B. Formulating and Solving Optimization Problem

We now formulate and solve the optimization problem of our interest. First, we re-express the approximated expression for R given in (5) as $R\left(T_t, T_f, \eta\right)$, i.e., a function of T_t, T_f, and η. Then, we formulate the joint optimization of T_t, T_f, and η to maximize $R\left(T_t, T_f, \eta\right)$ under the average transmit power constraint as

$$\max_{T_t, T_f, \eta}\ R\left(T_t, T_f, \eta\right) \tag{9a}$$

$$\text{s.t. } \rho_t T_t + \rho_d T_d \leq \rho T. \tag{9b}$$

Considering the practical scenario and the accuracy of (1), we only focus on the case where $T_t \geq N_A$ and $T_d > N_A$. For given codebook, the beamforming vector given in (2) is optimal for the above optimization problem. Motivated by the results in [12], we derive the optimal value of T_t which maximizes $R\left(T_t, T_f, \eta\right)$ for given ρT, denoted by T_t^*, in the following theorem.

Theorem 2: The optimal T_t that maximizes $R\left(T_t, T_f, \eta\right)$ for given ρT in the case of $T_t \geq N_A$ and $T_d > N_A$ is derived as

$$T_t^* = N_A. \tag{10}$$

We note that the optimal T_t is the same as the number of transmit antennas N_A, which physically means that the channels associated with all transmit antennas can be estimated during the channel training phase.

Proof: The first derivative of $R\left(T_t, T_f, \eta\right)$, given in (8), with respect to T_d is derived as

$$\frac{\partial R\left(T_t, T_f, \eta\right)}{\partial T_d} = \frac{1}{T}\mathbb{E}_{\hbar}\left[\log_2\left(1+\omega\right) - \frac{\alpha_1\sqrt{\omega\left(2+\omega\right)}}{1+\omega}\right] + \frac{1}{T}\mathbb{E}_{\hbar}\left[\frac{\alpha_1\alpha_2\omega}{\left(1+\omega\right)^2\sqrt{\omega\left(2+\omega\right)}} - \frac{\alpha_2\omega}{\left(1+\omega\right)\ln 2}\right],$$

where $\hbar = \left\|\bar{\mathbf{h}}_d\right\|^2$, $\omega = \tilde{\rho}_e\hbar$, $\alpha_1 = Q^{-1}(\epsilon)/(\sqrt{T}\ln 2)$, and $\alpha_2 = \frac{T_d}{T_d-N_A}\left(1-\sqrt{\frac{N_A(N_A+\rho T)}{T_d(T_d+\rho T)}}\right)$. We find that $\alpha_1 < 1$ and $\alpha_2 < 1$ due to $T_d > N_A$. Then, we need to show that $\Omega\left(\omega\right) = \log_2\left(1+\omega\right) - \frac{\sqrt{\omega(2+\omega)}}{1+\omega} - \frac{\omega}{(1+\omega)\ln 2} \geq 0$ for all $\omega \geq \omega_0$, where ω_0 is the solution to $\Omega\left(\omega\right) = 0$. We find that $\Omega\left(\omega\right) = 0$ at $\omega = \omega_0$ and its first derivative $d\Omega\left(\omega\right)/d\omega = \left(\frac{\omega}{\ln 2} - \frac{1}{\sqrt{\omega(2+\omega)}}\right)\frac{1}{(1+\omega)^2} > 0$ for all $\omega \geq \omega_0$. We also find that $d\Omega\left(\omega\right)/d\omega$ is a monotonically increasing function of ω for all $\omega \geq \omega_0$, where the value of ω_0 is relatively small compared to the required value of SNR in URLLC scenarios (e.g., the SNR $>$ 10 dB [14], [19], [20]). Therefore, we conclude that the optimal value of T_t is the minimum value of T_t for given ρT in the case of $T_t \geq N_A$ and $T_d > N_A$. This is due to the fact that (4) is a monotonically increasing function of T_d and keeping the minimum value of T_t maximizes $R\left(T_t, T_f, \eta\right)$ for given T_f. ■

Based on Theorem 2, it is clear that T_t^* is independent of η and T_f. Thus, the objective function in (9) is rewritten as

$$\max_{T_f, \eta}\ R\left(T_t^*, T_f, \eta\right) \tag{11a}$$

$$\text{s.t. } \rho_t T_t^* + \rho_d T_d \leq \rho T. \tag{11b}$$

To solve (11), we first determine the optimal η which maximizes $R\left(T_t^*, T_f, \eta\right)$ for given T_f with $T_t^* = N_A$, denoted by η^*. Then we perform a one-dimensional search to find the optimal T_f based on the obtained η^* and T_t^*. We note that the equality in (11b) is always guaranteed due to the fact that a larger T_d always leads to a higher R. This also indicates that we only need to determine the optimal values of η and T_f to find the optimal value of T_d as $T_d^* = T - T_t^* - T_f^*$. We next present the details for solving (11) with $T_t^* = N_A$ using a two-step approach.

Step 1: Find the optimal η for given T_f.

We note that R in (4) is a monotonically increasing function of γ when R is positive. Also, the expectation in (4) is relevant to channel realizations but independent of $\tilde{\rho}_e$. That is, the

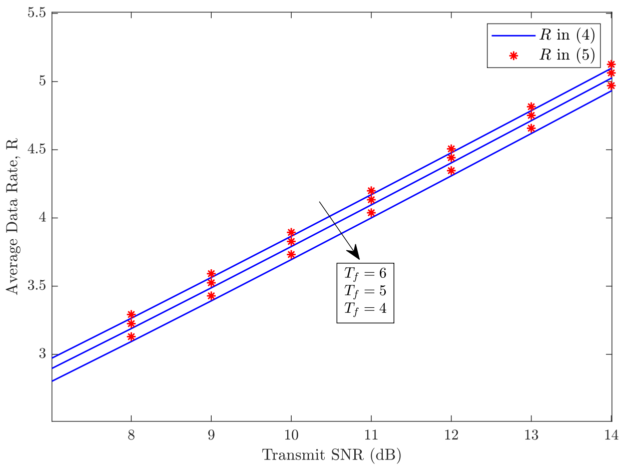

Fig. 1. The average data rate R versus transmit SNR for different feedback symbol period T_f with $T = 200$, $N_A = 4$ and $\epsilon = 10^{-6}$.

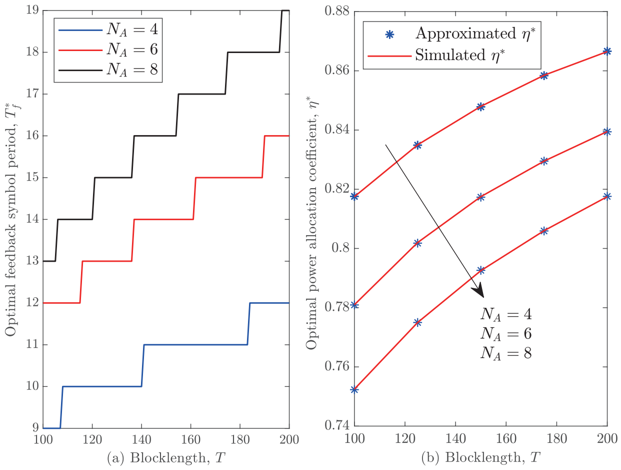

Fig. 2. The optimal feedback symbol period T_f^* and the optimal power allocation coefficient η^* versus T with $\rho = 10$ dB and $\epsilon = 10^{-9}$.

maximum average data rate is achieved by maximizing $\tilde{\rho}_e$ for given T_t and T_f. Based on (3), the effective SNR is given by

$$\tilde{\rho}_e = \frac{\mu P_d \left(1 - \sigma^2_{\hat{e}_d}\right)}{P_d \sigma^2_{\hat{e}_d} + \sigma^2} = \frac{\mu \rho T \eta (1-\eta)}{(T_d - N_A)(\nu - \eta)}, \tag{12}$$

where μ is defined below (5) and $\nu = \frac{\rho T + N_A}{\rho T (1 - N_A / T_d)}$.

By taking the second derivative of $\tilde{\rho}_e$ with respect to η, we find that $\frac{\partial^2 \tilde{\rho}_e}{\partial \eta^2} = \frac{\mu \rho T 2 \nu (\nu - 1)}{(T_d - N_A)(\eta - \nu)^3} < 0$, which confirms that $\tilde{\rho}_e$ is a concave function of η. As such, η^* can be found by numerically solving for $\partial \tilde{\rho}_e / \partial \eta = 0$, which gives

$$\eta^* = \nu - \sqrt{\nu^2 - \nu}. \tag{13}$$

Step 2: Find the optimal T_f.

We note that η^* given in (13) is a function of T and T_d (or equivalently, $T_t + T_f$), which is independent of the individual value of T_t or T_f. Since the optimal value of T_t is obtained, we can efficiently perform a one-dimensional numerical search to find the optimal T_f.

Overall, we first simplify the optimization problem by using $T_t^* = N_A$. Then, we maximize R over η for given T_f with T_t^*. After this, we find the optimal T_f, i.e., T_f^*, by using one-dimensional search. The complexity of our proposed method for solving (11) is low. Specifically, T_t^* can be obtained directly when N_A and T are determined. Based on this, for given T_f, we can obtain η^* according to (13). Finally, we perform a one-dimensional numerical search to find T_f^* within a finite range given by $T_f \in [T - N_A, T]$. Hence, when system parameters are determined, the optimization problem can be solved efficiently using our derived results. The effectiveness of our approach will be validated in Section IV.

IV. Numerical Results

Throughout this section, we consider the use of binary phase shift keying (BPSK) modulation for the feedback from the user to the AP such that $T_f = B$.

In Fig. 1, we demonstrate the accuracy of our derived closed-form expression for the lower bound on the data rate. The simulated and theoretical results are obtained from (4) and (5), respectively. The simulated points are averaged over 10,000 channel realizations, and the quantization codebook is generated based on the design criterion in [21]. In Fig. 1, we observe that the theoretical results precisely match the simulated ones during the whole SNR range, and the accuracy slightly increase when feedback symbol period increases. The observations imply that the quantization approximation has a almost negligible impact on the average data rate. Therefore, the closed-form expression derived in (5) serves as an accurate result for the average data rate with limited channel training and feedback under the consideration of the finite blocklength.

In Fig. 2(a), we plot the optimal symbol period for uplink feedback, T_f^*, versus T for different number of antennas at the AP, i.e., $N_A = 4$, 6, and 8. In this figure, we first observe that T_f^* increases as T increases. This observation is not surprising since more channel uses are allocated for downlink training and uplink feedback when T is larger. Also, we observe that T_f^* decreases as N_A decreases. This is due to the fact that decreasing N_A reduces the required channel uses for downlink training and uplink feedback. In Fig. 2(b), we plot the optimal power allocation coefficient, η^*, versus T for $N_A = 4$, 6, and 8. In this figure, we first confirm that the simulated curves exactly match the approximated values, demonstrating the correctness of our result derived in (13). We also observe that η^* increases as T increases. This is due to the fact that ρ_d remains stable and the ratio between T_d and T increases as T increases. Finally, we observe that η^* decreases as N_A increases. This is due to the fact that the number of channel uses for downlink training, T_t, increases with N_A, which reduces T_d.

In Fig. 3, we plot the transmit SNR for downlink training and data transmission versus T for different values of N_A, i.e., $N_A = 10$, 15, and 20. Here, we recall that $\rho_d = P_d/\sigma^2$ and $\rho_t = P_t/\sigma^2$. In this figure, we first observe that ρ_t gradually increases and ρ_d slightly decreases and tends to be constant as T increases. However, we confirm that $\rho_t T_t^*$ decreases and $\rho_d T_d$ increases as T increases. This implies that the transmit power allocated to downlink training decreases while the transmit power allocated to data transmission increases when T is larger. We also observe that ρ_t increases as N_A decreases. We further observe that ρ_d slightly decreases as N_A

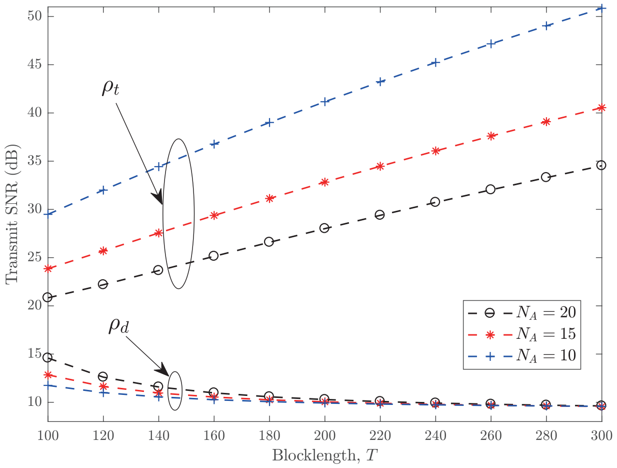

Fig. 3. The transmit SNR for downlink training and data transmission versus T for different N_A with $\rho = 10$ dB and $\epsilon = 10^{-9}$.

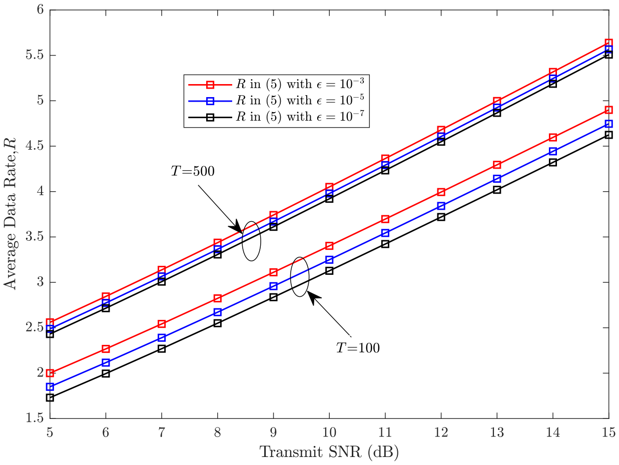

Fig. 4. The average data rate R versus transmit SNR for different values of ϵ and T with optimal η^* and $N_A = 4$.

decreases for small T but approaches almost the same value for large T. This observation can be explained by the fact that smaller N_A reduces the required T_t^* but leads to a negligible increase in T_d. This results in an increase in ρ_t and a minor reduction in ρ_d in order to guarantee $\rho_t T_t + \rho_d T_d = \rho T$.

Fig. 4 plots the lower bound on the data rate versus the transmit SNR for different values of ϵ and T. The curves are obtained from (5) with the optimal power allocation coefficient η^*. In this figure, we first observe that, for given T, the data rate decreases when the decoding error probability ϵ increases. It implies that the more strict requirement for reliability leads to a larger rate loss. Moreover, for the same ϵ, the data rate increases when the blocklength T increases as expected. We also find that the difference in data rates with different values of ϵ becomes negligible when T increases.

V. Conclusion

In this paper, we investigated the optimal resource allocation to maximize the average data rate in the MISO system which adopts short-packet communications. We proved that the optimal number of symbol periods allocated to downlink training is equal to the number of transmit antennas at the AP. We also derived the optimal power allocation between downlink training and data transmission at the AP in closed form. Our outcomes provide a guideline to assist the URLLC designers with the fundamental problem of transmit power and symbol period allocation to guarantee the advantage of short-packet communications in practice.

References

[1] 3GPP, “Study on scenarios and requirements for next generation access technologies,” 3rd Generation Partnership Project (3GPP), Technical Report (TR) 38.913, June 2017, Rel. 14.3.

[2] S. A. Ashraf, I. Aktas, E. Eriksson, K. W. Helmersson, and J. Ansari, “Ultra-reliable and low-latency communication for wireless factory automation: From LTE to 5G,” in *Proc. IEEE 21st Int. Conf. Emerging Technol. Factory Automation (ETFA)*, Berlin, Germany, Sept. 2016, pp. 1–8.

[3] Y. Polyanskiy, H. V. Poor, and S. Verdu, “Channel coding rate in the finite blocklength regime,” *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.

[4] G. Durisi, T. Koch, J. Östman, Y. Polyanskiy, and W. Yang, “Short-packet communications over multiple-antenna Rayleigh-fading channels,” *IEEE Trans. Commun.*, vol. 64, no. 2, pp. 618–629, Feb. 2016.

[5] X. Sun, S. Yan, N. Yang, Z. Ding, C. Shen, and Z. Zhong, “Short-packet downlink transmission with non-orthogonal multiple access,” *IEEE Trans. Wireless Commun.*, vol. 17, no. 7, pp. 4550–4564, July 2018.

[6] Y. Gu, H. Chen, Y. Li, L. Song, and B. Vucetic, “Short-packet two-way amplify-and-forward relaying,” *IEEE Signal Process. Lett.*, vol. 25, no. 2, pp. 263–267, Feb. 2018.

[7] L. Zhang, Y. Liang, and M. Xiao, “Spectrum sharing for Internet of Things: A survey,” *IEEE Wireless Commun. early access*, pp. 1–8, 2018.

[8] L. Zhang and Y. Liang, “Average throughput analysis and optimization in cooperative IoT networks with short packet communication,” *IEEE Trans. Veh. Technol.*, vol. 67, no. 12, pp. 11 549–11 562, Dec. 2018.

[9] O. L. A. López, E. M. G. Fernández, R. D. Souza, and H. Alves, “Ultra-reliable cooperative short-packet communications with wireless energy transfer,” *IEEE Sensors J.*, vol. 18, no. 5, pp. 2161–2177, Mar. 2018.

[10] C. She, C. Yang, and T. Q. S. Quek, “Radio resource management for ultra-reliable and low-latency communications,” *IEEE Commun. Mag.*, vol. 55, no. 6, pp. 72–78, June 2017.

[11] A. Lozano and N. Jindal, “Transmit diversity vs. spatial multiplexing in modern MIMO systems,” *IEEE Trans. Wireless Commun.*, vol. 9, no. 1, pp. 186–197, Jan. 2010.

[12] B. Hassibi and B. M. Hochwald, “How much training is needed in multiple-antenna wireless links?” *IEEE Trans. Inf. Theory*, vol. 49, no. 4, pp. 951–963, Apr. 2003.

[13] A. V. Oppenheim and G. C. Verghese, *Signals, systems and inference*. Upper Saddle River, NJ: Prentice Hallc, 2015.

[14] C. She, C. Yang, and T. Q. S. Quek, “Joint uplink and downlink resource configuration for ultra-reliable and low-latency communications,” *IEEE Trans. Commun.*, vol. 66, no. 5, pp. 2266–2280, May 2018.

[15] M. Kobayashi, N. Jindal, and G. Caire, “Training and feedback optimization for multiuser MIMO downlink,” *IEEE Trans. Commun.*, vol. 59, no. 8, pp. 2228–2240, Aug. 2011.

[16] I. S. Gradshteyn and I. M. Ryzhik, *Table of integrals, series, and products*, 7th ed. San Diego, CA: Academic, 2007.

[17] C. Li, S. Yan, and N. Yang, “On channel reciprocity to activate uplink channel training for downlink wireless transmission in Tactile Internet applications,” in *Proc. IEEE Int. Conf. Commun. (ICC) Workshop*, Kansas City, MO, May 2018, pp. 1–6.

[18] T. Yoo, N. Jindal, and A. Goldsmith, “Multi-antenna downlink channels with limited feedback and user selection,” *IEEE J. Sel. Areas Commun.*, vol. 25, no. 7, pp. 1478–1491, Sept. 2007.

[19] J. G. S. Schiessl and H. Al-Zubaidy, “Delay analysis for wireless fading channels with finite blocklength channel coding,” in *Proc. ACM MSWiM*, Cancun, Mexico, Nov. 2015, pp. 13–22.

[20] C. She, C. Yang, and T. Q. S. Quek, “Cross-layer optimization for ultra-reliable and low-latency radio access networks,” *IEEE Trans. Wireless Commun.*, vol. 17, no. 1, pp. 127–141, Jan. 2018.

[21] D. J. Love, R. W. Heath, and T. Strohmer, “Grassmannian beamforming for multiple-input multiple-output wireless systems,” *IEEE Trans. Inf. Theory*, vol. 49, no. 10, pp. 2735–2747, Oct. 2003.