A Low-Latency Precoding Strategy for In-Band Full-Duplex MIMO Relay Systems

Jacqueline R. Malayter, *Student Member, IEEE* and David J. Love, *Fellow, IEEE*

***Abstract*—Ultra-reliable low-latency communication (URLLC) and machine-to-machine (M2M) relaying are increasingly important. In-band full-duplex (IBFD) communication is desirable for low-latency communication because it can theoretically double the capacity of half-duplex (HD) communication, but is limited by self-interference (SI) in which the IBFD transmitter injects interference into its received signal due to large transmit and receive power differences. There are many methods to mitigate SI in IBFD, but we focus on multiple-input multiple-output (MIMO) precoding to spatially null SI, while also being careful to limit communication latency. Unlike works aiming to compute the highest achievable rate in IBFD relays with an essentially fixed precoder, we propose a strategy allowing the precoder to change on a per channel use basis to avoid rate bottlenecks at each hop. Our strategy generalizes to multiple relays. We frame finding the sequence of precoders to relay packets through N IBFD MIMO relays with the lowest communication latency as a shortest path problem. Specifically, we design a quantized covariance precoder codebook at each transmitter based on limiting the maximal SI power each precoder produces. Then, an iterative algorithm is used to optimize the selection of precoders over channel uses to relay packets in the fewest channel uses possible.**

***Index Terms*—MIMO, full-duplex, URLLC, relaying, short packets, wireless backhaul**

I. Introduction

As the vision of a smart, autonomous, and connected society is realized, communication systems must evolve to meet the new and distinct challenges that follow. On top of enhancing data rates, there are two additional features that must be incorporated into wireless networks: ultra-reliable low-latency communication (URLLC) and specialized short-packet communication for machine-to-machine (M2M) communication. URLLC is a type of communication with strict reliability and latency requirements subject to minimum throughput [2]. Meanwhile, M2M communication encompasses communication designed for large numbers of devices in dense areas. M2M communication is unique in that devices may use short packets, and M2M communication will likely require high reliability and low-latency [2], [3].

URLLC and M2M communication have value in a variety of applications. M2M communication will be pervasive in many scenarios, including in long-term low-power environmental sensing, in numerous applications in smart cities, and in partially or fully-autonomous factories, among other applications [4]. Similarly, URLLC will be used in critical disaster communication, drone and autonomous vehicle communication, and reliable cloud connectivity [2], [5]. It is clear M2M and URLLC are not generally exclusive.

Rural wireless communication and agricultural communication will especially benefit from both. Relay networks will extend and improve wireless broadband to isolated areas, as relay networks can provide wireless backhaul in areas where building cell towers is not economically feasible. Wireless backhaul will rely specifically on *low-latency* techniques to minimize delay introduced by processing at relay nodes, which is important for remote jobs, tele-education, telehealth, and more applications. Improved rural broadband, beyond improving quality of life for people in rural areas, opens opportunities for *smart agriculture*, where M2M and URLLC will be critical for data-driven agriculture, including in-field sensor networks and autonomous farm equipment [6].

Of great importance in analysis of URLLC/M2M systems is the idea that *short packets* will become an increasingly prevalent data type. For example, the main data traffic of sensors will be short packets [2]. Likewise, short packets may facilitate a reduction in the latency of relay networks. This is because, in decode-and-forward (DF) relaying, each relay node must aggregate an entire packet before decoding it and re-transmitting. If short packets are used, the time to accumulate an entire packet is reduced [7]. You may wonder how theoretical analyses of URLLC/M2M communication are possible given Shannon's main results on channel capacity explicity rely on sufficiently long blocklength [8]. More recent works by Polyanskiy *et. al* have extended channel coding results to the *finite-blocklength* regime, illustrating the trade-offs between blocklength, achievable rate, and error rate [9]. Following the seminal developments of Polyanskiy *et. al* [9], the finite blocklength capacity has been further explored for scalar fading channels, MIMO fading channels, MISO channels [3], [10]–[13].

An area of study in low-latency communication is finding achievable rates for the in-band full-duplex (IBFD) relay. The interest in IBFD for low-latency communication is because it could, in theory, double the spectral efficiency of a system by doing away with half-duplexing (HD), where uplink and downlink communication is performed in orthogonal time or frequency slots. Instead, receiving and transmitting signals occurs simultaneously in the same frequency band at a transceiver. The fundamental challenge with IBFD is that the self-interference (SI) from the transmit signal at a transceiver greatly overwhelms the receive signal due to a variety of reasons, including imperfect interference channel estimation

This work is in part supported by the National Science Foundation under the NSF Graduate Research Fellowship Program DGE1842166 and NSF grants CCF1816013 and EEC1941529 and by the Office of Naval Research under ONR Grant N00014-21-1-2472. This work has appeared in the 2022 IEEE Wireless Communications and Networking Conference (WCNC) [1].

J. R. Malayter and D. J. Love are with the Elmore Family School of Electrical and Computer Engineering at Purdue University.

and inadequate ADC dynamic range, among other hurdles [14]–[16]. There has been renewed interest in IBFD, however, because short-range systems are becoming more prevalent and it is important for future wireless backhaul. These communication architectures make SI suppression more manageable because the difference in power of the transmit and receive signal is less severe [14].

There are many works characterizing the performance of IBFD MIMO relays. In the work by Day *et al.* [15], the achievable end-to-end rate of an IBFD DF two-hop MIMO relay with limited dynamic range and channel estimation error over Rayleigh-fading channels is explored. To find the bound on achievable rate, the authors propose an algorithm for a non-convex optimization problem where the minimum mutual information (MI) of the two-hops is maximized. Although this work gives interesting guidelines on what conditions are ideal for HD versus IBFD, these results cannot be guaranteed to be globally maximal and assume a limited, stationary set of precoders is used throughout all channel uses. In Korpi *et. al* [17], the authors depart from using strictly IBFD *or* HD to find achievable rates, leading to the idea of *hybrid IBFD/HD* MIMO relaying. The authors found that some HD transmission is necessary to allow the first hop to transmit alone to avoid creating a rate bottleneck. Their results show that using a flexible hybrid scheme, that is, optimizing over HD versus IBFD time blocks to maximize rate, performed better than only using the IBFD scheme, such as in Day *et. al* [15]. Though Korpi *et. al* [17] takes a step towards creating a dynamic precoding IBFD relay architecture and improves upon the rates achieved in Day *et. al* [15], it is still unclear if this is the *lowest-latency* technique for relaying packets across a two-hop relay. Further, the precoder codebook presented in Korpi *et. al* [17] is restrictive– that is, the precoder at the relay transmitter is either off or the IBFD end-to-end rate maximizing precoder. It is unclear if the overall end-to-end delay could further benefit from more flexibility in precoder selection.

Rate analysis for IBFD relays with short-packets is a sparser research area. In Gu *et. al* [18], the authors introduce short packet expressions in HD versus IBFD analysis, although for a single antenna two-hop relay. For the single antenna case, it was found that IBFD was preferable in terms of overall delay for systems with lower transmit power, a less strict block error rate constraint, and, as expected, stronger SI suppression. In Hu *et. al* [19], the authors examine the performance of a single antenna IBFD amplify-and-forward (AF) and DF relay setup with reliability constraints. They found that shorter packets are more sensitive to SI and more resources should be spent on cancelling SI. As expected, lower SI values resulted in higher reliability relaying. We seek to extend the work on characterizing the minimum end-to-end delay for IBFD short-packet MIMO relay systems with reliability constraints.

In this paper, a framework is developed to minimize the overall delay in the transmission of a block of ℓ packets of size b bits across N fixed IBFD DF relays. By fixed, we assume that there is no choice in which relays to use, such as in the case of wireless backhaul networks or if the relays are chosen after some routing algorithm, for example. Each relay transmitter is equipped with a power-constrained precoder that can be used to optimize the achievable transmission rate or, inversely, spatially suppress SI at the relay receiver. We extend the idea of the hybrid precoding scheme in Korpi *et. al* [17] by allowing the precoders to change across channel uses, but we allow a more flexible selection of precoders. We develop a dynamic programming algorithm that determines the optimal set of precoders to apply to each relay in terms of reducing overall delay in the transmission. Upper and lower bounds on the overall transmission delay are derived, and the performance of our algorithm is compared with these bounds. Further, we substitute *finite block length* capacity expressions into our simulations and observe the overall delay for various block lengths and block error rate constraints.

Notation – In this work, $(\cdot)^T$, $(\cdot)^*$, and $\text{Tr}(\cdot)$ denotes the matrix transpose, conjugate transpose, and matrix trace respectively. $\lfloor \cdot \rfloor$ denotes the floor operator. $\mathcal{A}$ denotes a set, $\mathbf{A}$ denotes a matrix, $\mathbf{a}$ denotes a vector, $\mathbf{0}$ denotes the all zeros vector, and $\mathbf{a}[n]$ denotes the element in the nth index of $\mathbf{a}$. $\preceq$ and $\succeq$ denote the element-wise less than and greater than comparison of two vectors, respectively. $\text{card}(\cdot)$ denotes the cardinality of a set. $\mathcal{C} = \mathcal{A} \times \ldots \times \mathcal{B}$ means that $\mathcal{C} := \{(a_i, \ldots, b_i) : a_i \in \mathcal{A}, \ldots, b_i \in \mathcal{B}\}$.

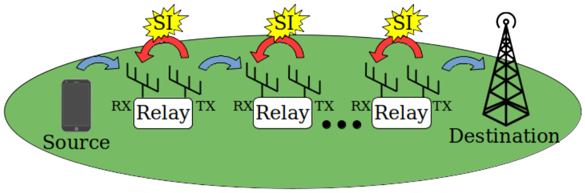

Fig. 1. High level diagram of IBFD relay network

II. Preliminaries

Consider a DF MIMO relay network with a source node, N IBFD relay nodes that produce self-interference (SI) from the relay transmitter to the receiver, and a destination node, where each node has multiple antennas with no direct link between non-adjacent nodes, such as the network shown in Fig. 1. The relay nodes of such a network are numbered consecutively from 1 to N, where relay 1 is adjacent to the source and relay N is adjacent to the destination. At the ith relay node, there are M^i_{tx} transmit antennas and M^i_{rx} receive antennas. The source has M^s_{tx} transmit antennas, and the destination has M^d_{rx} receive antennas. Each transmitter in the system is equipped with a $M^i_{tx} \times M^i_{tx}$ power-constrained precoder $\mathbf{F}^i_k$. System model details are discussed in detail in Section III-A.

A. Prior Work and Motivation

We begin by examining the current literature and performance of IBFD low-latency relaying and look to it for a motivating example. Results of interest to us are the end-to-end asymptotic rate-maximizing precoder result, such as is investigated in Day *et. al* [15]. In Day *et. al* [15], the authors model a two-hop MIMO IBFD relay system where the relay receiver experiences SI from its own transmitter. In essence, the authors proposed a method for numerically solving the non-convex optimization problem of finding the maximum end-to-end mutual information (which is the end-to-end capacity if channel estimation is assumed perfect) of the two-hop IBFD MIMO relay system. The optimization is performed with respect to the precoder covariance $\bar{\mathbf{Q}}[n] = \left(\mathbf{Q}_s[n], \mathbf{Q}_r[n]\right)$, which is a vector of two power constrained covariances matrices, the covariance of the source and the relay precoder, respectively, optimized over two communication epochs (that is, $n \in \{1, 2\}$). The two epoch scheme allows for a HD solution.

$$\mathcal{I}\left(\bar{\mathbf{Q}}\right) = \max_{\bar{\mathbf{Q}} \in Q_\tau} \min\left(\mathcal{I}_{sr,\tau}\left(\bar{\mathbf{Q}}\right), \mathcal{I}_{rd,\tau}\left(\bar{\mathbf{Q}}\right)\right) \tag{1}$$

where

$$\mathcal{I}_{sr,\tau} = \sum_{n=1}^{2} \tau[n] \mathcal{I}_{sr}\left(\bar{\mathbf{Q}}[n]\right) \tag{2}$$

and

$$\mathcal{I}_{rd,\tau} = \sum_{n=1}^{2} \tau[n] \mathcal{I}_{rd}\left(\bar{\mathbf{Q}}[n]\right) \tag{3}$$

are the mutual information of the source to relay and relay to destination link, respectively. The authors approach solving (1) by converting it into a weighted sum-rate optimization problem which in the end only guarantees a *local maximum* in a generally non-convex problem. The reader is directed to Day *et. al* [15] for further details on the proposed algorithm.

1) Implications on Latency with Practical Systems: By design, the above algorithm aims to find the *link-equalizing* precoder design (unless the source-to-relay link is much, much better than the relay-to-destination link). But what if we were transmitting ℓ packets across a DF single relay system? Consider why the solution to (1) might not be the optimal selection of precoders to deliver the ℓ packets to the destination. We define the number of channel uses required to get the ℓ packets to the relay as our metric of latency. That is, if transmitting over some fixed bandwidth $\mathcal{W}$ from a source to destination node, then the *lowest-latency* transmission method is the transmission method that delivers the ℓ packets in the shortest amount of time, or, alternatively, the *smallest number of channel uses*. Now, assume there is no direct link between the source and the destination node, that is, all packets must flow through the relay.

Clearly, to maintain a causal system, any packet arriving at the destination must have already been received and decoded in full at the relay first. From this perspective, for a DF relaying regime, the transmit precoder at the relay should suppress SI completely for the first $b/C^{s,r}$ channel uses, where we remind the reader that b is the number of bits per packet and $C^{s,r}$ is the capacity of the source to relay link. This is due to the causality of the system, as before the first packet has completely arrived at the relay, there is no information to be sent to the destination yet. This insight is somewhat incompatible with the algorithm in Day *et. al* [15], because for the first $b/C^{s,r}$ channel uses, it does not make much sense to design the source and relay precoders to be link-equalizing if it is not possible to send information over the relay-to-destination link. Instead, we should design the source and relay precoder to ensure the best source-to-relay rate to get the first packet to the relay as quickly as possible. Otherwise, the relay system would encounter a *bottleneck* where the relay must wait longer than necessary to transmit the first packet to the destination.

It is pointed out in Korpi *et. al* [17] the potential bottlenecking issue and propose using HD for some time before switching to FD for the rest of the transmission. But in the case for both works, there is no so-called *causality constraint* explicitly included in either algorithm. That is, throughout any point in the transmission of ℓ packets across a real separated relay system, it must hold that the relay has accumulated the same or more packets than the destination. It is unclear how this practical constraint affects how the source and relay precoders should be designed for a practical, low-latency relaying scheme. In this manuscript, we are tasked with building a low-latency relaying algorithm that explicitly considers such a constraint.

2) Complexity for N > 1 Relays: The algorithm presented in Day *et. al* [15] is explicitly designed for an $N = 1$ relay system. However, for example, in a rural setting or extremely dense urban setting where relay networks could provide wireless backhaul, using only one relay may not be a realistic assumption. The algorithm in Day *et. al* [15] is already somewhat computationally complex and is approximated for ease of implementation in subsequent sections of the manuscript, but it is not clear how this could generalize to $N > 1$ relays. Therefore, in addition to considering a causality constraint discussed in the previous subsection, we are motivated to develop a low-latency relaying framework that generalizes more easily to an arbitrary number of relays. In Section III-B, we discuss how we can formulate the low-latency relaying problem as a shortest path problem so that we may include a causality constraint and generalize to $N > 1$ relays. In our conference paper Malayter *et. al* [1], we demonstrated this result for $N = 1$ relays. In this work, we extend our findings to $N > 1$ relays, including methods to reduce computational complexity.

III. Channel Model and Problem Formulation

A. Propagation Channels

In general, we describe the channel from relay i to relay i' as the $M_{rx}^{i'} \times M_{tx}^{i}$ channel matrix $\mathbf{H}^{i,i'}$ and its reverse (not necessarily reciprocal) channel as $\mathbf{H}^{i',i}$, where $i < i'$ (where we adopt the convention that the channel from the source to relay i is $\mathbf{H}^{s,i}$ and the channel from relay i to the destination is $\mathbf{H}^{i,d}$). The reverse channels are only present between relays since these are the only nodes operating in IBFD and thus causes *inter-relay interference* (IRI). It is assumed the channel matrices of non-adjacent nodes are zero, that is, $\mathbf{H}^{i,i'} = \mathbf{H}^{i',i} = \mathbf{0}$ when $i' \neq i + 1$. Additionally, we model the SI as a *loopback interference* channel that describes the electromagnetic leakage from the transmitter to the receiver at a given relay. The SI channel at each of the relay nodes is described by the $M_{rx}^{i} \times M_{tx}^{i}$ matrix $\mathbf{G}^{i}$, where it is assumed that the entries of $\mathbf{G}^{i}$ are circularly symmetric complex Gaussian with zero mean and unit variance. Therefore, the system is an $N+1$ hop separated relay network, and each packet of information originating from the source must travel through the N relays consecutively to reach the destination.

As the primary objective of this manuscript is to optimize the precoders at each node's transmitter in terms of overall latency in the transmission of ℓ packets, we equip the transmitter at the source and each relay with an $M^i_{tx} \times M^i_{tx}$ precoder $\mathbf{F}^i_k$, where i denotes the node and k denotes the time index. The precoder power constraint is normalized such that $\mathrm{Tr}\left(\mathbf{F}^i_k \mathbf{F}^{i}_k{}^*\right) \leq 1$, where the power scaling of the $(i-1)$th precoder is lumped into the signal-to-noise ratio (SNR) term at the ith node, ρ_i. At the receiver of node i, the incoming signal is impaired by independent complex Gaussian noise $\mathbf{n}^i_k$ distributed $\mathcal{CN}(0, \mathbf{I})$.

Further, we define the signal vector transmitted at time k from relay i as the $M^i_{tx} \times 1$ vector $\mathbf{x}^i_k$, where the $\mathbf{x}^i_k$'s are drawn from independent complex Gaussian codebooks. Similarly, the signal vector transmitted at time k from the source is the $M^s_{tx} \times 1$ vector $\mathbf{x}^r_k$. It then follows that the received signal at relay i at time index k can be written as

$$\begin{aligned}\mathbf{y}^i_k =& \sqrt{\rho_i}\mathbf{H}^{i-1,i}\mathbf{F}^{i-1}_k\mathbf{x}^{i-1}_k + \sqrt{\eta_i}\mathbf{G}^i\mathbf{F}^i_k\mathbf{x}^i_k \\ &+ \sqrt{\alpha_i}\mathbf{H}^{i+1,i}\mathbf{F}^{i+1}_k\mathbf{x}^{i+1}_k + \mathbf{n}^i_k,\end{aligned} \tag{4}$$

where we slightly modify the notation for the received signal at relay 1 and the destination as

$$\mathbf{y}^1_k = \sqrt{\rho_1}\mathbf{H}^{s,1}\mathbf{F}^s_k\mathbf{x}^s_k + \sqrt{\eta_1}\mathbf{G}^1\mathbf{F}^1_k\mathbf{x}^1_k + \sqrt{\alpha_1}\mathbf{H}^{2,1}\mathbf{F}^2_k\mathbf{x}^2_k + \mathbf{n}^1_k$$

and

$$\mathbf{y}^d_k = \sqrt{\rho_d}\mathbf{H}^{N,d}\mathbf{F}^N_k\mathbf{x}^N_k + \mathbf{n}^d_k,$$

respectively. In (4), ρ_i and ρ_d are the SNR at relay i's receiver and the destination receiver, respectively. We define η_i is the interference-to-noise ratio (INR) of the loopback interference channel at the ith relay receiver and the term α_i is the INR of the IRI caused by the transmitter at the next consecutive relay node. It can be seen from (4) that the first term $\sqrt{\rho_i}\mathbf{H}^{i-1,i}\mathbf{F}^{i-1}_k\mathbf{x}^{i-1}_k$ is the signal of interest at the receiver, whereas the last three terms $\sqrt{\eta_i}\mathbf{G}^i\mathbf{F}^i_k\mathbf{x}^i_k$, $\sqrt{\alpha_1}\mathbf{H}^{i+1,i}\mathbf{F}^{i+1}_k\mathbf{x}^{i+1}_k$, and $\mathbf{n}^i_k$ are undesirable SI, IRI, and noise terms, respectively. Therefore, the SI and IRI terms are treated as additional noise. The second to last term in (4) is non-zero only if there are 2 or more relays, in which case the interference affects relays 1 through $N-1$. Thus in the separated relay case, the second to last term may be omitted. The covariance of the noise and SI terms is given by

$$\mathbf{\Sigma}^i_k = \eta_i\mathbf{G}^i\mathbf{Q}^i_k\mathbf{G}^{i*} + \alpha_i\mathbf{H}^{i+1,i}\mathbf{Q}^{i+1}_k\mathbf{H}^{i+1,i*} + \mathbf{I}, \tag{5}$$

where $\mathbf{Q}^i_k = \mathbf{F}^i_k\mathbf{F}^{*i}_k$ is the precoder covariance. Equation (4) gives intuition about the design problem for the precoders in the relay network. For example, it can be seen at node i, $\mathbf{F}^i_k$ can be designed to minimize or eliminate the SI term $\sqrt{\eta_i}\mathbf{G}^i\mathbf{F}^i_k\mathbf{x}^i_k$. In the $N > 1$ case, designing the precoder $\mathbf{F}^i_k$ to reduce the IRI experienced at the $(i-1)$th node is futile since in practical systems $\eta_i \gg \alpha_i$, so we do not design the precoder to reduce IRI. At the same time, $\mathbf{F}^i_k$ can also be designed to improve the signal term at the following relay. In the following section, we produce an algorithm that determines how much the precoder $\mathbf{F}^i_k$ should be biased towards reducing SI at node i versus how much should $\mathbf{F}^i_k$ be biased towards improving the received signal quality at node $i+1$. As a final note about the system model, it is assumed there is full channel state information (CSI) at the transmitter and the receiver. The system model block diagram is summarized in Fig. 2.

B. Shortest Path Problem Formulation

The primary goal of the relay system is to leverage the transmit precoders at each relay to deliver ℓ packets of size b bits across N IBFD relays to the destination with the lowest overall latency possible. In this manuscript, the metric of latency we use is *channel uses*, since, for a fixed bandwidth, the more channel uses required to relay information corresponds to a longer overall transmission time. Therefore, we want to reduce the number of channel uses required to deliver $B = \ell b$ bits across the N-relay system. At the end of this subsection, we show how this minimization of channel uses can be expressed in a linear programming (LP) formulation of the shortest path problem.

We first introduce the notion of a *bit-deficit*. Let Δ^s_k, Δ^i_k, and Δ^d_k be the bit-deficit of the source, relay i, and the destination, respectively, which indicates how many bits the given node has yet to receive after k total channel uses across the relay. For example, the bit deficits before any transmission occurs are $\Delta^s_0 = 0$, $\Delta^i_0 = B$, and $\Delta^d_0 = B$, because the source has all of the packets and neither the relay(s) nor the destination have received any information. On the other hand, after some minimum, but unknown, amount of channel uses T, the bit deficits will be $\Delta^s_T = \Delta^1_T = \ldots = \Delta^N_T = \Delta^d_T = 0$, meaning all of the nodes have received all B bits and the transmission of information from source to destination is complete.

Since each relay can only transmit to the next consecutive relay, any information arriving at a relay comes only from its preceding relay (or the source, in the case of the first relay). Therefore, after any number of channel uses k, the bit deficits will naturally obey the inequality

$$\Delta^s_k \leq \Delta^1_k \leq \ldots \leq \Delta^N_k \leq \Delta^d_k \qquad \forall k \geq 0. \tag{6}$$

However, since the relay network is operating in the DF regime, then (6) must be strengthened, because the DF relay requires that every *packet* must arrive in full at a relay in order to be decoded and retransmitted. The number of *fully received* packets received at relay i and the destination are $\left\lfloor \frac{\Delta^i_k}{b} \right\rfloor$ and $\left\lfloor \frac{\Delta^d_k}{b} \right\rfloor$, respectively. It follows that the number of fully decoded packets relay i has received is $\left\lfloor \frac{\Delta^i_k}{b} \right\rfloor$ since a packet must be received in full to be decoded. Thus (6) is replaced with the stronger set of constraints

$$\begin{array}{lc} \left\lfloor \frac{\Delta^s_k}{b} \right\rfloor \leq \frac{\Delta^1_k}{b} & \text{and} \\ \left\lfloor \frac{\Delta^1_k}{b} \right\rfloor \leq \frac{\Delta^2_k}{b} & \text{and} \\ \quad\vdots & \\ \left\lfloor \frac{\Delta^N_k}{b} \right\rfloor \leq \frac{\Delta^d_k}{b} & \forall k \geq 0, \end{array} \tag{7}$$

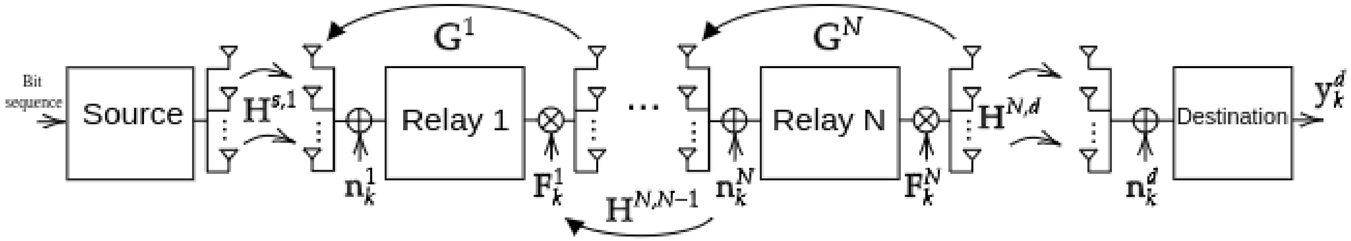

Fig. 2. Block Diagram of the IBFD MIMO separated relay network

since packets must be received in full at a given node before they can be transmitted to the next node. Due to the nature of the above constraint, we can conclude that the transmission is complete when $\Delta_k^d = 0$, because this implies all of the other nodes' bit deficits are also zero.

Clearly, after a channel use while $\Delta_k^d \neq 0$, it is expected that one (or more) of the Δ_k^i's or Δ_k^d will reduce. However, which of the Δ_k^i's or Δ_k^d will reduce and to what extent is dependent on which precoder is used at each relay and how much information is already accumulated at each node after channel use $k-1$. Let

$$\mathbf{s}_k = \left(\Delta_k^s, \Delta_k^1, \ldots, \Delta_k^d\right) \tag{8}$$

be the *vector state representation* of the bit deficits of each relay node after k channel uses, and denote $\mathcal{S}_k$ as the set of all possible vector state representations of the relay after k channel uses. We also denote $\mathcal{S}_0^k = \bigcup_0^k \mathcal{S}_k$ as the set of all possible states up to channel use k. Let $\mathcal{Q}^i$ denote the set of all possible precoder covariances at relay i. Then, we will refer to

$$u^i = \{\mathbf{Q}^{n_s}, \mathbf{Q}^{n_1}, \ldots, \mathbf{Q}^{n_N}\} \tag{9}$$

as a *control action* that describes precoder covariances that the source and relay(s) transmitters use during a channel use, where $\mathbf{Q}^{n_s} \in \mathcal{Q}^s$ and $\mathbf{Q}^{n_i} \in \mathcal{Q}^i$ are the source and ith relay's precoder, respectively. We denote $\mathcal{U}$ as the set of all possible control actions. Further, define

$$\pi(\mathbf{s}_k^j) = \{u^i, \ldots, u^t\} \tag{10}$$

as a *policy*, that is, a set of controls actions that when applied over each channel use after starting at $\mathbf{s}_0 = (0, B, \ldots, B)$ result in the state $\mathbf{s}_k^j$. Let T_{upper} be some finite, upper bound on the number of channel uses required to relay B bits to the destination (for example, if a *half-duplex* relaying policy is used then this is certainly an upper bound that is easily computed). Then there exists at least one policy $\pi(\mathbf{s}_{T_{upper}})$ where achieving a terminal state $\mathbf{s}_{T_{upper}} = (0, \ldots, 0)$ is possible (though, using our algorithm, the goal is to find the policy $\pi(\mathbf{s}_T)$ with the smallest T such that $\mathbf{s}_T = (0, \ldots, 0)$). Suppose then $G(\mathcal{S}_0^{T_{upper}}, \mathcal{E})$ is a directed graph that connects the initial relay state $\mathbf{s}_0 = \big(0, B, \ldots, B\big)$ and the terminal state $\mathbf{s}_{T_{upper}} = \big(0, \ldots, 0\big)$, where $\mathcal{E}$ is the set of all possible ordered pairs $(\mathbf{s}_k^i, \mathbf{s}_{k+1}^j)$ associated with a control action u^i that forms a directed arc between two elements of $\mathcal{S}_0^{T_{upper}}$.

With the above graph formulation in mind, we formalize the objective function to minimize latency. Let T be the minimum latency in terms of overall channel uses required to relay B bits from the source to the destination while satisfying the constraints in (7). By the way $G(\mathcal{S}_0^{T_{upper}}, \mathcal{E})$ is constructed, the constraints in (7) are implicit. That is, there only exist elements of $\mathcal{S}_0^{T_{upper}}$ that are physically possible given the set of control actions $\mathcal{U}$, and there only exist edges in $\mathcal{E}$ that are associated with possible state changes governed by the control actions in $\mathcal{U}$. Defining $\mathcal{E}_0 \subset \mathcal{E}$ as the set of all arcs that connect $\mathbf{s}_0 = \big(0, B, \ldots, B\big)$ and $\mathbf{s}_{T_{upper}} = (0, \ldots, 0)$, let δ_{ij} be an indicator variable such that

$$\delta_{ij} = \begin{cases} 1 & \text{if arc } (i,j) \in \mathcal{E}_0 \text{ is on a route to } \mathbf{s}_{T_{upper}} \\ 0 & \text{else} \end{cases} \tag{11}$$

and let d_{ij} be the delay of traversing arc (i,j) in terms of channel uses. We condition d_{ij} such that for every arc $(\mathbf{s}_k^i, \mathbf{s}_{k+1}^j)$

$$d_{ij} = \begin{cases} 1, & \mathbf{s}_k^i \neq (0, \ldots, 0) \\ 0, & \text{else} \end{cases} \tag{12}$$

Note that if there exists a policy $\pi(\mathbf{s}_T)$ with $T < T_{upper}$ such that $\mathbf{s}_T = (0, \ldots, 0)$, then the arcs associated with the controls of $\pi(\mathbf{s}_T)$ must also be on a route to $\mathbf{s}_{T,upper}$.

Then, T is the result of the LP formulation of the shortest path problem

$$\begin{aligned} T = \quad & \text{minimize } \textstyle\sum_{(i,j)\in\mathcal{E}_0} \; d_{ij}\delta_{ij} \\ & \text{subject to } \sum_{\alpha=\mathbf{s}_1^i\in\mathcal{S}_1} \delta_{\mathbf{s}_0\alpha} = 1, \\ & \sum_{\alpha=\mathbf{s}^i_{T_{upper}-1}\in\mathcal{S}_{T_{upper}+1}} \delta_{\alpha,\mathbf{s}_{T_{upper}}} = 1, \\ & \sum_{\alpha=\mathbf{s}^i_{k-1}\in\mathcal{S}_{k-1}} \delta_{\alpha\mathbf{s}^i_k} - \sum_{\beta=\mathbf{s}^i_{k+1}\in\mathcal{S}_{k+1}} \delta_{\mathbf{s}^i_k\beta} = 0 \end{aligned} \tag{13}$$

Note that the constraints in (13) ensure conservation-of-flow, guaranteeing continuity of a path through a given relay state in the graph $G(\mathcal{S}_0^{T_{upper}}, \mathcal{E})$ and ensuring the shortest path starts at $\mathbf{s}_0$ and ends at the terminal state [20]. We constructed the graph network such that each state occurs after one channel use. It follows that the solution of (13) with d_{ij} defined as above gives precisely the minimum latency to relay b bits from the source to the relay, because it finds the path that reaches the all zeros state *first*, which we denote $\mathbf{s}_T = (0, \ldots, 0)$, where T is the minimum number of channel uses to relay B bits to the destination.

Although (13) elucidates the optimization problem at hand, it is *highly* impractical to solve in the given form since it involves generating an entire graph $G(\mathcal{S}_0^{T_{upper}}, \mathcal{E})$ of large size (since T_{upper} is merely an upper bound) and brute force computing the cost of all of the paths to find which path reaches a terminal state $\mathbf{s}_T = (0, \ldots, 0)$ first, $T \leq T_{upper}$. In Section V, we propose an algorithm to compute the precoding policy that solves (13) in a more tractable manner.

IV. Rate and Latency Analysis

In this section we discuss a lower bound on the latency encountered relaying B bits across the separated relay system which is achievable in the very special case where there is no SI or IRI. An upper bound is presented in the following section under Lemma 3 as it relies on some knowledge of our approach to the optimization problem. We use the traditional Shannon capacity expression in this analysis, which are somewhat optimistic for finite blocklength codes but are traditionally used in capacity analysis. With this, the mutual information (MI) of any given channel from relay i to $i+1$ is given by [21]

$$\begin{aligned}&\mathcal{I}^{i,i+1}\left(\mathbf{Q}_k^i, \mathbf{Q}^{i+1}{}_k, \mathbf{Q}^{i+2}{}_k\right) \\ &\quad= \log\left|\mathbf{I} + \rho_{i+1}\mathbf{H}^{i,i+1}\mathbf{Q}_k^i\mathbf{H}^{i,i+1*}(\mathbf{\Sigma}_k^{i+1})^{-1}\right|,\end{aligned} \tag{14}$$

where the dependence on $\mathbf{Q}^{i+1}{}_k, \mathbf{Q}^{i+2}{}_k$ is implicit in Σ_k^{i+1} as defined in (5). Note that (14) assumes the presence of IRI; in the case where there is no IRI, $\mathbf{Q}_k^{i+2}$ can simply be set to $\mathbf{0}$. The capacity of the channel from relay i to $i+1$ is given by

$$\mathcal{C}^{i,i+1} = \log\left|\mathbf{I} + \rho_{i+1}\mathbf{H}^{i,i+1}\mathbf{Q}_{wf}^i\mathbf{H}^{i,i+1*}\right| \tag{15}$$

where $\mathbf{Q}_{wf}^i$ is the classical water-filling optimized covariance matrix [22]. The capacity in (15) is achievable when the SI and IRI is suppressed completely, or, mathematically, when $\mathbf{\Sigma}_k^{i+1} = \mathbf{I}$. This could occur, for example, when $\mathbf{F}_k^{i+1}$ is in the null space of $\mathbf{G}^{i+1}$ and $\mathbf{F}_k^{i+2}$ is in the null space of $\mathbf{H}^{i+2,i+1}$ for some non-zero η_{i+1}, α_{i+1}. The end-to-end (E2E) capacity of the separated relay with SI is given by [23]

$$\begin{aligned}C_{E2E} = &\max_{\mathbf{Q}^s, \mathbf{Q}^1, \ldots, \mathbf{Q}^d} \min\Big(\mathcal{I}^{s,1}\left(\mathbf{Q}^s, \mathbf{Q}^1, \mathbf{Q}^2\right), \\ &\mathcal{I}^{1,2}\left(\mathbf{Q}^1, \mathbf{Q}^2, \mathbf{Q}^3\right), \ldots, \mathcal{I}^{N,d}\left(\mathbf{Q}^N, \mathbf{Q}^d, \mathbf{0}\right)\Big),\end{aligned} \tag{16}$$

where $\mathbf{Q}^d$ is also set to $\mathbf{0}$ because the destination does not have a transmitter. In most cases, (16) is not a trivial optimization problem to solve because (14) is convex in $\mathbf{\Sigma}_k^{i+1}$ and concave in $\mathbf{Q}_k^i$ [22], [24]. However, a relatively simple upper bound on the E2E relay capacity can be formulated from (15) and (16). In the case there is no SI or IRI on any link, that is, $\mathbf{\Sigma}_k^j = \mathbf{I}, \ \forall \ k, j$, (16) becomes

$$C_{E2E}^u = \min\left(C^{s,1}, \ldots, C^{N,d}\right). \tag{17}$$

It is clear to see that the maximum E2E relay capacity is *bottlenecked* by the worst individual link. From this we can construct a (potentially very loose) lower bound on the number of channel uses required to relay ℓ packets of size b bits across the N separated relays by ignoring SI, which we will call $T_{N,lower}^{(\ell)}$. Let

$$t^{i,j} = \frac{b}{C^{i,j}} \tag{18}$$

be the minimum amount of channel uses required to relay one packet across the i,jth hop, that is, when there is no SI from the jth node and the precoder at the ith node is $\mathbf{Q}_{wf}^i$. It is not hard to see the minimum channel uses to relay one packet across the network is simply

$$T_{N,lower}^{(1)} = t^{s,1} + t^{1,2} + \ldots + t^{N,d}$$

because there is no decoding bottleneck encountered with a single packet since once the packet arrives in full at any relay it can immediately begin transmission to the next relay. However, there will be a bottleneck with $\ell > 1$ packets and therefore we can write a more general expression for a lower bound on the channel uses required to relay ℓ packets from source to destination through N relays.

Lemma 1: A lower bound on the number of channel uses T required to relay ℓ packets across the relay network with a packet size of b bits is given by

$$\begin{aligned}T_{N,lower}^{(\ell)} = &t^{s,1} + t^{1,2} + \ldots + t^{N,d} \\ &+ (\ell - 1)\max\{t^{s,1}, t^{1,2}, \ldots, t^{N,d}\}\end{aligned}$$

where $t^{i,j}$ is defined in (18). See Appendix A for proof.

Indeed, Lemma 1 is potentially quite optimistic for a lower bound, since it assumes the absence of SI and IRI, but it provides intuition of the notion of the rate bottleneck caused by the worst hop. Our algorithm gets closer to the true lower bound on the number of channel uses required by accounting for SI.

V. Dynamic Programming Optimization Approach

In Section III-B, we discussed how we can think of the low-latency relaying problem as a shortest path problem. In (13), we posed the objective function to solve for the lowest possible latency required to relay ℓ packets across the N separated relay network. However, solving (13) in its current form is massively inefficient. Instead, in this section, we discuss how we take a dynamic programming approach to solving (13). Specifically, since we are faced with a deterministic shortest path problem, we take an iterative approach where we begin with $\mathbf{s}_0$ and generate a directed graph $G(\mathcal{S}_0^{T_{upper}}, \mathcal{E})$ until a state is produced such that $\mathbf{s}_T = \mathbf{0}$, indicating the shortest path has been found. Beginning with $\mathbf{s}_0$, every control action $u_k^i \in \mathcal{U}$ is used to direct an arc from $\mathbf{s}_0$ to a new state $\mathbf{s}_1^i$, creating a new state set of states S_1 after one channel use. We do the same for each state $\mathbf{s}_1^i \in S_1$, and so on until the first instance of a state satisfying $\mathbf{s}_T = \mathbf{0}$ after some T channel uses. See Fig. 3 for a visual representation of $G(\mathcal{S}_0^{T_{upper}}, \mathcal{E})$. As discussed in Section III-B, the set of control actions $\mathcal{U}$ that govern the state changes in $G(\mathcal{S}_0^{T_{upper}}, \mathcal{E})$ correspond to which precoders are used during channel use k and therefore in Section V-A we show how we generate the set of usable precoders. In Section V-B we show how the state space is reduced for computational tractability. Finally in Section V-C we outline the algorithm in detail.

This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2023.3292985

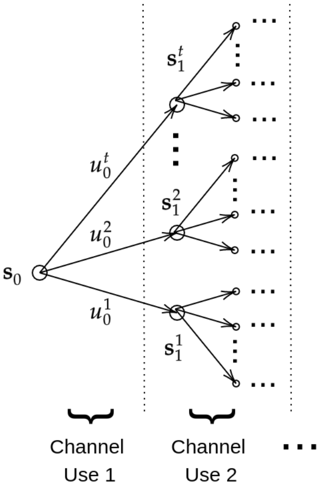

Fig. 3. Illustration of iteratively produced directed graph

A. Generating a Precoder Codebook

We address the first complexity issue with solving (13). The relay system's original state is $\mathbf{s}_0 = (0, B, \ldots, B)$. Now consider the cardinality of the state space $\mathcal{S}_1$, that is, the number of possible vector state representations of the relay after one channel use. This depends on the cardinality of $\mathcal{U}$, the set of all possible control actions which is equivalent to the number of precoder combinations we can choose to use at each transmitter during the first channel use. In other words, if the number of possible precoders is $\text{card}(\mathcal{Q}^i)$ at relay i, then $\text{card}(\mathcal{U}) = \text{card}(\mathcal{Q}^s)\prod_{t=1}^{N}\text{card}(\mathcal{Q}^t)$. If the only constraint we place on the precoders is that $Tr\left(\mathbf{F}_k^i\mathbf{F}_k^i\right) \leq 1$, then $\text{card}(\mathcal{U})$ is infinite, and, hence, $\text{card}(\mathcal{S}_1)$ is also infinite.

In order to solve (13) computationally tractibly, $\text{card}(\mathcal{U})$ should be finite, therefore we look to quantizing the set of precoder covariances a given transmitter may use. We address this problem by using the widely implemented *codebook-based precoding* approach, which limits the set of usable precoders to a finite set. The so-called quantized covariance precoding approach has in the past been used in limited feedback scenarios [21], [25]–[28]. In the case of this manuscript, we construct the codebook using a *SI thresholding approach* based off the joint precoding approach in Huberman *et. al* [29]. Such a codebook construction allows the maximization of the achievable rate of the ith relay with respect to the power constrained precoder, subject to an additional limitation on the power of the SI produced by the relay. In this construction, the codebook contains precoders that generate maximal SI and precoders that generate very minimal SI using *quantized* SI threshold increments. In a sense, the presented quantized covariance method allows computational feasibility at the expense of an *approximation* of the optimal solution to (13).

Denote the codebook at the ith relay transmitter by $\mathcal{Q}^i$ and let

$$\mathcal{I}^i\left(\mathbf{Q}^i\right) = \log\left|\mathbf{I} + \rho_i\mathbf{H}^{i,i+1}\mathbf{Q}^i\mathbf{H}^{i,i+1*}\right|$$

Then, we construct $\mathcal{Q}^i$ by solving the convex optimization problem

$$\max_{\mathbf{Q}^i} \mathcal{I}^i\left(\mathbf{Q}^i\right) \tag{19}$$

$$\text{subject to}$$

$$\text{Tr}\left(\mathbf{G}^i\mathbf{Q}^i\mathbf{G}^{i*}\right) \leq \tau, \text{Tr}\left(\mathbf{Q}^i\right) \leq 1$$

for each transmitting node i and for various values of τ depending on the desired codebook cardinality $|\mathcal{Q}^i|$. The SI term $\text{Tr}\left(\mathbf{G}^i\mathbf{Q}^i\mathbf{G}^{i*}\right) \leq \tau$ comes from $\text{Tr}(\mathbf{\Sigma}^i) = \text{Tr}(\mathbf{I} + \eta_i\mathbf{G}^i\mathbf{Q}^i\mathbf{G}^{i*} + \alpha_i\mathbf{H}^{i+1,i}\mathbf{Q}^{i+1}\mathbf{H}^{i+1,i*}) \approx \text{Tr}(\mathbf{I} + \eta_i\mathbf{G}^i\mathbf{Q}^i\mathbf{G}^{i*}) \leq \tau' \implies \text{Tr}(\mathbf{G}^i\mathbf{Q}^i\mathbf{G}^{i*}) \leq \tau$ by the idea that $\eta_i \gg \alpha_i$ in practical systems and the linearity of the matrix trace. We chose (19) in part due to its *convexity*, so it can be solved by off the shelf convex optimization software. In our simulations, (19) is solved using MATLAB CVX and the MOSEK solver with a CPU time to solve approximately .5 seconds on a Lenovo ThinkPad X1 Carbon Gen 8 with an Intel® Core$^{\text{TM}}$ i7-10510U CPU [30]–[32]. We note below a property of this codebook construction that allows the simplification of the state space.

Lemma 2: The objective function in (19) monotonically increases in $\tau > 0$ given a fixed precoder power constraint.

Proof: Consider the objective function in (19) and assume that the constraint $\text{Tr}(\mathbf{Q}) \leq 1$ is fixed. We first note that τ should always be greater than or equal to 0 since $\mathbf{\Sigma}^i$ is positive semi-definite. Let $\tau_{\max} = \text{Tr}(\mathbf{G}^i\mathbf{Q}_{wf}\mathbf{G}^{i*})$, which corresponds to the power of the SI produced when the optimal waterfilling precoder discussed in Section IV is used. Note that for any $\tau \geq \tau_{\max}$, the maximizing precoder of (19) will still be $\mathbf{Q}_{wf}$ since the power constraint is assumed fixed, which shows the result of the objective function is constant for $\tau \geq \tau_{max}$.

Now consider the behavior of (19) with $\tau \in [0, \tau_{\max})$. Denote the maximizer of (19) with SI threshhold τ as $\hat{\mathbf{Q}}_\tau$. Clearly for any $\tau \leq \tau_{wf}$ under the fixed precoder power constraint, we have that

$$\mathcal{I}^i(\hat{\mathbf{Q}}_\tau) \leq \mathcal{I}^i(\mathbf{Q}_{wf}).$$

Fix a $\tau \in [0, \tau_{\max})$ and let $\varepsilon \in [1, \frac{\tau_{\max}}{\tau}]$. Suppose the SI constraint is increased to $\tau\varepsilon$ (that is, $\text{Tr}(\mathbf{G}^i\hat{\mathbf{Q}}^i\mathbf{G}^{i*}) \leq \tau\varepsilon$). Then it follows that $\varepsilon\hat{\mathbf{Q}}_\tau$ (a simple power scaling of $\hat{\mathbf{Q}}_\tau$) is a feasible solution of (19) since

$$\text{Tr}(\mathbf{G}^i\hat{\mathbf{Q}}_\tau\mathbf{G}^{i*}) \leq \tau \implies \text{Tr}(\mathbf{G}^i\varepsilon\hat{\mathbf{Q}}_\tau\mathbf{G}^{i*}) \leq \varepsilon\tau.$$

Further, we have that

$$\mathcal{I}^i(\hat{\mathbf{Q}}_\tau) \leq \mathcal{I}^i(\varepsilon\hat{\mathbf{Q}}_\tau).$$

Since $\varepsilon\hat{\mathbf{Q}}_\tau$ is in the feasible region of (19) it follows that

$$\mathcal{I}^i(\varepsilon\hat{\mathbf{Q}}_\tau) \leq \mathcal{I}^i(\hat{\mathbf{Q}}_{\tau\varepsilon}) \implies \mathcal{I}^i(\hat{\mathbf{Q}}_\tau) \leq \mathcal{I}^i(\hat{\mathbf{Q}}_{\tau\varepsilon})$$

where $\hat{\mathbf{Q}}_{\tau\varepsilon}$ is the maximizer of (19) with SI constraint $\tau\varepsilon$. Thus we have shown that (19) increases monotonically with a fixed precoder power constraint. ■

As we will see, Lemma 2 allows us to reduce the state space. Lemma 2 also informs us that when we design our codebook we should constrain against values of $\tau \in [0, \tau_{\max}]$, since, as shown in the proof of *Lemma 2*, constraining SI to values of $\tau > \tau_{\max}$ in (19) will not produce a solution that can perform better in terms of maximum achievable rate than $\mathbf{Q}_{wf}$ under the power constraint $\text{Tr}(\mathbf{Q}^i) \leq 1$.

We also note an upper bound on the number of channel uses required to relay ℓ packets of size b bits based on the codebook construction.

Lemma 3: The HD solution is an upper bound on the number of channel uses required to relay ℓ packets across the relay network.

Proof: Include $\hat{\mathbf{Q}}_0$ (a precoder that produces no SI) and $\mathbf{Q}_{wf}$ in the codebooks at each relay, $\mathcal{Q}^i$ for $i = 1, \ldots, N$, and, at each relay 2 through $N-1$, include a precoder that produces no IRI in the codebook. Then the optimal relaying policy, at worst, will perform the same as a HD relaying scheme in terms of channel uses, because a HD scheme is implicit in the solution set. This is because the inclusion of one of the aforementioned precoders in each codebook guarantees that the capacity of each link is achievable since surrounding relays can be configured to produce no SI/IRI, as would be possible if a HD scheme were being utilized. ■

Therefore, if we include the aforementioned precoders in each relay's codebook, we guarantee the solution of our algorithm performs at least as well as HD, so that the latency incurred by HD is *indeed an upper bound* of the latency incurred in our algorithm. At relays 2 through N, a precoder that does not produce IRI can be designed by solving (19) except replacing the constraint $\text{Tr}(\mathbf{G}^i\mathbf{Q}^i\mathbf{G}^{i*}) \leq \tau$ with $\text{Tr}(\mathbf{H}^{i,i-1}\mathbf{Q}^i\mathbf{H}^{i,i-1*}) \leq \tau$ and evaluating at $\tau = 0$.

B. State Space Reduction

Since $\text{card}(\mathbf{Q}_i)$ is designed to be finite, $\text{card}(\mathcal{U})$ is also finite. However, the state space grows exponentially with respect to $\text{card}(\mathcal{U})$ as the number of channel uses increases if no state space simplification is done. For example, if $G(\mathcal{S}_0^{T_{upper}}, \mathcal{E})$ is constructed as described at the beginning of this section, without any sort of state pruning, $\text{card}(\mathcal{S}_k) = \text{card}(\mathcal{U})^k$. To mitigate this complexity issue, we show how we distinguish whether one state $\mathbf{s}_k^i$ is optimal over another state $\mathbf{s}_k^j$, $i \neq j$, after k channel uses such that we may omit suboptimal states to generate a reduced state space, $\mathcal{S}_k^{red.}$, so that $\text{card}(\mathcal{S}_k^{red.}) \ll \text{card}(\mathcal{U})^k$.

1) $\mathbf{s}_k^i = \mathbf{s}_k^j$: It may be that two distinct paths across $G(\mathcal{S}_0^{T_{upper}}, \mathcal{E})$ end in the same state after k channel uses. In such a case, if the cost of being at $\mathbf{s}_k^i$ is the same as the cost of being at $\mathbf{s}_k^j$, which they will have the same cost based on our definition of cost in (12), then $\mathbf{s}_k^i$ and $\mathbf{s}_k^j$ are equally optimal and therefore one may be omitted arbitrarily. This is an example of Bellman's Principle of Optimality [33].

2) $\mathbf{s}_k^i \neq \mathbf{s}_k^j$: The following insight also follows from Bellman's Principle of Optimality: if there are two states on a directed graph $G(\mathcal{S}_0^{T_{upper}}, \mathcal{E})$ describing an N relay system, then if it holds that $\mathbf{s}_i^k \preceq \mathbf{s}_j^k$, then $\mathbf{s}_i^k$ is optimal in terms of latency than $\mathbf{s}_j^k$. To see why this is the case, let $\mathbf{b}(\mathbf{s}_k^i)$ represent the current number of bits at the source and the relays after k channel uses (noting that this does not include the bits at the destination), where each entry of $\mathbf{b}(\mathbf{s}_k^i)$ can be computed by $\mathbf{b}(\mathbf{s}_k^i)[n] = \mathbf{s}_k^i[n+1] - \mathbf{s}_k^i[n]$. Now suppose that $\mathbf{b}(\mathbf{s}_k^i)[n] = \mathbf{b}(\mathbf{s}_k^j)[n]$ for $n \geq 2$ and $\mathbf{b}(\mathbf{s}_k^i)[1] < \mathbf{b}(\mathbf{s}_k^j)[1]$. Then, if it takes a minimum of m channel uses to get $\mathbf{b}(\mathbf{s}_k^i)[1]$ bits from the source to the destination (noting that the bits at the relays must reach the destination before the bits at the source due to causality), then it will take greater than m channel uses to deliver $\mathbf{b}(\mathbf{s}_k^j)[1]$ bits to the destination. Hence, $\mathbf{s}_k^i$ is optimal with respect to $\mathbf{s}_k^j$ in terms of latency.
Suppose now that $\mathbf{b}(\mathbf{s}_k^i)[n] \leq \mathbf{b}(\mathbf{s}_k^j)[n]$ for any $n \geq 2$ and still $\mathbf{b}(\mathbf{s}_k^i)[1] < \mathbf{b}(\mathbf{s}_k^j)[1]$. Then the minimum number of channel uses to get $\mathbf{b}(\mathbf{s}_k^i)[1]$ bits from the source to destination is less than or equal to m, since the bits at the relay have to wait for less bits at the relays to be delivered to the destination first. Hence, $\mathbf{s}_k^i$ remains the optimal relative to $\mathbf{s}_k^j$. From the perspective of the bit deficit, this corresponds to when $\mathbf{s}_k^i \preceq \mathbf{s}_k^j$.

3) Further Simplifications: Noting that (19) is monotonically increasing in τ, we exploit this to reduce the state space as we generate the new states. Let the current state under inspection be $\mathbf{s}_k^j$ and again let the bits at the source and each relay of $\mathbf{s}_k^j$ be represented by $\mathbf{b}(\mathbf{s}_k^j)$. Since we established by causality that we cannot send bits that have not already arrived at the relay, the only precoders of interest at relay i are those precoders in the set

$$\left\{\mathbf{Q} \in \mathcal{Q}^i : \ \mathcal{I}^i(\mathbf{Q}) < \mathbf{b}(\mathbf{s}_k^j)[n]\right\}$$

and

$$\mathbf{Q} = \arg\min_{\mathbf{Q} \in \mathcal{Q}^i} \mathcal{I}^i(\mathbf{Q}) : \ \mathcal{I}^i(\mathbf{Q}) \geq \mathbf{b}(\mathbf{s}_k^j)[n].$$

We are interested in the last precoder because it is the precoder that allows sending all the bits at relay i with the least amount of interference generated. The former set of precoders allows transmission with limited SI at the expense of the transmission of only a portion of the bits at the relay. We denote this subset of $\mathcal{Q}^i$ as $\hat{\mathcal{Q}}^i(\mathbf{s}_k^j)$, and note that using any precoders $\mathbf{Q} \in \mathcal{Q}^i \setminus \hat{\mathcal{Q}}^i(\mathbf{s}_k^j)$ simply generates extra SI while not providing any rate gain at the ith relay. Therefore, when generating $\mathcal{S}_{k+1}$, we restrict the control actions connecting the $\mathbf{s}_k^j$ to new states in $\mathcal{S}_{k+1}$ to $\hat{\mathcal{U}}(\mathbf{s}_k^j) = \mathcal{Q}^r \times \hat{\mathcal{Q}}^1(\mathbf{s}_k^j) \times \ldots \times \hat{\mathcal{Q}}^N(\mathbf{s}_k^j)$.

C. Proposed Algorithm

With the notion of a precoder codebook and a method to reduce state space dimensionality, we describe the proposed algorithm in detail. For an N relay system, given parameters $\mathbf{G}^i$, $\mathbf{H}^{i,j}$, ℓ, and b, the algorithm is an iterative process to find the policy $\pi(\mathbf{s}_T) = \{u_i, \ldots, u_t\}$ that ends in the terminal state in the fewest number of channel uses T.

First, the codebook is constructed based on the channel and the desired codebook dimensionality. For a desired $\text{card}(\mathcal{Q}^i)$, a codebook at each relay transmitter is constructed by solving (19), starting with $\tau = \text{Tr}(\mathbf{G}^i\mathbf{Q}_{wf}\mathbf{G}^{i*})$ and decreasing τ until $\tau = 0$ in intervals such that the τ values satisfy equally spaced rate intervals from 0 to $\log|\mathbf{I} + \mathbf{H}^{i,i+1}\mathbf{Q}_{wf}\mathbf{H}^{i,i+1*}|$. Next, the initial parameters are set. We set the initial state to be $\mathbf{s}_0 = (0, B, \ldots, B)$, meaning that all of the information is at the source node and neither the relays 'nor the destination have received any packets. From here, we construct the directed graph $G(\mathcal{S}_0^{T_{upper}}, \mathcal{E})$ on a per channel use basis, and terminate when a finite-cost path terminates in $\mathbf{s}_T = (0, 0, \ldots, 0)$. That is, beginning from $\mathbf{s}_0 = (0, B, \ldots, B)$, each $u^t \in \hat{\mathcal{U}}(\mathbf{s}_0)$ is used to generate the next set of states after one channel use, $\mathcal{S}_1$. Since each control action is an $N+1$-tuple of precoders, where each element represents the precoder to be used at the source, relay 1, and so on, then the expression to generate a new state $\mathbf{s}^j_{k+1}$ originating from $\mathbf{s}^i_k$ using a control action $u^t = (\mathbf{Q}^r, \ldots, \mathbf{Q}^N) \in \hat{\mathcal{U}}(\mathbf{s}^i_k)$ is

$$\mathbf{s}^j_{k+1}[n] = \mathbf{s}^i_k[n] - b_{rx}\left(\mathbf{s}^i_k[n-1]\right)$$

where b_{rx} is the bits received at relay n after channel use $k+1$ which can be computed as

$$b_{rx}\left(\mathbf{s}^i_k[n-1]\right) = \min\Bigg(\mathcal{I}^{n-1}(\mathbf{Q}^{n-1}, \mathbf{Q}^n), b\left\lfloor \mathbf{s}^i_k[n-1]/b \right\rfloor \\ - \sum_{t=n+1}^{N+2} \mathbf{s}^i_k[t]\Bigg)$$

where $\mathcal{I}^{n-1}(\mathbf{Q}^{n-1}, \mathbf{Q}^n)$ is the maximum achievable rate from relay $n-1$ to relay n and $b\left\lfloor \mathbf{s}^i_k[n-1]/b \right\rfloor - \sum_{t=n+1}^{N+2} \mathbf{s}^i_k[t]$ gives the bits at the $n-1$th relay that are available to send (those that have been decoded from a fully-received packet).

At this point, $\mathcal{S}_1$ is reduced by removing all states $\mathbf{s}^j_k$ such that $\mathbf{s}^i_k \preceq \mathbf{s}^j_k$ leaving a reduced state space $\mathcal{S}^{red.}_1$. Then, for each $\mathbf{s}^i_k \in \mathcal{S}^{red.}_1$, each control action $u^t \in \hat{\mathcal{U}}(\mathbf{s}^i_k)$ is used to generate the set of states after two channel uses $\mathcal{S}_2$, and the process repeats. The algorithm ends when, after T iterations, there exists a state $\mathbf{s}_T$ such that $\mathbf{s}_T = (0, 0, \ldots, 0)$. T reflects the minimum number of channel uses required to relay the ℓ packets to the destination using the codebooks $\mathcal{Q}^i$ that we constructed. Further, the selection of precoders associated with the arcs in $G(\mathcal{S}^{T_{upper}}_0, \mathcal{E})$ connecting $\mathbf{s}_0$ to $\mathbf{s}^i_T$ is the lowest latency relaying policy given the constructed codebooks. The algorithm is summarized in Algorithm 1.

VI. Simulation Results

In this section, we simulate the dynamic programming algorithm. We first test the performance of the algorithm for the simple separated relay case, and then extend our results to two and three relay systems where we use the traditional infinite blocklength rate expressions in (14). Finally, we substitute finite blocklength expressions in our algorithm and observe the performance degradation. In the simulations, we normalize against $2T^{(\ell)}_{N,lower}$ (the HD bound since orthogonal communication resources can be used). Thus the upper bound on channel uses is 1 and the lower bound is 0.5.

A. Separated Relay: Channel Uses versus SI (η)

In this section, we compare the performance of our proposed method compared to the channel uses required to relay the packets using HD for a range of SI levels. We set $N=1$, which corresponds to a separated relay. We examine the cases where $M^s_{tx}{=}M^1_{rx}{=}M^1_{tx}{=}M^d_{rx}{=}\ 2$, $M^s_{tx}{=}M^1_{rx}{=}M^1_{tx}{=}M^d_{rx}{=}$ 4, and $M^s_{tx}{=}M^1_{rx}{=}M^1_{tx}{=}M^d_{rx}{=}\ 8$. The SNR at each link is set to 10 dB, that is, $\rho_1 = \rho_d = 10$, and the INR (η) is swept from 20 to 50 dB at the relay transmitter. It is assumed that $\mathbf{H}^{s,1}$, $\mathbf{H}^{1,d}$, and $\mathbf{G}^1 \sim \mathcal{C}N(0, \mathbf{I})$. We test three sets of packet size configurations, including the case where $\ell = 10$ and $b = 10$ (equal number of packets and bits per packet), $\ell = 25$ and

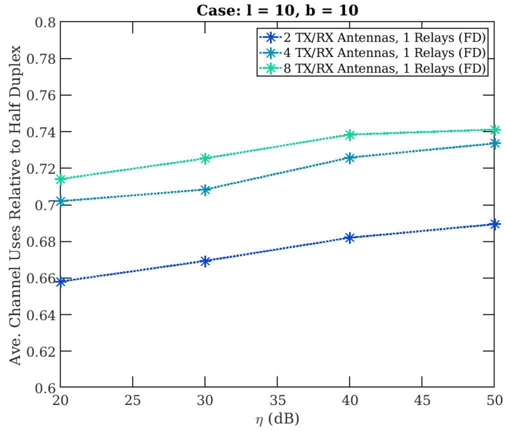

Fig. 4. Average number of channel uses required to relay packets across separated relay for 2, 4, and 8 antenna setups and $\ell = 10$, $b = 10$.

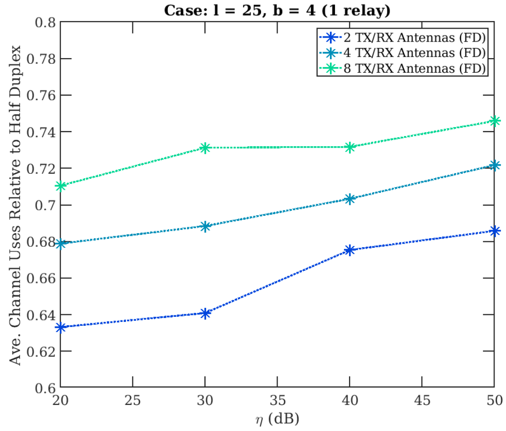

Fig. 5. Average number of channel uses required to relay packets across separated relay for 2, 4, and 8 antenna setups and $\ell = 25$, $b = 4$.

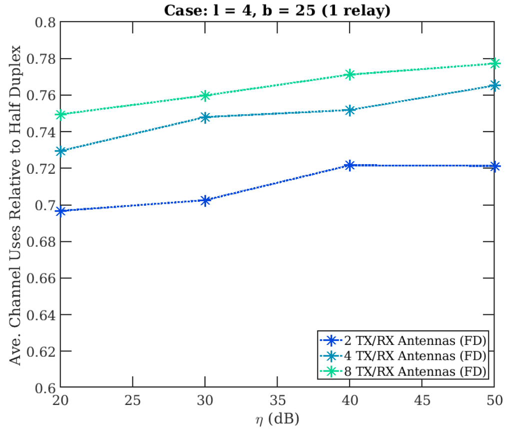

Fig. 6. Average number of channel uses required to relay packets across separated relay for 2, 4, and 8 antenna setups and $\ell = 4$, $b = 25$.

Algorithm 1: Finding Optimal Precoding Strategy

```
Initialize
  $\mathbf{H}^{s,1},\ldots,\mathbf{H}^{N,d},\mathbf{H}^{2,1},\ldots,\mathbf{H}^{N-1,N},\mathbf{G}^{1},\ldots,\mathbf{G}^{N}$,
  $\rho_r,\rho_1,\ldots,\rho_N,\eta_1,\eta_2,\ldots,\eta_N,\alpha_1,\ldots,\alpha_{N-1},b,\ell$;
Step 1 Generate finite set of precoder covariances at
  each relay, $\mathcal{Q}^i$ for $i=1,\ldots,N$:
for $\tau = 0 \to Tr\left(\boldsymbol{G}^i\boldsymbol{Q}_{wf}\boldsymbol{G}^{i*}\right)$ do
    $\mathbf{Q}^t \in \mathcal{Q}^i = \begin{array}{ll}\max\limits_{\mathbf{Q}} & \mathcal{I}^i(\mathbf{Q}) \\ \text{s. t.} & Tr\left(\mathbf{G}^i\mathbf{Q}\mathbf{G}^{i*}\right)\leq\tau,\ Tr\left(\mathbf{Q}\right)\leq 1\end{array}$
end
if $N > 1$ then
    For relays 2 through $N-1$ replace
      $\mathrm{Tr}(\mathbf{G}^i\mathbf{Q}^i\mathbf{G}^{i*})\leq\tau$ with
      $\mathrm{Tr}(\mathbf{H}^{i,i-1}\mathbf{Q}^i\mathbf{H}^{i,i-1*})\leq\tau$ in (19) and evaluate at
      $\tau=0$. Add resulting precoder to relay codebook.
end
Step 2 Find lowest latency precoder sequence:
Set initial state $\mathbf{s}_0=(0,B,\ldots,B)$, initial policy
  $\pi(\mathbf{s}_0)=\{\}$, and M = 1 as initial number of states;
while $\boldsymbol{s}_k^i \neq \mathbf{0}\ \forall\ i=1,\ldots,M$ do
    Apply precoder covariances to each existing state
      $\mathbf{s}_k^\alpha \in S_k$:
    Initialize new state counter $x=0$;
    for Each $\boldsymbol{s}_k^\alpha \in S_k$ do
        Get reduced set of control actions $\hat{\mathcal{U}}(\mathbf{s}_k^\alpha)$;
        for Each $u^t \in \hat{\mathcal{U}}(\boldsymbol{s}_k^\alpha)$ do
            Update state counter: $x$++;
            Generate new state:
              $\mathbf{s}_{k+1}^x[n]=\mathbf{s}_k^\alpha[n]-b_{rx}\left(\mathbf{s}_k^\alpha[n-1]\right)$;
            Record control actions for new state:
              $\pi\left(\mathbf{s}_{k+1}^x\right)=\{\pi\left(\mathbf{s}_k^\alpha\right),u^t\}$;
        end
    end
    Reduce State Space:
    if $\boldsymbol{s}_k^i \preceq \boldsymbol{s}_k^j,\ i\neq j$ then
        Delete $\mathbf{s}_k^i$;
    end
    Update the new current number of states: $M=x$;
end
Get optimal precoder sequence:
Find $\mathbf{s}_k^i$ such that $\mathbf{s}_k^i=\mathbf{0}$;
Optimal sequence of precoders is $\pi(\mathbf{s}_k^i)$;
```

$b=4$ (more packets of smaller size), and $\ell=4$ and $b=25$ (fewer packets of larger size). Finally, card$(\mathcal{Q}^1)=10$ at the relay transmitter codebook and card$(\mathcal{Q}^s)=1$ (that is, it always transmits using $\mathbf{Q}_{wf}$ since it does not produce SI). We average the results over 100 random channel realizations per antenna regime, SI level, and packet size configuration.

We note that the performance of the algorithm, in general, is in general slightly better for the $\ell=25$ and $b=4$ case, and then slightly degrades with the increase of packet size with the $\ell=10$ and $b=10$ case slightly worse than the $\ell=25$ and $b=4$ case, and the $\ell=4$ and $b=25$ case performing the worst. In each case, the more antennas each transmitter is equipped with, the lesser improvement there is *relative* to HD. This could be due to the notion that the capacity scales positively with more TX/RX antennas and so the overall HD performance is better with more antennas. As well, there is an expected performance degradation when η is increased since SI deteriorates IBFD relay performance. In any case, there is performance gain when using the dynamic precoding algorithm relative to HD, regardless of the self-interference level η, which follows from Lemma 3 since the HD precoding strategy is in the solution space. In any of the packet size regimes, there is approximately a 20% to 40% reduction in channel uses relative to HD, with the highest reduction in channel uses in the 2 antenna case and lowest INR. This suggests that our algorithm would be especially beneficial in networks that contain smaller devices that have limited numbers of antennas.

B. Extension to Multiple Relays

In this section, we demonstrate the extension of the algorithm to $N > 1$ relays, showing the performance of the algorithm for multiple relays. Specifically, we show the performance of the algorithm for the case of M_{tx}^s=M_{rx}^d=M_{rx}^i=M_{tx}^i=2 when there are 1, 2, and 3 relays. In this simulation, we assume $\ell=20$ and $b=1$, card$(\mathcal{Q}^i)=10$, card$(\mathcal{Q}^s)=1$ (for the same reason as the previous subsection), and $\mathbf{H}^{s,1}$, $\mathbf{H}^{i-1,i}$, $\mathbf{H}^{1,d},\mathbf{H}^{2,1},\mathbf{H}^{3,2}$ and $\mathbf{G}^i\sim\mathcal{C}N(0,\mathbf{I})$. It is justifiable to assume that $\mathbf{H}^{2,1},\mathbf{H}^{3,2}$ are not the reciprocal channel of $\mathbf{H}^{1,2},\mathbf{H}^{2,3}$ since the RF chain of the transmitter and receiver at each node are assumed to be separate. We assume $\rho_i=\rho_d=\alpha_i=10$ dB. For each simulation case, we average the results over 100 random channel realizations.

Interestingly, the best performance gain relative to HD is demonstrated in the single relay case, and then the performance relative to HD diminishes (though a performance gain is still observed in all cases). In general, the 1 relay, 2 relay, and 3 relay case demonstrates a $28-33\%$, $21-23\%$, and $10-15\%$ reduction in channel uses, respectively. This is potentially due to the notion that in HD, data can still be sent across non-adjacent hops without interference occurring in the $N > 1$ cases but also because in the $N > 1$ relays, there is ISI present on top of SI. For example, in the 2 relay case, data can be sent across the first and third hop simultaneously without the first relay receiver experiencing interference from the second relay transmitter by assumption of the system model. In each of the cases, there is performance gain observed over HD, but it is much less dramatic than the $N=1$ case, indicating that our algorithm may be especially useful in networks with fewer hops such as shorter range M2M networks. Regardless, a performance gain would still be observed for multiple relay systems such as rural wireless backhaul.

C. Effect of Short Packet Capacity Expressions

In this section, we show the performance difference of our relaying scheme when using finite blocklength capacity expressions for various decoding error probabilities ε [9]. For the simulation results in Sections VI-A and VI-B, the capacity expression in (15) was used to model the achievable rate of

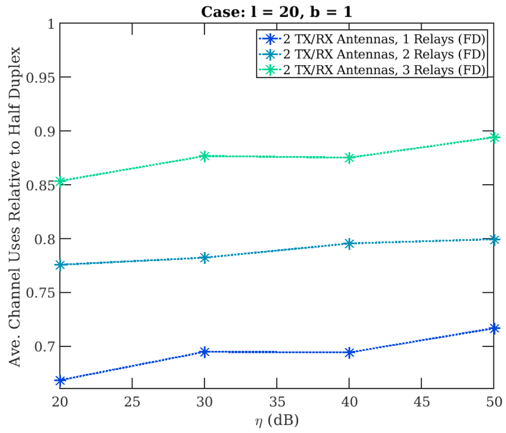

Fig. 7. Average number of channel uses required to relay packets across 1, 2, 3 relays for 2 antenna setups and $\ell = 20$, $b = 1$.

the channels, which assumes asymptotically large blocklength. In this simulation, we still design the codebook using the optimization problem in (19) with infinite blocklength expressions. However, when solving the shortest path problem as outlined in Algorithm 1, we model the channel using finite blocklength expressions and observe the performance loss of the algorithm. The non-asymptotic achievable rate for the parallel AWGN channel the rate is given by [34] [Thm. 78]

$$R^*(n,\varepsilon) = C_L(P) - \sqrt{V_L(P)/n}Q^{-1}(\varepsilon) + O\left(\frac{\log n}{n}\right) \tag{20}$$

where

$$C_L(P) = \sum_{i=1}^{L} C\left(\frac{W_i}{\sigma_i^2}\right) \tag{21}$$

and

$$V_L(P) = \sum_{i=1}^{L} V\left(\frac{W_i}{\sigma_i^2}\right) \tag{22}$$

where V_L and C_L are the capacity and dispersion of the AWGN channels, respectively, W_i is the power allocation for the ith parallel channel, σ_i^2 is the noise power of the ith parallel channel, and ε is the average desired error probability. In the simulation we approximate the remainder terms with normal approximation [2], [34].

We fit our model into the above expressions as so. Note that for any fixed $\mathbf{Q}_k^i$ and $\mathbf{Q}_k^j$, we have that the mutual information is given by

$$\mathcal{I}^{ij}\left(\mathbf{Q}_k^i, \mathbf{Q}_k^j\right) = \log\left|\mathbf{I} + \rho_j \mathbf{H}^{ij}\mathbf{Q}_k^i\mathbf{H}^{ij*}(\Sigma_k^j)^{-1}\right|$$

and note that $(\Sigma_k^j) = \mathbf{R}^*\mathbf{R}$ given by the Cholesky decomposition so then $(\Sigma_k^j)^{-1} = \mathbf{R}^{-1}(\mathbf{R}^*)^{-1}$. Hence

$$\begin{aligned}\mathcal{I}^{ij}\left(\mathbf{Q}_k^i, \mathbf{Q}_k^j\right) &= \log\left|\mathbf{I} + \rho_j(\mathbf{R}^*)^{-1}\mathbf{H}^{ij}\mathbf{Q}_k^i\mathbf{H}^{ij*}\mathbf{R}^{-1}\right| \\ &= \log\left|\mathbf{I} + \rho_j((\mathbf{R}^*)^{-1}\mathbf{H}^{ij}\mathbf{F}_k^i)((\mathbf{R}^{-1})^*\mathbf{H}^{ij}\mathbf{F}_k^i)^*\right|\end{aligned}$$

which can be diagonalized and expressed as

$$\mathcal{I}^{ij}\left(\mathbf{Q}_k^i, \mathbf{Q}_k^j\right) = \sum_t \log(1 + \rho_j\lambda_t^2)$$

where λ_t is the tth singular value of $((\mathbf{R}^*)^{-1}\mathbf{H}^{ij}\mathbf{F}_k^i)$. Referring back to (20), $C_L(P) = \sum_t \log(1 + \rho_j\lambda_t^2)$ and $V_L(P) = \sum_t \frac{\rho_j\lambda_t^2(\rho_j\lambda_t^2+2)}{(\rho_j\lambda_t^2+1)^2}\log^2 e$ [34]. Given an expression to compute the finite blocklength capacity of the channel, we may now compute the performance of our algorithm using (20).

In our simulation, we assume that $\varepsilon = 10^{-5}$ and 10^{-3} and that $\ell = b = 10$, $N = 1$, $\rho_1 = \rho_d = 10$ dB, $\text{card}(\mathcal{Q}^1) = 10$, $\text{card}(\mathcal{Q}^s) = 1$ as before, and $\mathbf{H}^{s,1}$, $\mathbf{H}^{1,d}$, and $\mathbf{G}^1 \sim CN(0, \mathbf{I})$. We also assume that M_{tx}^s=M_{rx}^d=M_{rx}^1=M_{tx}^1=2. We average the results over 100 random channel realizations for each SI value and ε value.

Relative to HD, it can be seen that the algorithm performance is slightly worse when the simulations use short packet expressions than when asymptotic blocklength capacity expressions are used, with the worst performance degradation observed with the lowest target decoding error probability $\varepsilon = 10^{-5}$. In any case, the performance degradation relative to HD is remarkably mild. Indeed, when $\varepsilon = 10^{-3}$, the algorithm performs approximately $1 - 1.5\%$ worse relative to HD than when the achievable rate of the channels is modeled with asymptotic blocklength expressions, and when $\varepsilon = 10^5$, the algorithm performs approximately $1 - 2\%$ worse relative to HD than when the achievable rate of the channels is modeled with asymptotic blocklength expressions. This suggests that a codebook design using asymptotic capacity expressions such as presented in (19) will still yield a $30 - 33\%$ channel use reduction compared to using a HD relaying scheme.

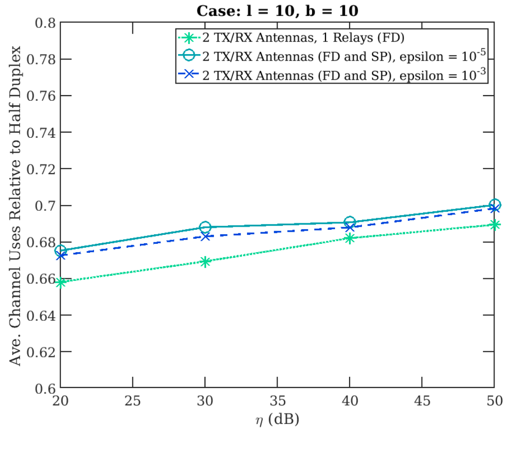

Fig. 8. Average number of channel uses required to relay packets across one relay with 2 antenna setups and $\ell = 10$, $b = 10$, where SP stands for *Short Packet* capacity expressions.

VII. Conclusion

In this manuscript, we framed the task of precoding for low-latency communication as a shortest path problem. We presented an iterative dynamic programming based algorithm for determining the lowest-latency selection of precoders at the source and relay transmitters for an N separated relay system

This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2023.3292985

on a per channel uses basis, allowing for a more flexible precoder selection than previous works. Our simulation results show the best performance improvement for the case of 2 antennas and 1 relay, whereas increasing relays and antennas still exhibits performance gains but in diminishing proportion to HD. This seems to suggest that the presented algorithm would be best suited for devices limited in antennas and for shorter range M2M systems, though a performance gain would still be observed for longer range systems such as rural wireless backhaul. Our simulation results also show that our algorithm and presented codebook design would still offer $\sim 30\%$ reduction in channel uses relative to HD when we model the channel with short packet capacity expressions.

This work offers *numerous* interesting future research directions. For example, this work assumes a fixed packet size, but an extension that would allow dynamic packet sizing would be useful for practical systems. This algorithm also does not scale well with large N in that the state space may grow prohibitively large even given the heuristic simplifications presented. In cases such as simple integrated access and backhaul (IAB) with two hops or frugal-networks aimed at bringing rural access with constraints on number of relay nodes, it may not be necessary to scale with large N [35]–[37]. In other cases, further complexity reduction without greatly sacrificing algorithm performance is a valuable future direction. Additionally, this work takes a centralized approach in finding the lowest-latency precoding policy across an N separated relay network. That is, the computer solving the optimization problem is assumed to know all of the propagation channels along the N relay system. In the future, it may be interesting to devise a *decentralized* algorithm where each node selects the precoder with limited knowledge of the other nodes' channels. Though our algorithm may fare well in situations where wireless rural backhaul, some urban backhaul deployments, and static sensor network deployments will have longer channel coherence times and thus have more reliable channel state information tracking, scenarios where relay nodes may be moving quickly may necessitate a more efficient codebook optimization problem than (19) .

Appendix A
Proof of Lemma 1

We prove by induction the lower bound on T for N relays and ℓ packets, given by $T^{(\ell)}_{N,lower}$, as discussed in Lemma 1. Assume that once a packet has arrived at a given relay, it immediately begins transmission to the next relay as long as another packet is not also transmitting to the next relay (in other words, there is no processing delay and the source of delay in this system results only from waiting for prior packets to complete transmission).

A. One Relay Base Case

We consider first the simple case of lower bounding the channel uses required to relay ℓ packets through one relay to the destination. Clearly, the first packet traverses the first hop in $t^{s,1}$ channel uses and the second hop in $t^{1,d}$ channel uses. Therefore, packet 1 arrives at the destination in $t^{s,1} + t^{1,d}$ channel uses. Packet two must wait $t^{s,1}$ channel uses before it can traverse the first hop because it must wait for packet 1 to complete transmission. After $2t^{s,1}$ channel uses, packet 2 arrives at the first relay.

Now we observe two cases: $t^{s,1} \geq t^{1,d}$ or $t^{s,1} < t^{1,d}$. If $t^{s,1} \geq t^{1,d}$, then packet 1 will have already arrived at the destination by time packet 2 arrives at relay 2, and, thus, packet 2 may immediately begin transmission across the second hop. Packet 2 arrives at the destination in $2t^{s,1} + t^{1,d}$ channel uses. However, if $t^{s,1} < t^{1,d}$, then packet 2 must wait $(t^{1,d} + t^{s,1}) - 2t^{s,1} = t^{1,d} - t^{s,1}$ channel uses before it can transmit across the second hop. Thus, packet 2 arrives at the destination in $t^{s,1} + 2t^{1,d}$ channel uses. We can write $T^{(2)}_{1,lower}$ more generally as

$$T^{(2)}_{1,lower} = t^{s,1} + t^{1,d} + \max\{t^{s,1}, t^{1,d}\}. \tag{23}$$

Then we see that for ℓ packets we have that it takes the ℓth packet $\ell t^{s,1}$ channel uses to reach the first relay. Again, if $t^{s,1} \geq t^{1,d}$, then the $\ell - 1$ previous packets will have already arrived at the destination, hence, $T^{(\ell)}_{1,lower} = \ell t^{s,1} + t^{1,d}$. For the case $t^{s,1} < t^{1,d}$, we have that $T^{(\ell)}_{1,lower} = \ell t^{s,1} + t^{1,d} + (T^{\ell-1}_{1,lower} - \ell t^{s,1})$ where the last grouping of terms is the channel uses the ℓth packet must wait before it can begin transmission to the destination. Hence

$$\begin{aligned} T^{(\ell)}_{1,lower} &= t^{1,d} + T^{(\ell-1)}_{1,lower} \\ &= t^{1,d} + (\ell - 3)t^{1,d} + T^{(2)}_{1,lower} \\ &= t^{1,s} + \ell t^{1,d}. \end{aligned}$$

Therefore, for the *separated relay*, we have

$$T^{(\ell)}_{1,lower} = t^{s,1} + t^{1,d} + (\ell - 1)\max\{t^{s,1}, t^{1,d}\}. \tag{24}$$

B. N Relay, ℓ Packet Lower Bound

Now we generalize to N relays where we will use (24) as the base case for the following inductive argument. Suppose the channel uses required for ℓ packets to reach the nth relay in full through $n - 1$ relays is given by

$$\begin{aligned} T^{(\ell)}_{n-1,lower} &= t^{s,1} + t^{1,2} + \ldots + t^{n-1,n} \\ &\quad + (\ell - 1)\max\{t^{s,1}, t^{1,2}, \ldots, t^{n-1,n}\}. \end{aligned} \tag{25}$$

Then, it follows that the channel uses required for ℓ packets to reach the $n + 1$th relay through n relays is given by

$$\begin{aligned} T^{(\ell)}_{n,lower} &= t^{s,1} + t^{1,2} + \ldots + t^{n,n+1} \\ &\quad + (\ell - 1)\max\{t^{s,1}, t^{1,2}, \ldots, t^{n,n+1}\}. \end{aligned} \tag{26}$$

To see why this is, note that by hypothesis packet ℓ arrives at relay n in full in $T^{(\ell)}_{n-1,lower}$ channel uses. Likewise, packet $\ell - 1$ arrives at relay $n + 1$ in full in $T^{(\ell-1)}_{n,lower}$ channel uses. We observe two cases: 1) *packet $\ell - 1$ has arrived in full at relay $n + 1$ by the time packet ℓ has arrived in full at relay n* or 2) *packet $\ell - 1$ has NOT arrived in full at relay $n + 1$*

by the time packet ℓ has arrived in full at relay n. In the first case, it follows that

$$\begin{aligned} T^{(\ell-1)}_{n,lower} &\leq T^{(\ell)}_{n-1,lower} \\ &\implies t^{n,n+1} + (\ell-2)\max\{t^{s,1},\ldots,t^{n,n+1}\} \\ &\qquad \leq (\ell-1)\max\{t^{s,1},t^{1,2},\ldots,t^{n-1,n}\} \\ &\implies t^{n,n+1} \leq \max\{t^{s,1},t^{1,2},\ldots,t^{n-1,n}\}. \end{aligned} \tag{27}$$

Since the ℓth packet does not have to wait for the $\ell-1$th packet to transmit to the $n+1$th relay, we have that

$$\begin{aligned} T^{(\ell)}_{n,lower} &= t^{n,n+1} + T^{(\ell)}_{n-1,lower} \\ &= t^{s,1} + t^{1,2} + \ldots + t^{n,n+1} \\ &\qquad + (\ell-1)\max\{t^{s,1},t^{1,2},\ldots,t^{n-1,n}\} \\ &= t^{s,1} + t^{1,2} + \ldots + t^{n,n+1} \\ &\qquad + (\ell-1)\max\{t^{s,1},t^{1,2},\ldots,t^{n,n+1}\} \end{aligned} \tag{28}$$

where the last equality comes from (27).

In the second case, it follows that

$$\begin{aligned} T^{(\ell-1)}_{n,lower} &> T^{(\ell)}_{n-1,lower} \\ &\implies t^{n,n+1} + (\ell-2)\max\{t^{s,1},\ldots,t^{n,n+1}\} \\ &\qquad > (\ell-1)\max\{t^{s,1},t^{1,2},\ldots,t^{n-1,n}\} \\ &\implies t^{n,n+1} = \max\{t^{s,1},\ldots,t^{n,n+1}\}. \end{aligned} \tag{29}$$

Further, packet 2 must wait $T^{(\ell-1)}_{n,lower} - T^{(\ell)}_{n-1,lower}$ channel uses before it can begin transmission to the $n+1$th relay. Hence, for the second case

$$\begin{aligned} T^{(\ell)}_{n,lower} &= T^{(\ell)}_{n-1,lower} + \left(T^{(\ell-1)}_{n,lower} - T^{(\ell)}_{n-1,lower}\right) + t^{n,n+1} \\ &= t^{s,1} + t^{1,2} + \ldots + t^{n,n+1} \\ &\qquad + (\ell-1)\max\{t^{s,1},t^{1,2},\ldots,t^{n,n+1}\} \end{aligned}$$

since $t^{n,n+1} = \max\{t^{s,1},t^{1,2},\ldots,t^{n,n+1}\}$ by (29). Therefore, we have proved (25) implies (26).

Since we showed (25) was true for the base case $N=1$ relays in (24), it follows the channel uses required for ℓ packets to reach the destination through N relays is given by

$$\begin{aligned} T^{(\ell)}_{N,lower} =& t^{s,1} + t^{1,2} + \ldots + t^{N,d} \\ &+ (\ell-1)\max\{t^{s,1},t^{1,2},\ldots,t^{N,d}\} \end{aligned}$$

which is the desired result. ■

References

[1] J. R. Malayter and D. J. Love, "Precoding for low-latency full-duplex MIMO relays: A dynamic approach," in *2022 IEEE Wireless Commun. Netw. Conf.*, March 2022, pp. 2417–2422.

[2] G. Durisi, T. Koch, and P. Popovski, "Toward massive, ultrareliable, and low-latency wireless communication with short packets," *Proc. IEEE*, vol. 104, no. 9, pp. 1711–1726, Sept. 2016.

[3] 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.

[4] C. Bockelmann, N. Pratas, H. Nikopour, K. Au, T. Svensson, C. Stefanovic, P. Popovski, and A. Dekorsy, "Massive machine-type communications in 5G: physical and MAC-layer solutions," *IEEE Commun. Magaz.*, vol. 54, no. 9, pp. 59–65, Sept. 2016.

[5] P. Popovski, Č. Stefanović, J. J. Nielsen, E. De Carvalho, M. Angjelichinoski, K. F. Trillingsgaard, and A.-S. Bana, "Wireless access in ultra-reliable low-latency communication (URLLC)," *IEEE Trans. Commun.*, vol. 67, no. 8, pp. 5783–5801, Aug. 2019.

[6] Y. Zhang, D. J. Love, J. V. Krogmeier, C. R. Anderson, R. W. Heath, and D. R. Buckmaster, "Challenges and opportunities of future rural wireless communications," *IEEE Commun. Mag.*, vol. 59, no. 12, pp. 16–22, Dec. 2021.

[7] C.-C. Wang, D. J. Love, and D. Ogbe, "Transcoding: A new strategy for relay channels," in *2017 55th Annu. Allerton Conf. Commun. Control Comput.* IEEE, Oct. 2017, pp. 450–454.

[8] C. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, July 1948.

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

[10] Y. Polyanskiy and S. Verdú, "Scalar coherent fading channel: Dispersion analysis," in *2011 IEEE Int. Symp. Inf. Theory - Proc.* IEEE, Aug. 2011, pp. 2959–2963.

[11] S. Vituri and M. Feder, "Dispersion of infinite constellations in MIMO fading channels," in *2012 Proc. Conv. Electr. Electron. Eng. Israel*, Nov. 2012, pp. 1–5.

[12] A. Collins and Y. Polyanskiy, "Orthogonal designs optimize achievable dispersion for coherent MISO channels," in *2014 IEEE Int. Symp. Inf. Theory.* IEEE, July 2014, pp. 2524–2528.

[13] J. Hoydis, R. Couillet, and P. Piantanida, "The second-order coding rate of the MIMO quasi-static rayleigh fading channel," *IEEE Trans. Inf. Theory*, vol. 61, no. 12, pp. 6591–6622, Dec. 2015.

[14] A. Sabharwal, P. Schniter, D. Guo, D. W. Bliss, S. Rangarajan, and R. Wichman, "In-band full-duplex wireless: Challenges and opportunities," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 9, pp. 1637–1652, Sept. 2014.

[15] B. P. Day, A. R. Margetts, D. W. Bliss, and P. Schniter, "Full-duplex MIMO relaying: Achievable rates under limited dynamic range," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 8, pp. 1541–1553, July 2012.

[16] B. P. Day, A. R. Margetts, D. W. Bliss, and P. Schniter, "Full-duplex bidirectional MIMO: Achievable rates under limited dynamic range," *IEEE Trans. Signal Process.*, vol. 60, no. 7, pp. 3702–3713, Apr. 2012.

[17] D. Korpi, T. Riihonen, K. Haneda, K. Yamamoto, and M. Valkama, "Achievable transmission rates and self-interference channel estimation in hybrid full-duplex/half-duplex MIMO relaying," in *2015 IEEE Veh. Technol. Conf.*, Sept. 2015, pp. 1–5.

[18] Y. Gu, H. Chen, Y. Li, and B. Vucetic, "Ultra-reliable short-packet communications: Half-duplex or full-duplex relaying?" *IEEE Wireless Commun. Lett.*, vol. 7, no. 3, pp. 348–351, June 2018.

[19] Y. Hu, E. Jorswieck, and A. Schmeink, "Full-duplex relay in high-reliability low-latency networks operating with finite blocklength codes," in *2019 Int. Symp. Wirel. Commun. Syst.*, Aug. 2019, pp. 367–372.

[20] R. L. Rardin, *Optimization in operations research.* Prentice Hall Upper Saddle River, NJ, 1998, vol. 166.

[21] R. S. Blum, "MIMO capacity with interference," *IEEE J. Sel. Areas Commun.*, vol. 21, no. 5, pp. 793–801, June 2003.

[22] E. Telatar, "Capacity of multi-antenna gaussian channels," *Eur. Trans. Telecommun.*, vol. 10, no. 6, pp. 585–595, Nov. 1999.

[23] B. Wang, J. Zhang, and A. Host-Madsen, "On the capacity of MIMO relay channels," *IEEE Trans. Inf. Theory*, vol. 51, no. 1, pp. 29–43, Jan. 2005.

[24] K.-K. K. Kim, "Optimization and convexity of log det$(I + KX^{-1})$," *Int. J. Control Autom. Syst.*, vol. 17, no. 4, pp. 1067–1070, Feb. 2019.

[25] D. Love and R. Heath, "Limited feedback unitary precoding for spatial multiplexing systems," *IEEE Trans. Inf. Theory*, vol. 51, no. 8, pp. 2967–2976, Aug. 2005.

[26] D. J. Love, R. W. Heath, V. K. N. Lau, D. Gesbert, B. D. Rao, and M. Andrews, "An overview of limited feedback in wireless communication systems," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 8, pp. 1341–1365, Oct. 2008.

[27] V. K. Lau, Y. Liu, and T.-A. Chen, "Capacity of memoryless channels and block-fading channels with designable cardinality-constrained channel state feedback," *IEEE Trans. Inf. Theory*, vol. 50, no. 9, pp. 2038–2049, Sept. 2004.

[28] V. Lau, Y. Liu, and T.-A. Chen, "On the design of MIMO block-fading channels with feedback-link capacity constraint," *IEEE Trans. Commun.*, vol. 52, no. 1, pp. 62–70, Jan. 2004.

[29] S. Huberman and T. Le-Ngoc, "Self-interference-threshold-based MIMO full-duplex precoding," *IEEE Trans. Veh.*, vol. 64, no. 8, pp. 3803–3807, Sept. 2015.

[30] *CVX: MATLAB*, (2020). CVX Research. [Online]. Available: http://cvxr.com/cvx/.

[31] M. Grant and S. Boyd, "Graph implementations for nonsmooth convex programs," in *Recent Advances in Learning and Control*, V. Blondel,

S. Boyd, and H. Kimura, Eds. Springer-Verlag Limited, 2008, pp. 95–110.

[32] *MOSEK: MATLAB*, (2019). MOSEK ApS. [Online]. Available: http://cvxr.com/cvx/doc/mosek.html.

[33] R. Bellman, "The theory of dynamic programming," *Bull. Amer. Math. Soc.*, vol. 60, no. 6, pp. 503–515, Nov. 1954.

[34] Y. Polyanskiy, "Channel coding: Non-asymptotic fundamental limits," Ph.D. dissertation, Dept. Electr. Eng., Princeton Univ., Princeton, NJ, USA, Nov. 2010.

[35] M. Polese, M. Giordani, T. Zugno, A. Roy, S. Goyal, D. Castor, and M. Zorzi, "Integrated access and backhaul in 5g mmwave networks: Potential and challenges," *IEEE Commun. Mag.*, vol. 58, no. 3, pp. 62–68, March 2020.

[36] M. Khaturia, K. Appaiah, and A. Karandikar, "On efficient wireless backhaul planning for the "frugal 5g" network," in *2019 IEEE Wireless Commun. and Netw. Conf. Workshop*, 2019, pp. 1–6.

[37] M. Khaturia, P. Jha, and A. Karandikar, "Connecting the unconnected: Toward frugal 5g network architecture and standardization," *IEEE Commun. Stand. Mag.*, vol. 4, no. 2, pp. 64–71, June 2020.

Jacqueline Malayter Jacqueline Malayter received her B.S. (with distinction) in electrical engineering in December of 2020 and began pursuing her Ph.D. degree in electrical engineering in January of 2021, both from Purdue University, West Lafayette. She was awarded the National Science Foundation Graduate Research Fellowship (NSF GRFP) in the Spring of 2022. Her research interests include wireless communications and signal processing, multiple-input multiple-output (MIMO), software defined radios (SDRs), and wireless relay networks.

David Love David J. Love (S'98 - M'05 - SM'09 - F'15) received the B.S. (with highest honors), M.S.E., and Ph.D. degrees in electrical engineering from the University of Texas at Austin in 2000, 2002, and 2004, respectively. Since 2004, he has been with the Elmore Family School of Electrical and Computer Engineering at Purdue University, where he is now the Nick Trbovich Professor of Electrical and Computer Engineering. He served as a Senior Editor for IEEE Signal Processing Magazine, Editor for the IEEE Transactions on Communications, Associate Editor for the IEEE Transactions on Signal Processing, and guest editor for special issues of the IEEE Journal on Selected Areas in Communications and the EURASIP Journal on Wireless Communications and Networking. He was a member of the Executive Committee for the National Spectrum Consortium. He holds 32 issued U.S. patents. His research interests are in the design and analysis of broadband wireless communication systems, beyond-5G wireless systems, multiple-input multiple-output (MIMO) communications, millimeter wave wireless, software defined radios and wireless networks, coding theory, and MIMO array processing.

Dr. Love is a Fellow of the American Association for the Advancement of Science (AAAS) and was named a Thomson Reuters Highly Cited Researcher (2014 and 2015). Along with his co-authors, he won best paper awards from the IEEE Communications Society (2016 Stephen O. Rice Prize and 2020 Fred W. Ellersick Prize), the IEEE Signal Processing Society (2015 IEEE Signal Processing Society Best Paper Award), and the IEEE Vehicular Technology Society (2010 Jack Neubauer Memorial Award).