Training and Decodings for Cooperative Network with Multiple Relays and Receive Antennas
Sun Sun, and Yindi Jing, Member, IEEE

Abstract—In this paper, channel training and coherent decodings under channel estimation error are investigated for relay networks with one single-antenna transmitter, \( R \) single-antenna relays, and one \( R \)-antenna receiver. A two-stage training scheme is proposed to estimate both the relay-receiver and the transmitter-relay channels at the receiver, which are commonly required in amplify-and-forward (AF) relay networks. We use distributed space-time coding (DSTC) for data transmission and investigate the effect of channel estimation errors on network performance. Two coherent decodings are considered: mismatched decoding in which channel estimations are treated as if perfect, and matched decoding in which estimation error is taken into consideration. We show that for full diversity, with mismatched decoding, at least \( 3R \) symbol intervals are required for training; while with matched decoding, \( R+2 \) symbol intervals for training is enough. The complexities of the decoding schemes are investigated. To achieve a balance between performance and complexity, an adaptive decoding scheme is proposed. Simulated error rates are shown to justify the analytical results.

Index Terms—MIMO relay network, channel training, LMMSE estimation, diversity, distributed space-time coding.

I. INTRODUCTION

Cooperative relay network is of great interest nowadays since it can provide possible spatial diversity through the cooperation of the relays. In early researches, coherent cooperative schemes have been proposed (e.g., amplify-and-forward (AF), decode-and-forward (DF), distributed space-time coding (DSTC), beamforming), in which perfect and global channel state information (CSI) is assumed at some or all of the network nodes (e.g. [1]–[4]). In reality, a training process is usually needed for the receiver to estimate the required CSI, and channel estimation is imperfect [5].

In recent years, researches on channel training and communication with imperfect CSI in cooperative relay network are growing (e.g. [6]–[16]). These problems are particularly interesting for AF relay network [7]–[15]. In such network, the end-to-end channel is a concatenation of channels of different transmission phases, making these problems distinct from those in a point-to-point multi-input-multi-output (MIMO) system. For single-relay single-antenna AF network, these problems are investigated in [7]–[11]. Using DSTC, [7] and [8] consider training schemes for the receiver to estimate the end-to-end channel matrix and analyze the network diversity under estimation error. [9] proposes two training schemes for orthogonal AF: One is on the estimation of the end-to-end channel vector at the destination; while for the other, although the relay-receiver channel is estimated at the receiver, the transmitter-relay channel is estimated at the relay then forwarded to the receiver. [10] investigates the SNR-maximizing power allocations between training and data transmission and between the broadcasting phase and the relaying phase in training. In [11], the joint optimization of the number of training symbols, the power allocation between training and data transmission, and the power allocation between the transmitter and the relay that maximizes a mutual information lower bound is investigated. In [12]–[14], training for multi-relay single-antenna network is studied. For AF with matched filter, [12] proposes a training scheme for the receiver to estimate the end-to-end channel coefficient. In [13], [14], the training of the end-to-end channel vector at the receiver is investigated for DSTC.

None of the aforementioned papers consider network with multiple receive antennas, nor do they consider matched decoding, in which the estimation error is taken into consideration in the decoding at the receiver. In [16], [17], we study the training design for the general multi-relay multi-antenna networks. Using DSTC, training and channel estimation schemes are proposed for the receiver to estimate both the transmitter-relay channel matrix and the relay-receiver channel matrix. The achievable diversity under channel estimation error and the training length requirement for full diversity are investigated. In these papers, however, only mismatched decoding is considered. For multiple-antenna system, the maximum-likelihood (ML) decoding under channel estimation error is studied in [18] and compared with mismatched ML decoding. It is proved that both decodings can ensure full diversity provided that the channel estimation error is small.

This paper investigates the training and decodings of a relay network with a single-antenna transmitter, \( R \) single-antenna relays, and a \( R \)-antenna receiver. This scenario mimics the uplink communication from a single-antenna mobile to a \( R \)-antenna base station with the help of \( R \) single-antenna mobiles or relay stations. We propose a training scheme for the receiver to estimate both the relay-receiver and the transmitter-relay channels. This global CSI requirement at the receiver, i.e., the receiver knows not only the relay-receiver channels but also the transmitter-relay channels, is common for non-regenerative cooperative schemes such as AF [1], DSTC [19], and beamforming [21]. It is noteworthy that the proposed scheme requires neither feedback nor communication of channel coefficients among network nodes. We then investigate different coherent decodings for DSTC under imperfect channel estimations. The main contributions are summarized below:

- A two-stage training scheme is proposed to estimate both the transmitter-relay and the relay-receiver channels at the receiver, where the estimation of the former is based on the estimation of the latter. Compared to [16], here, in
estimating the transmitter-relay channels, the estimation error on the relay-receiver channels is considered.

- For mismatched ML decoding, we show analytically that with the shortest training length, \( R + 2 \) symbol intervals, the achievable diversity is at most 1. Simulation illuminates that the minimum requirement on the training time for full diversity is 3\( R \) symbol intervals.
- A matched decoding that takes into account the channel estimation error is proposed. Simulation shows that, with this decoding, the minimum training period of \( R + 2 \) symbol intervals is enough for full diversity. Thus, matched decoding shortens the requirement on training time without sacrificing diversity.
- An adaptive decoding is proposed to balance the network performance and decoding complexity.

The following notation are used in this paper. The conjugate, transpose, Hermitian, Frobenius norm, determinant, inverse, and trace of matrix \( A \) are denoted by \( \overline{A}, A^t, A^H, \| A \|_F, \det(A), A^{-1}, \) and \( \text{tr}(A) \) respectively; \( A \) denotes the column vector by stacking the columns of \( A \) into one column; \( A_{ij} \) denotes the element of \( A \) in the \( i \)-th row and \( j \)-th column; \( A \succ B \) means that \( A - B \) is a positive definite matrix; \( \mathbb{E}(A) \) represents the expectation of \( A \), with each entry the expectation of the corresponding entry of \( A; \mathbb{I}_R \) is the \( R \times R \) identity matrix; \( \otimes \) denotes the Kronecker product; \( \text{diag}(a_1, \cdots, a_n) \) denotes a block diagonal matrix whose \( i \)-th diagonal block is \( a_i; \Re(\cdot) \) denotes the real part of a complex number; \( g(x) = \Re(f(x)) \) indicates \( \lim_{x \to \infty} g(x) / f(x) = c \) with \( c \) a non-zero constant; \( |x| \) denotes the absolute value of a complex scalar \( x \).

The rest of the paper is organized as follows. The network model is described in the following subsection. In Section II, we propose a training scheme and the channel estimations at the receiver. The behaviour of the estimation error is also analyzed. Section III devotes to decodings for DSTC network with estimated CSI, including mismatched, matched, and adaptive decodings. The achievable diversity and training time requirement for full diversity of these decodings are also investigated. Simulation results are shown in Section IV and we conclude in Section V.

A. Network Model

In this subsection, we describe the network model. Consider a relay network with a single-antenna transmitter, \( R \) single-antenna relays, and a \( R \)-antenna receiver as shown in Fig. 1.

![Fig. 1. Network model.](image)

We assume that \( R \geq 2 \). Denote the \( R \times 1 \) transmitter-relay channel vector as \( f = [f_1 \cdots f_R]^t \), where \( f_i \) is the channel from the transmitter to the \( i \)-th relay; denote the \( R \times R \) relay-receiver channel matrix as \( G = [g_{i \cdot}] \), where \( g_i = [g_{i1} \cdots g_{iR}] \) is the \( 1 \times R \) channel vector from the \( i \)-th relay to the receiver. All channel coefficients, \( f_i \) and \( g_{ij} \) for \( i, j = 1, \cdots, R \), are assumed to be i.i.d. quasi-stationary Rayleigh flat-fading with the distribution \( \mathcal{CN}(0, 1) \). We use a block-fading model by assuming that the channels remain invariant during a coherence interval. Each coherence interval is split into the training phase during which channels are estimated and the data transmission phase during which data are sent.

II. Training Scheme

In this section, a training scheme is proposed for the receiver to estimate the relay-receiver channel matrix \( G \) and the transmitter-relay channel vector \( f \). Then properties of the training scheme are discussed.

A. Training of \( G \)

The training of \( G \) is straightforward as the relay-receiver link is a virtual multiple-antenna system, whose training has been well-investigated [22]. Using results in [22], the pilot \( \sqrt{RT_1}{\bar{G}} \) is sent from the relays, where \( P_1 \) is the average transmit power at each relay for each transmission for both the training and data transmission phases\(^2\). Denote the received matrix at the receiver as \( Y_p \). We have \( \hat{Y}_p = \sqrt{RT_1}{\bar{G}} + \tilde{W}_g \), where \( \tilde{W}_g \) is the \( R \times R \) noise matrix at the receiver. In this paper, all the noises are assumed as i.i.d. Gaussian following \( \mathcal{CN}(0, 1) \). The linear minimum-mean-square-error (LMMSE) estimation of \( \bar{G} \) at the receiver is thus:

\[
\hat{G} = \frac{\sqrt{RT_1}}{1 + P_1} \hat{Y}_p. \tag{1}
\]

Since \( \hat{G} \) and \( \hat{Y}_p \) are jointly Gaussian, (1) is also the MMSE estimation.

Due to the Gaussian model, it follows that \( \hat{G} \sim \mathcal{CN} \left( 0, \frac{\sqrt{RT_1}}{1 + P_1} \mathbb{I}_R \right) \). Let \( \Delta G \triangleq G - \hat{G} \), which is the estimation error on \( G \). Invoking the geometric property of the LMMSE estimator that \( \Delta G \) and \( \hat{G} \) are uncorrelated, we have that \( \Delta G \sim \mathcal{CN} \left( 0, \frac{1}{1 + P_1} \mathbb{I}_R \right) \). This training stage takes \( R \) symbol intervals.

B. Training of \( f \) with Estimated \( G \)

To estimate \( f \) at the receiver, we employ the two-step DSTC scheme [16], [19]. Let \( s_p \) be the \( N_p \times 1 \) pilot vector satisfying \( s_p^H s_p = 1 \). This training stage takes \( 2N_p \) symbol intervals, with \( N_p \) symbol intervals for each step. In the first step, \( \sqrt{P N_p s_p} \) is sent by the transmitter, where \( P \) is the average power of the transmitter for both the training and the data transmission phases. Denote the signal received at the \( i \)-th relay by \( r_{i,p} \). We have \( r_{i,p} = \sqrt{P N_p s_p f_i} + v_{i,p} \), where \( v_{i,p} \) is the noise at the

\(^2\)In this paper, to simplify the presentation, we assume that the average power of each node in the network keeps the same for both training and data transmission phases. Optimal power allocation between the training and data transmission phases is important but beyond the scope of this paper.
where the observation model in (4) is calculated to be

\[ X_p = \beta_p S_p H + W_p, \]

where \( \beta_p \triangleq \sqrt{\frac{P + N_p}{P + R}} \), and \( S_p \triangleq [A_{1,p} s_p \ldots A_{R,p} s_p] \), which is the distributed space-time codeword formed at the receiver, \( W_p \triangleq \alpha_p [A_{1,p} v_{1,p} \cdots \cdot A_{R,p} v_{R,p}] G + W_{r,p} \), which is the equivalent noise matrix with \( W_{r,p} \) the noise matrix at the receiver, and

\[ H \triangleq [(f_{1} g_{1})^t \cdots (f_{R} g_{R})^t]^t, \]

which is the \( R \times R \) end-to-end channel matrix.

Stacking the columns of \( X_p \) into one column, we have

\[ \overrightarrow{X_p} = \beta_p [(G^t \otimes I_{N}) S_p] f + \overrightarrow{W_p}, \]

where \( S_p \triangleq \text{diag}\{A_{1,p}s_p, \ldots, A_{2,p}s_p\} \). When \( G \) is perfectly known, \( (G^t \otimes I_{N}) S_p \) can function as the pilot signal in estimating \( f \). In [16], assuming that \( G \) is known perfectly, we propose training and estimation schemes for \( f \). In this paper, in the estimation of \( f \), we assume that only the estimation of \( f \) is known and take the estimation error \( \Delta G \) into consideration. The following theorem is proved.

**Theorem 1:** Given \( G \) in (1), the LMMSE estimation of \( f \) with the observation model in (4) is

\[ \hat{f} = \beta_p \left( R_{\overrightarrow{W}_{e,p}}^{-1} \right)^{-1} \overrightarrow{Z_p} R_{\overrightarrow{w}_{e,p}}^{-1} \overrightarrow{X_p}, \]

where \( \overrightarrow{Z_p} \triangleq (G^t \otimes I_{N}) S_p \), and

\[ R_{\overrightarrow{W}_{e,p}} \triangleq \frac{\beta_p^2}{1 + P} \left( I_{R} \otimes S_p S_p \right) + \frac{R \alpha_p^2}{1 + P} I_{R N_p} + \left( I_{R} \otimes S_p S_p \right) \otimes \mathcal{I}_{N_p}, \]

(6)

Let \( \Delta f \triangleq f - \hat{f} \), which is the estimation error on \( f \). \( \Delta f \) has zero mean i.e., \( \mathbb{E}(\Delta f) = 0 \). Its covariance matrix is

\[ R_{\Delta f} \triangleq \left( R_{\overrightarrow{W}_{e,p}}^{-1} \right)^{-1}. \]

(7)

**Proof:** Replacing \( G \) with \( \hat{G} + \Delta G \) in (4), we get

\[ \overrightarrow{X_p} = \beta_p \overrightarrow{Z_p} f + \beta_p [(G^t \otimes I_{N}) S_p] f + \overrightarrow{W}_{g,1,p} + \overrightarrow{W}_{g,2,p} + \overrightarrow{W}_{r,p}, \]

(8)

where \( \overrightarrow{W}_{g,1,p} \triangleq \alpha_p [A_{1,p} v_{1,p} \cdots \cdot A_{R,p} v_{R,p}] \hat{G} \) and \( \overrightarrow{W}_{g,2,p} \triangleq \alpha_p [A_{1,p} v_{1,p} \cdots \cdot A_{R,p} v_{R,p}] \Delta G \). Define

\[ \overrightarrow{W}_{e,p} \triangleq \beta_p [(G^t \otimes I_{N}) S_p] f + \overrightarrow{W}_{g,1,p} + \overrightarrow{W}_{g,2,p} + \overrightarrow{W}_{r,p}, \]

which is the equivalent noise in the training equation. It has zero-mean. Its covariance matrix can be calculated as follows.

\[ R_{\overrightarrow{W}_{e,p}} = \mathbb{E}(\overrightarrow{W}_{e,p} \overrightarrow{W}_{e,p}^*), \]

\[ = \mathbb{E}(\beta_p^2 [(G^t \otimes I_{N}) S_p]^* [(G^t \otimes I_{N}) S_p]^* + \overrightarrow{W}_{g,1,p} \overrightarrow{W}_{g,1,p}^* + \overrightarrow{W}_{g,2,p} \overrightarrow{W}_{g,2,p}^* + \overrightarrow{W}_{r,p} \overrightarrow{W}_{r,p}^*), \]

(9)

which is (6). The second equality is derived since \( [(G^t \otimes I_{N}) S_p]^* \). In this paper, since only \( G \) is available at the receiver, the expression of \( R_{\Delta f} \) is even more involved, which largely complicates the

### C. Discussions

In Sections II-A and II-B, the proposed training scheme and estimations at the receiver of the network channels, including of the transmitter-relay channel vector \( f \) and the relay-receiver channel matrix \( G \), were discussed. In this subsection we discuss properties of the proposed training and estimation schemes including the training time requirement, the training code design, and the behavior of the estimation error.

1) **Discussion on Training Time:** First, we discuss the length of the training period. The total length of the training, including both the training of \( G \) and the training of \( f \), is \( R + 2 N_p \). From the estimation \( \hat{f} \) in (5), there is no requirement on \( N_p \) to conduct the LMMSE estimation. For good performance, however, the number of independent training equations in (8) should be no less than that of the unknowns. An interesting thing is that the equivalent pilot in the training equation, \( \overrightarrow{Z_p} = (G^t \otimes I_{N}) S_p \), is a random matrix whose property depends on the realization of \( G \). We consider the best scenario of \( G \) being full rank to obtain a lower bound on \( N_p \). When \( G \) is full rank, with a properly designed \( S_p \), the number of independent equations in (8) is \( R N_p \). Since the number of unknowns in \( f \) is \( R \), a lower bound on \( N_p \) is 1, i.e., \( N_p \geq 1 \), which is innocuous. A lower bound on the total training length is thus \( R + 2 \).

2) **Discussion on Pilot Design:** The next to discuss is the pilot design problem, which is to find \( S_p \) and \( A_{i,p} \) such that the power of the estimation error on \( f \) is minimized, i.e.,

\[ \min_{A_{i,p}, S_p} \text{tr}(R_{\Delta f}) \quad \text{s.t.} \quad S_p S_p = 1 \quad \text{and} \quad A_{i,p} A_{i,p} = I_{N_p}, \]

(9)

where \( R_{\Delta f} \) is given in (7). The problem (9) is analyzed in [17] for the case where \( G \) is perfectly known at the receiver. In this paper, since only \( G \) is available at the receiver, the expression of \( R_{\Delta f} \) is even more involved;
optimization problem in (9). For simplicity, we adopt the pilot designs in [17], which minimize the mean square error (MSE) or an upper bound on the MSE of \( f \) when the estimation error of \( G \) is zero. Thus, when the estimation of \( G \) has good quality, for example, when the average power \( P \) is large, these pilot designs are expected to perform close to the optimal.

A case of special interest in this paper is when \( N_p = 1 \). This results in \( R + 2 \) symbol intervals in training, which is the shortest training period. In this case, \( s_p \) and \( A_{i,p} \) reduce to scalars. An obvious code design is \( s_p = A_{i,p} = 0 \) to satisfy the power constraints.

3) Discussion on Estimation Quality of \( f \): In the following, we investigate the estimation error on \( f \). Define the MSE of the estimation \( \hat{f} \) as \( MSE(\hat{f}) \triangleq \mathbb{E}_G(\text{tr}(R_{\Delta f})) \), which is the average power of the estimation error. Note that this definition is slightly different from the conventional one due to the extra average over \( G \). This is because in the training of \( f \), the equivalent pilot is a function of the random matrix \( G \), while conventionally the pilot is a fixed scalar or matrix. We analyze the behaviour of \( MSE(\hat{f}) \) under the aforementioned training and pilot designs. The results are essential to the diversity derivation in the next section.

**Theorem 2:** With the estimation of \( f \) in (5) and the aforementioned pilot designs, when \( N_p = 1 \), \( MSE(\hat{f}) = 2R^2 \log P/P + O(1/P) \), when \( N_p \geq R \), \( MSE(\hat{f}) = O(1/P) \).

**Proof:** See Appendix A.

The different scalings of \( MSE(\hat{f}) \) for different values of \( N_p \) is due to the randomness of the equivalent pilot in (8) and can be intuitively explained as follows. When \( N_p = 1 \), with the proposed pilot designs, (8) can be rewritten as \( \hat{X}_p = \beta_p G^{t}f + \hat{W}_{e,p} \). When \( G \) is full rank, there are \( R \) independent equations. When \( G \) is rank deficient, the number of independent equations is less than \( R \), which consequently leads to inaccurate estimations since there are \( R \) unknown parameters. Although \( G \) is full rank with probability 1, it can get infinitely close to singular with non-zero probability, which on average affects the MSE behaviour. When \( N_p \geq R \), the number of independent training equations in (8) is at least \( R \) regardless of the rank of \( G \). Since there are \( R \) unknown elements in \( f \), reliable estimation is expected.

When \( N_p = 1 \), other than the scalings discussed above, we can also see from Theorem 2 that \( MSE(\hat{f}) \) is proportional to \( R^2 \) when \( \log P \gg 1 \). Since there are \( R \) relays in total, the average \( MSE(\hat{f}) \) per relay is linear in \( R \). This implies that, as the network has more relays and receive antennas, the estimation quality gets worse.

4) Discussion on Applications and Extension: In the training scheme we proposed, all channels, including the transmitter-relay channels and the relay-receiver channels, are estimated at the receiver with no channel coefficient forwarding at the relays and no feedback. Moreover, with these estimations, the end-to-end channel matrix \( H \) can be estimated at the receiver as

\[
H = [(f_1 \hat{g}_1)^t \cdots (f_R \hat{g}_R)^t]^t.
\]

Although in this paper we constrain the relay network to have a single-antenna transmitter, \( R \) single-antenna relays, and a \( R \)-antenna receiver, the training scheme and estimation rule (5) can be applied directly when the numbers of relays and receive antennas are not equal. Nevertheless, the analysis of the training properties can be further involved. The proposed training scheme can also be straightforwardly generalized to the relay network with an \( M \)-antenna transmitter, \( R \) single-antenna relays, and an \( N \)-antenna receiver. The estimation of \( G \), which is \( R \times N \) in this general case, is the same as the point-to-point MIMO channel estimation. Given the estimation \( G \), the \( R \times 1 \) channel vector from each transmit antenna to the relays can be estimated in turn using the rule in (5).

The proposed training design can be applied to many non-regenerative cooperative schemes such as AF [1], DSTC [19], and beamforming [21], in which global CSI at the receiver is required. In this paper, we focus on DSTC scheme.

### III. Decoding and Diversity Analysis of DSTC under Estimation Errors

The channel training and estimations were discussed in Section II. In this section, we study the data transmission of the relay network under channel estimation error, focusing on decoding designs and performance analysis. For the data transmission, only DSTC is considered, which is shown to achieve full diversity with global CSI at the receiver only [19].

#### A. Review on DSTC Decoding and Diversity Results with Perfect CSI

Before proceeding to the decoding and performance analysis with estimation errors, we review the decoding and diversity results of DSTC with perfect CSI [19]. Consider the two-step DSTC with \( T \geq R \) symbol intervals for each step. The notation follow those in Section II-B. Denote the information vector as \( s = [s_1 \cdots s_T]^t \), which satisfies \( E(s^*s) = 1 \). Orthogonal distributed space-time codeword is used, i.e., \( S^*S = I_R \) where \( S = [A_1s \cdots A_Rs] \) [20]. Define \( \beta \triangleq \sqrt{\frac{P}{R(P + 1)}} \) and \( \alpha \triangleq \sqrt{\frac{P}{R(P + 1)}} \). Recall that \( P \) is the transmit power of the transmitter, which is also the total transmit power of the relays. Following the derivation of system equation in (4), the signal received at the receiver can be written as

\[
\tilde{X} = \beta((G^t \otimes I_T)S)f + \tilde{W},
\]

where \( \tilde{X} \triangleq \text{diag} \{A_1s, \cdots, A_Rs\} \) and \( \tilde{W} \triangleq \alpha \{A_1v_1 \cdots A_Rv_R\}G + \tilde{W}_R \) with \( v_i \) the noise at the \( i \)-th relay, \( i = 1, \cdots, R \), and \( \tilde{W}_R \) the noise at the receiver. \( \tilde{W} \) is the equivalent noise at the receiver during data transmission. Define \( Z \triangleq (G^t \otimes I_T)S \) and \( R_{\tilde{W}} \triangleq (I_R + \alpha^2 G^tG) \otimes I_T \) which can be shown to be the covariance matrix of \( \tilde{W} \). With perfect CSI, the ML decoding is derived in [19] as:

\[
\text{DEC}_0 \triangleq \arg \min_s \left( \tilde{X} - \beta Zf \right)^* R_{\tilde{W}}^{-1} \left( \tilde{X} - \beta Zf \right).
\]
S. Let $T \triangleq \mathbf{H}^\dagger [\mathbf{A}_1 \cdots \mathbf{A}_R]$, which is $R \times RT$. Decompose $T$ into blocks of size $1 \times T$ and denote the block of the $i$-th row and $j$-th column as $T_{b,ij}$ for $i, j = 1, \cdots, R$. It can be shown with straightforward algebra that the simplified decoding is:

$$
\text{DEC}_{0,\text{simp}} : \arg \max_{s_j} \mathbb{E} \left[ s_j \left( \sum_{i=1}^{R} T_{b,ii} \right) \right], j = 1, \cdots, T. \quad (13)
$$

This decoding can be performed symbol-by-symbol, thus has much lower complexity. Simulation shows that $\text{DEC}_{0,\text{simp}}$ performs almost the same as the optimal decoding $\text{DEC}_0$.

B. Training-Based Decoding

In this section, we investigate the training-based decoding, where the estimations $\mathbf{G}$ and $\mathbf{f}$ are available in data transmission. The mismatched decoding and the matched decoding are considered first, then an adaptive decoding is proposed to balance the performance and complexity of the network. The complexities of the proposed decoding schemes are analyzed.

1) Mismatched Decoding: Given the estimations $\mathbf{G}$ and $\mathbf{f}$ in (1) and (5) respectively at the receiver, the most straightforward decoding is to ignore the estimation error and replace the channels with their estimations in the coherent ML decoding $\text{DEC}_0$, which yields:

$$
\text{DEC}_1 : \arg \min_{s} \left( \hat{\mathbf{X}} - \beta \hat{\mathbf{Z}}\hat{\mathbf{f}}^\dagger \right)^\dagger \mathbf{R}^{-1}_{W} \left( \hat{\mathbf{X}} - \beta \hat{\mathbf{Z}}\hat{\mathbf{f}}^\dagger \right), \quad (14)
$$

where $\hat{\mathbf{Z}} \triangleq \left( \mathbf{G}^\dagger \otimes \mathbf{I}_T \right) \hat{\mathbf{S}}$ and $\mathbf{R}^{-1}_{W} \triangleq (\mathbf{I}_R + \alpha^2 \mathbf{G}^\dagger \mathbf{G}) \otimes \mathbf{I}_T$.

In analyzing the performance of the decoding $\text{DEC}_1$, we replace $\mathbf{G}$ and $\mathbf{f}$ in the transmission equation (11) by $\mathbf{G} + \Delta \mathbf{G}$ and $\mathbf{f}^\dagger + \Delta \mathbf{f}$, respectively, to obtain the following training-based transmission equation:

$$
\hat{\mathbf{X}} = \beta \hat{\mathbf{Z}}\hat{\mathbf{f}}^\dagger + \hat{\mathbf{W}}_e, \quad (15)
$$

where $\hat{\mathbf{W}}_e$ is the noise plus estimation error term defined as

$$
\hat{\mathbf{W}}_e \triangleq \beta \left[ \Delta \mathbf{f}^\dagger + (\Delta \mathbf{G}^\dagger \otimes \mathbf{I}_T) \hat{\mathbf{S}} \mathbf{f}^\dagger + (\Delta \mathbf{G}^\dagger \otimes \mathbf{I}_T) \hat{\mathbf{S}} \hat{\mathbf{f}} \right] + \hat{\mathbf{W}}_{g1} + \hat{\mathbf{W}}_{g2} + \hat{\mathbf{W}}_r,
$$

with $\hat{\mathbf{W}}_{g1} \triangleq \alpha [\mathbf{A}_1 \mathbf{v}_1 \cdots \mathbf{A}_R \mathbf{v}_R] \hat{\mathbf{G}}^\dagger$ and $\hat{\mathbf{W}}_{g2} \triangleq \alpha [\mathbf{A}_1 \mathbf{v}_1 \cdots \mathbf{A}_R \mathbf{v}_R] \Delta \mathbf{G}$. The noise covariance term in $\text{DEC}_1$, $\mathbf{R}^{-1}_{W}$, actually only corresponds to the noise component in $\hat{\mathbf{W}}_e$, while the estimation error component is ignored. The decoding is thus termed as mismatched decoding since it does not match the training-based system equation.

From the definition of $\hat{\mathbf{W}}_e$ we can see that it is not Gaussian, which makes further analysis intractable. In what follows, we study the performance of $\text{DEC}_1$ using the MSE results in Theorem 2. We consider the behaviour of $\hat{\mathbf{W}}_e$ for different values of $N_p$, corresponding to different lengths for training. It can be expected that larger $N_p$ leads to better performance.

The first to consider is the case of $N_p = 1$, corresponding to the shortest training length $R+2$. Theorem 2 shows that when $N_p = 1$, in the high transmit power range, i.e., $P \gg 1$, MSE($\hat{\mathbf{f}}$) scales as $\log P/P$. From Section II-A we know that MSE($\mathbf{G}$) scales as $1/P$. As $\beta^2 = \frac{T}{P}P + O(1)$ and $\alpha^2 = 1/R + O(1/P)$, the highest order of the average power in $\hat{\mathbf{W}}_e$ is $\log P$, which the complexity of the proposed decoding schemes are analyzed.

$$
\text{DEC}_0,\text{simp} : \arg \max_{s_j} \mathbb{E} \left[ s_j \left( \sum_{i=1}^{R} T_{b,ii} \right) \right], j = 1, \cdots, T. \quad (13)
$$

It can be shown straightforwardly that $\hat{\mathbf{W}}'_e$ has zero-mean. Its covariance matrix is calculated to be

$$
\mathbf{R}^{-1}_{\hat{\mathbf{W}}'_e} = \beta^2 \mathbf{Z} \mathbf{R} \Delta \mathbf{f}^\dagger + \mathbf{I}_R. \quad (17)
$$

By further treating $\Delta \mathbf{f}$ as Gaussian for tractable analysis, the following diversity result is proved.

**Theorem 3:** With the approximate system model and $N_p = 1$, the achieved diversity of $\text{DEC}_1$ is no larger than 1.

**Proof:** See Appendix B.

Although the diversity result in Theorem 3 is derived with an approximate system equation, simulation on block error rate (BLER) in Section IV shows its validity. Thus, with the minimum training period $R + 2$, i.e., $N_p = 1$, mismatched decoding loses diversity. This can be explained via the behaviour of the MSE of the channel estimations. Through the derivation of the system model (16), we know when $N_p = 1$, the equivalent noise $\hat{\mathbf{W}}'_e$ in data transmission is dominated by the term related to the estimation error in $\mathbf{f}$, $\Delta \mathbf{f}$, which scales as $\log P$. Notice that the remaining noise power scales as 1 and the signal power scales as $P$. Treating $\mathbf{f}$ as if it perfectly ignores the $\Delta \mathbf{f}$ term, hence induces diversity loss.

Now we consider the case of $N_p \geq R$. The total training length is thus no shorter than $3R$. In this case, from Theorem 2, all estimation error related terms in $\hat{\mathbf{W}}_e$ have the average power in the order of 1 or lower. Thus, ignoring the estimation error in decoding will not degrade the network diversity. Our simulation shows that in this case, full diversity can be achieved with mismatched decoding.

For the case of $1 < N_p < R$, the MSE and diversity analysis is even more challenging due to the concatenation of channel matrices and DSTC matrix. Our simulation on limited network scenarios shows that the mismatched decoding $\text{DEC}_1$ cannot achieve full diversity. Hence, with $\text{DEC}_1$, at least $3R$ symbol intervals are needed for training to ensure full diversity.

Similar to the perfect CSI case, by approximating $\hat{\mathbf{G}}^\dagger \mathbf{G}$ with its expectation $\hat{\mathbf{G}}^\dagger \hat{\mathbf{G}}$, we consider the orthogonal structure of $\hat{\mathbf{S}}$, we obtain a simplified mismatched decoding $\text{DEC}_{1,\text{simp}}$:

$$
\text{DEC}_{1,\text{simp}} : \arg \max_{s_j} \mathbb{E} \left[ s_j \left( \sum_{i=1}^{R} T_{b,ii} \right) \right], j = 1, \cdots, T, \quad (18)
$$

where analogous to the definitions in Section III-A, $\hat{T}_{b,ij}$ is the $ij$-th $1 \times T$ block element of $\hat{T}$ defined as $\hat{T} = \mathbf{H}^\dagger [\mathbf{A}_1 \cdots \mathbf{A}_R]$. The definition of $\hat{\mathbf{H}}$ is in (10). Simulation shows that the difference in BLER performance between $\text{DEC}_1$ and $\text{DEC}_{1,\text{simp}}$ is negligible.

2) Matched Decoding: Having shown mismatched decoding loses diversity when $N_p = 1$ due to the neglect of the estimation error in decoding, our particular interest is: can
full diversity be achieved when \( N_p = 1 \)? For this concern, we consider the approximate training-based system equation (16).

If \( \Delta f \) is Gaussian, the probability density function (PDF) of \( \hat{X} | \hat{G}, \hat{f} \) is

\[
\mathbb{P}(\hat{X} | \hat{G}, \hat{f}) = \frac{1}{\pi^R \text{det}(\hat{R}_W)} e^{-\frac{1}{2} (\hat{X} - \beta \hat{Z} \hat{f})^T \hat{R}^{-1}_W (\hat{X} - \beta \hat{Z} \hat{f})}.
\]

Maximizing this PDF, the following decoding rule is obtained:

\[
\text{DEC}_2 : \arg \min_s \left( \ln \text{det}(\hat{R}_W) + (\hat{X} - \beta \hat{Z} \hat{f})^T \hat{R}^{-1}_W (\hat{X} - \beta \hat{Z} \hat{f}) \right).
\]

Since \( \Delta f \) is non-Gaussian, this decoding is not the optimal ML decoding of the approximate training-based transmission equation, but a suboptimal decoding. In \( \text{DEC}_2 \), the covariance matrix of the estimation error term \( \beta \hat{Z} \Delta f \) is incorporated through \( \hat{R}_W \). We call it matched decoding, because as opposed to \( \text{DEC}_1 \), it takes into account the estimation error and matches the approximate training-based transmission equation. Simulation shows that \( \text{DEC}_2 \) achieves full diversity even when \( N_p = 1 \). Thus, with the proposed matched decoding \( \text{DEC}_2 \), \( R + 2 \) symbol intervals are enough for the training phase to achieve full diversity in data transmission, while with mismatched decoding \( \text{DEC}_1 \), the minimum training length for full diversity is \( 3R \) symbol intervals.

However, this reduction in training time or improvement in diversity comes with a price on computational complexity. It is noteworthy that \( \hat{R}_W \), in (17) depends on the transmitted signal through \( \hat{Z} \). \( \text{DEC}_2 \) thus cannot be reduced to decoupled symbol-by-symbol decoding. The update of \( \hat{R}^{-1}_W \) during decoding further aggravates the computational load.

3) Adaptive Decoding: In a training-based communication system, it is desirable to achieve reliable communication with both short training period and low decoding complexity. From previous studies, we see that \( \text{DEC}_{1, \text{simp}} \) can be performed symbol-wise, but cannot achieve full diversity with the minimum training length, \( R + 2 \) symbol intervals; \( \text{DEC}_2 \) can achieve full diversity with the minimum training length but it requires joint decoding of all information symbols. In this subsection, we propose an adaptive decoding (A-DEC) by switching between \( \text{DEC}_{1, \text{simp}} \) and \( \text{DEC}_2 \) based on the quality of \( \hat{G} \) to achieve a balance between performance and complexity with the shortest training length.

The idea of A-DEC is to use the more complicated but reliable matched decoding \( \text{DEC}_2 \) only when necessary. Since we target at the shortest training, only the \( N_p = 1 \) case is considered in this part. As discussed at the end of Section II-B, when \( \hat{G} \) is close to singular, the equations in the training model (8) are close to dependent, thus large estimation error occurs. Let \( \hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_R \) be the ordered eigenvalues of \( \hat{G} \hat{G}^* \). We thus use \( \hat{\lambda}_R \), the smallest eigenvalue of \( \hat{G} \hat{G}^* \), to indicate the quality of \( \hat{G} \). When \( \hat{\lambda}_R \) is less than a pre-designed threshold \( \epsilon \), we consider \( \hat{G} \) as ill-conditioned (close to singular) and adopt \( \text{DEC}_2 \); otherwise, the low-complexity decoding \( \text{DEC}_{1, \text{simp}} \) is used. A-DEC is also described in Algorithm 1.

With A-DEC, the balance between performance and complexity can be controlled by adjusting the value of \( \epsilon \). When \( \epsilon = 0 \), A-DEC reduces to \( \text{DEC}_{1, \text{simp}} \); when \( \epsilon \) approaches infinity, it becomes \( \text{DEC}_2 \). With a smaller \( \epsilon \), A-DEC has lower complexity but achieves less reliability, and vice versa. Simulation shows that for any positive \( \epsilon \), full diversity can be obtained when \( N_p = 1 \).

Algorithm 1 Adaptive decoding (A-DEC)

1. Choose the positive threshold \( \epsilon \).
2. Let \( \hat{\lambda}_R \) be the smallest eigenvalue of \( \hat{G} \hat{G}^* \). If \( \hat{\lambda}_R < \epsilon \), employ \( \text{DEC}_2 \); otherwise, employ \( \text{DEC}_{1, \text{simp}} \).

\[
\epsilon = 0, \text{A-DEC reduces to} \text{DEC}_{1, \text{simp}}; \text{when } \epsilon \text{ approaches infinity, it becomes } \text{DEC}_2. \text{ With a smaller } \epsilon \text{, A-DEC has lower complexity but achieves less reliability, and vice versa. Simulation shows that for any positive } \epsilon \text{, full diversity can be obtained when } N_p = 1.
\]

4) Complexity Analysis: We now study the complexity of the proposed decoding schemes, including the mismatched decodings (\( \text{DEC}_1 \) and \( \text{DEC}_{1, \text{simp}} \)), the matched decoding (\( \text{DEC}_2 \)), and the adaptive decoding (A-DEC).

To measure complexity, the unit of flop is employed, which is defined as the amount of the calculation associated with one elementary operation (addition or multiplication) [23], [24]. The following are the rules we adopt in calculating the number of flops. To calculate \( \text{AB} + \text{C} \) where \( \text{A}, \text{B}, \text{and C} \) are \( m \times n, n \times p \), and \( m \times p \) matrices respectively, \( 2mnp \) flops are needed. To save flops, the calculation of the inverse of an \( n \times n \) non-singular matrix is performed by solving linear equations, where \( 2n^3/3 \) flops are required. To calculate the determinant of an \( n \times n \) matrix, \( 2n^3/3 \) flops are required.

Let \( C \) be the modulation for the information symbols \( s_i \)'s and \( |C| \) the cardinality of \( C \). Assume that a rate-1 (with respect to each step) orthogonal DSTC is used in data transmission, e.g., Alamouti code. Orthogonal DSTC with other rates can be treated similarly. By the aforementioned rules, the number of flops needed for performing \( \text{DEC}_1 \), \( \text{DEC}_{1, \text{simp}} \), and \( \text{DEC}_2 \) can be calculated to be:

\[
\text{Comp}(\text{DEC}_1) = R^3 \left[ \left( \frac{2}{3} T + 2 \right) T |C|^T + 2 \right] + 2R^2T |C|^T + 2RT |C|^T,
\]

\[
\text{Comp}(\text{DEC}_{1, \text{simp}}) = R^3 \left( 2T^2 + 2T + 1 \right) + RT + T \left( |C| - 1 \right),
\]

\[
\text{Comp}(\text{DEC}_2) = R^3 \left[ \left( \frac{4}{3} T^2 + 4T + 2 \right) T |C|^T + 10 \right] + 2R^2T |C|^T + 2RT |C|^T + |C|^T,
\]

respectively, where \( \text{Comp}(\cdot) \) indicates the complexity of a decoding. For \( \text{DEC}_1 \) and \( \text{DEC}_2 \), the associated complexities are cubic in the network size \( R \) and exponential in \( T \) due to the joint symbol decoding; for \( \text{DEC}_{1, \text{simp}} \), its complexity is quadratic in \( R \) and \( T \) due to symbol-wise decoding.

For A-DEC, the average number of flops can be derived approximately as follows:

\[
\mathbb{E}[\text{Comp}(\text{A-DEC})] = \mathbb{P}(\hat{\lambda}_R < \epsilon) \text{Comp}(\text{DEC}_2) + \mathbb{P}(\hat{\lambda}_R \geq \epsilon) \text{Comp}(\text{DEC}_{1, \text{simp}}).
\]

Since \( \hat{G} \hat{G}^* \) is a Wishart matrix [27], it can be shown that \( \hat{\lambda}_R \) is exponentially distributed with mean \( P/(R(P+1)) \). Thus \( \mathbb{P}(\hat{\lambda}_R < \epsilon) = 1 - e^{-\epsilon R(P+1)/P} \approx Re \) for small \( \epsilon \), which means \( \text{DEC}_2 \) will be adopted with an approximate probability \( Re \).

In Table I and II, we list the required numbers of flops for \( \text{DEC}_1 \), \( \text{DEC}_{1, \text{simp}} \), \( \text{DEC}_2 \), and A-DEC with several common
TABLE I

<table>
<thead>
<tr>
<th></th>
<th>DEC$_1$</th>
<th>DEC$_{1,\text{simp}}$</th>
<th>DEC$_2$</th>
<th>A-DEC $\epsilon = 0.1$</th>
<th>A-DEC $\epsilon = 0.01$</th>
<th>A-DEC $\epsilon = 0.001$</th>
</tr>
</thead>
<tbody>
<tr>
<td>BPSK</td>
<td>539</td>
<td>58</td>
<td>1162</td>
<td>279</td>
<td>81</td>
<td>61</td>
</tr>
<tr>
<td>4QAM</td>
<td>2107</td>
<td>62</td>
<td>4406</td>
<td>931</td>
<td>149</td>
<td>71</td>
</tr>
<tr>
<td>16QAM</td>
<td>33467</td>
<td>86</td>
<td>69285</td>
<td>13926</td>
<td>1470</td>
<td>225</td>
</tr>
</tbody>
</table>

TABLE II

<table>
<thead>
<tr>
<th></th>
<th>DEC$_1$</th>
<th>DEC$_{1,\text{simp}}$</th>
<th>DEC$_2$</th>
<th>A-DEC $\epsilon = 0.1$</th>
<th>A-DEC $\epsilon = 0.01$</th>
<th>A-DEC $\epsilon = 0.001$</th>
</tr>
</thead>
<tbody>
<tr>
<td>BPSK</td>
<td>$3.38 \times 10^4$</td>
<td>385</td>
<td>$6.98 \times 10^4$</td>
<td>$2.12 \times 10^4$</td>
<td>$2.47 \times 10^4$</td>
<td>594</td>
</tr>
<tr>
<td>4QAM</td>
<td>$5.41 \times 10^5$</td>
<td>393</td>
<td>$1.11 \times 10^6$</td>
<td>$3.34 \times 10^5$</td>
<td>$3.38 \times 10^4$</td>
<td>$3.73 \times 10^3$</td>
</tr>
<tr>
<td>16QAM</td>
<td>$1.38 \times 10^8$</td>
<td>441</td>
<td>$2.84 \times 10^8$</td>
<td>$8.54 \times 10^7$</td>
<td>$8.54 \times 10^6$</td>
<td>$8.55 \times 10^5$</td>
</tr>
</tbody>
</table>

constellations when $R = 2$ and $R = 3$ respectively. $T$, the length of each of the two phases in data transmission, is set to be 2 when $R = 2$ and 4 when $R = 3$.

The results on the training time requirement, diversity, and decoding complexity in this section are only for networks employing DSTC for the data transmission. Different results may be obtained by employing other cooperative schemes.

IV. SIMULATION RESULTS

In this section, simulation results are exhibited. BPSK is used for the data transmission for all simulated networks. For two-relay networks, i.e., $R = 2$, Alamouti code is used, while for three-relay networks, i.e., $R = 3$, a generalized rate-1 real orthogonal code is used [28].

In Fig. 2, to verify our results in Theorem 2, MSE(\hat{f}) for $N_p = 1, 2,$ and 3 are shown for the network with $R = 2$. The scaling behaviour of MSE(\hat{f}) shown in Theorem 2 (when $N_p = 1$, MSE(\hat{f}) scales as $\log P/P$; and when $N_p \geq 2$, MSE(\hat{f}) scales as $1/P$) can be confirmed by reading the slopes of the MSE curves in Fig. 2. We can also see that in the high power region, to achieve the same level of MSE(\hat{f}), by increasing $N_p$ from 1 to 2, 10dB power can be saved in training. But by further increasing $N_p$ from 2 to 3, only 0.5dB power can be saved.

In Fig. 3, the performance of mismatched decodings (DEC$_1$(and DEC$_{1,\text{simp}}$), matched decoding (DEC$_2$), and perfect CSI decodings (DEC$_0$ and DEC$_{0,\text{simp}}$) are shown for the network with $R = 2$. First, we observe that the differences between DEC$_1$ and DEC$_{1,\text{simp}}$, and between DEC$_0$ and DEC$_{0,\text{simp}}$ are negligible. But note that, from Table I, the complexity of the simplified decodings in terms of the number of flops is only 1/10 of the original ones. Second, when $N_p = 1$, DEC$_1$ achieves diversity 1 only, which conforms with our analytical result in Theorem 3. When $N_p = 2$, it achieves full diversity 2. Third, for DEC$_2$, full diversity is achieved even for $N_p = 1$, and the performance is about 11dB better than that of DEC$_1$ at the BLER of $10^{-4}$. When $N_p = 2$, DEC$_2$ is only slightly better than DEC$_1$. But the complexity of DEC$_2$ is
From Table I, we can see that this improvement is gained with training length, better at the BLER of $10^{-5}$. The network with $\mathcal{R} = 2$ decreases much faster due to its full diversity. It is about twice of DEC$^{1,\text{simp}}$, which is consistent with Theorem 3. When $N_p = 1$, the diversity of DEC$^{1,\text{simp}}$ is only 1, which is consistent with Theorem 3. When $N_p = 2$, DEC$^{1,\text{simp}}$ achieves diversity 2. When $N_p = 3$, it achieves the full diversity, which is 3. Thus, for network with $R = 3$, at least 9 symbol intervals are required for full diversity if the mismatched decoding is applied. For DEC$^2$, full diversity is achieved when $N_p = 1$. When $N_p$ increases from 1 to 2, the performance of DEC$^2$ is improved by about 0.5dB. Very little improvement can be obtained with further increase in $N_p$. For $N_p = 1$, the advantage of DEC$^2$ over DEC$^{1,\text{simp}}$ is about 22.5dB at the BLER level of $2 \times 10^{-5}$, while the difference reduces to 3dB when $N_p = 2$, and is negligible when $N_p = 3$. From Table II, the complexity of DEC$^2$ is 181 times that of DEC$^{1,\text{simp}}$. The perfect CSI decoding DEC$_0$ is shown as a benchmark. Due to imperfect training, when $N_p = 3$, the degradation of DEC$_{1,\text{simp}}$ is about 2.5dB worse than the matched decoding DEC$_2$, but its complexity is only 1/4 of DEC$_2$. When $\epsilon = 0.1$, A-DEC performs almost the same as DEC$_2$, but its complexity is less than 1/4 of DEC$_2$. Therefore, the proposed A-DEC is an efficient decoding scheme. By choosing a proper $\epsilon$, we can balance reliability and complexity.

In Fig. 5, the performance of the mismatched decoding DEC$^{1,\text{simp}}$, matched decoding DEC$^2$, and perfect CSI decoding DEC$_0$ are shown for the network with $R = 3$. First it can be observed that when $N_p = 1$, the diversity of DEC$^{1,\text{simp}}$ is only 1, which is consistent with Theorem 3. When $N_p = 2$, DEC$^{1,\text{simp}}$ achieves diversity 2. When $N_p = 3$, it achieves the full diversity, which is 3. Thus, for network with $R = 3$, at least 9 symbol intervals are required for full diversity if the mismatched decoding is applied. For DEC$^2$, full diversity is achieved when $N_p = 1$. When $N_p$ increases from 1 to 2, the performance of DEC$^2$ is improved by about 0.5dB. Very little improvement can be obtained with further increase in $N_p$. For $N_p = 1$, the advantage of DEC$^2$ over DEC$^{1,\text{simp}}$ is about 22.5dB at the BLER level of $2 \times 10^{-5}$, while the difference reduces to 3dB when $N_p = 2$, and is negligible when $N_p = 3$. From Table II, the complexity of DEC$^2$ is 181 times that of DEC$^{1,\text{simp}}$. The perfect CSI decoding DEC$_0$ is shown as a benchmark. Due to imperfect training, when $N_p = 3$, the degradation of DEC$_{1,\text{simp}}$ is about 2.5dB worse than the matched decoding DEC$_2$, but its complexity is only 1/4 of DEC$_2$. When $\epsilon = 0.1$, A-DEC performs almost the same as DEC$_2$, but its complexity is less than 1/4 of DEC$_2$. Therefore, the proposed A-DEC is an efficient decoding scheme. By choosing a proper $\epsilon$, we can balance reliability and complexity.

In Fig. 6, the performance of A-DEC is shown for the network with $R = 3$. $N_p$ is set to be 1, i.e., the minimum training length, 5 symbol intervals, is applied. $\epsilon$ is chosen as 0.001, 0.01, and 0.1. It is observed that all three A-DECs achieve full diversity. As $\epsilon$ increases, the network BLER decreases. BLERs of the mismatched and matched decodings are also shown for comparison. A-DEC has about the same performance as the mismatched decoding DEC$_{1,\text{simp}}$ in the low power region. But as the power increases, its BLER decreases much faster due to its full diversity. It is about 3dB better at the BLER of $10^{-4}$ even when $\epsilon$ is as small as 0.001. From Table I, we can see that this improvement is gained with only 3 extra flops in each block decoding. When $\epsilon = 0.01$, A-DEC is about 2.5dB worse than the matched decoding DEC$_2$, but its complexity is only 1/4 of DEC$_2$. When $\epsilon = 0.1$, A-DEC performs almost the same as DEC$_2$, but its complexity is less than 1/4 of DEC$_2$. Therefore, the proposed A-DEC is an efficient decoding scheme. By choosing a proper $\epsilon$, we can balance reliability and complexity.

In Fig. 6, the performance of A-DEC is shown for the network with $R = 3$. $N_p$ is set to be 1, i.e., the minimum training length, 5 symbol intervals, is applied. Similar observations to those in Fig. 4 can be seen. All A-DECs achieve the full diversity, which is 3. In contrast with DEC$_{1,\text{simp}}$ at the BLER of $2 \times 10^{-5}$, A-DEC with $\epsilon = 0.001$ has about 7dB advantage and 1.5 times complexity; A-DEC with $\epsilon = 0.01$ has about 12.5dB advantage and 6.4 times complexity; A-DEC with $\epsilon = 0.1$ has about 18.5dB advantage and 55 times complexity. In contrast with DEC$_2$ at the BLER of $10^{-5}$, A-DEC with $\epsilon = 0.001$ has about 7dB advantage and 1.5 times complexity; A-DEC with $\epsilon = 0.01$ has about 12.5dB advantage and 6.4 times complexity; A-DEC with $\epsilon = 0.1$ has about 18.5dB advantage and 55 times complexity.
DEC with $\epsilon = 0.001$ has about $15.5$dB disadvantage and $1/118$ complexity; A-DEC with $\epsilon = 0.01$ has about $10$dB disadvantage and $1/28$ complexity; A-DEC with $\epsilon = 0.1$ has about $4$dB disadvantage and $1/3$ complexity.

V. CONCLUSIONS

In this paper, we consider the training in a relay network with one single-antenna transmitter, $R$ single-antenna relays, and one $R$-antenna receiver. A two-stage training scheme is proposed to estimate both the transmit-relay-relay receiver channels at the receiver. Two coherent decodings, a mismatched decoding and a matched decoding are considered in DSTD network. We show that at least $3R$ symbol intervals of training are required for mismatched decoding to achieve full diversity, while $R + 2$ symbol intervals for training are enough for the matched decoding to achieve full diversity. But the complexity of the matched decoding is exponential in the data rate. An adaptive decoding is thus proposed to balance the BLER performance and decoding complexity.

APPENDIX

A. Proof of Theorem 2

Recall that from Theorem 1, with the imperfect estimation $\hat{G}$, $R_{\Delta f} = \left( \mathbb{I}_R + \frac{\beta_2^2}{\alpha_1^2} \hat{Z}_p \mathbb{I}_{\hat{W}_{e.p}} \right)^{-1}$, where $\hat{Z}_p \triangleq (G^t \otimes \mathbb{I}_{N_p}) \hat{S}_p$ and $R_{\hat{W}_{e.p}} = \frac{\beta_2^2}{R} (G^t \otimes S_p S_p^t) + \frac{\alpha_2^2}{R} \mathbb{I}_{RN_p} + (\mathbb{I}_R + \alpha_2^2 G^t G) \mathbb{I}_{N_p}$.

We first consider the case of $N_p = 1$. Using the proposed training code, $\hat{Z}_p = G^t$ and $R_{\hat{W}_{e.p}} = kI_R + \alpha_2^2 \hat{G}^t \hat{G}$ where $k \triangleq 1 + \frac{R(\alpha_2^2 + \beta_2^2)}{1 + \beta_2^2} = 2 + O \left( \frac{1}{P} \right)$. Thus,

$$R_{\Delta f} = \left[ I_R + \frac{1}{k} \hat{G} \left( \hat{G}^t + \frac{1}{k} \alpha_2^2 \hat{G}^t \hat{G} \right)^{-1} \hat{G}^t \right]^{-1}. \quad (19)$$

From (19), the eigenvalues of $R_{\Delta f}$ are $\frac{k + \alpha_2^2 \lambda_{un,i}}{k + (\alpha_2^2 + \beta_2^2) \lambda_{un,i}}$ for $i = 1, \ldots, R$ with $\lambda_{un,i}$'s the unordered eigenvalues of $\hat{G} \hat{G}^t$, since $\hat{G} \left( \hat{G}^t + \frac{1}{k} \alpha_2^2 \hat{G}^t \hat{G} \right)^{-1} \hat{G}^t$ and $\left( \hat{G}^t + \frac{1}{k} \alpha_2^2 \hat{G}^t \hat{G} \right)^{-1} \hat{G}^t$ have the same eigenvalues. Thus,

$$\text{MSE}(\hat{f}) = \mathbb{E}_{\lambda_{un,i}} \left[ \sum_{i=0}^{R} \frac{k + \alpha_2^2 \lambda_{un,i}}{k + (\alpha_2^2 + \beta_2^2) \lambda_{un,i}} \right]. \quad (20)$$

Since entries of $\hat{G}$ are i.i.d. complex Gaussian, the unordered eigenvalues of the Wishart matrix $\hat{G} \hat{G}^t$ have the same PDF, which was derived in [29] to be

$$p_{\lambda_{un,i}}(x) = \frac{1}{R} \left( 1 + \frac{2}{P} \right) e^{-x} \left[ \sum_{i=0}^{R-1} L_i(x) \right],$$

where $L_i(x)$ is the Laguerre polynomial of order $i$ defined as $L_i(x) = \sum_{k=0}^{i} (-1)^k \binom{i}{k} \frac{x^k}{k!}$. After some straightforward transformation, when $P \gg 1$, $p_{\lambda_{un,i}}(x)$ can be rewritten as

$$p_{\lambda_{un,i}}(x) = e^{-x} + \frac{1}{R} e^{-x} \sum_{j=1}^{2(R-1)} r_j x^j$$

+ lower order terms in $P$, (21)

where $r_j$'s are constants irrelevant of $P$.

Let $f(x) = \frac{k + \alpha_2^2}{k + (\alpha_2^2 + \beta_2^2)}$. From (20), we have

$$\text{MSE}(\hat{f}) = \mathbb{E}_{\lambda_{un,i}} f(\lambda_{un,i}) \quad (22)$$

First we calculate the second integral in (22). For $j \geq 1$,

$$\int_0^\infty f(x)x^j e^{-x} dx = \frac{\Gamma(1+j) + \alpha_2^2 E_{2+j} \left( \frac{k}{\alpha_2^2 + \beta_2^2} \right) \Gamma(2+j)}{\alpha_2^2 + \beta_2^2} \quad (23)$$

where $E_n(x) \triangleq \int_0^\infty e^{-xt}/t^n dt$ is the exponential integral, and $\Gamma(x) \triangleq \int_0^\infty e^{-t}/t^{x-1} dt$ is the Gamma function. Since $\alpha_2^2 = 1/R + O(1/P)$ and $\beta_2^2 = P/R + O(1)$ (notice that $N_p = 1$), we can show that when $P \gg 1$ and $j \geq 1$, $E_1+j \left( k/(\alpha_2^2 + \beta_2^2) \right) = 1/(1+j) + O(1/P)$. So, when $P \gg 1$,

$$\int_0^\infty f(x)x^j e^{-x} dx = \frac{1}{P^2} + O \left( \frac{1}{P} \right). \quad (24)$$

When $P \gg 1$, $E_1 \left( \frac{k}{\alpha_2^2 + \beta_2^2} \right) = \log P + O(1).$ So,

$$\int_0^\infty f(x)x^j e^{-x} dx = 2R \log^2 P + O \left( \frac{1}{P} \right).$$

Applying the results in (23) and (24) to (21), we have

$$\text{MSE}(\hat{f}) = 2R^2 \log^2 P + O(1/P).$$

Next we consider the case of $N_p \geq R$. We have for $P \gg 1$, $\beta_2^2 = \frac{N_p}{R} P + O(1)$. To prove $\text{MSE}(\hat{f}) = O(1/P)$, we will find a lower bound and an upper bound on $\text{MSE}(\hat{f})$ both of order $1/P$.

We start by bounding $R_{\hat{W}_{e.p}}$. It is obvious that $R_{\hat{W}_{e.p}} \geq \mathbb{I}_{RN_p}$. Applying the lower bound on $R_{\hat{W}_{e.p}}$ in $R_{\Delta f}$ given in (7), $R_{\Delta f}$ can be lower bounded as

$$R_{\Delta f} \geq \mathbb{I}_R + \beta_2^2 \mathbb{I}_R \hat{Z}_p^{-1} \quad (25)$$

where $\hat{Z}_p = (\mathbb{I}_R \otimes S_p) \left( G_1^t, \cdots, G_R^t \right)^t$ with $G_i \triangleq \mathbb{E}_{\lambda_i} \left( \hat{g}_1, \cdots, \hat{g}_{R_i} \right)$, and $S_p S_p^t = \mathbb{I}_R$. The equality is derived by using the facts that $\hat{Z}_p$ can be rewritten as $\hat{Z}_p = (\mathbb{I}_R \otimes S_p) \left( G_1^t, \cdots, G_R^t \right)^t$ with $G_i \triangleq \mathbb{E}_{\lambda_i} \left( \hat{g}_1, \cdots, \hat{g}_{R_i} \right)$, and $S_p S_p^t = \mathbb{I}_R$.
Since \( \text{MSE}(\hat{f}) = \mathbb{E}_G (\text{tr}(R_{\Delta f})) \), using (25), \( \text{MSE}(\hat{f}) \) can be lower bounded as follows.

\[
\text{MSE}(\hat{f}) \geq \mathbb{E}_G \left[ \frac{R}{1 \! \! \! \! \! /} \left( 1 + \beta^2 P \| g_t \|_F^2 \right)^{-1} \right]
\]

\[
= \mathbb{E}_G \left[ \frac{R}{1 \! \! \! \! \! /} \left( 1 + \frac{\beta^2 P}{P + 1} \| g_t \|_F^2 \right)^{-1} \right]
\]

\[
= \frac{R^2}{N_p (R - 1)} + \text{lower order terms in } P, \tag{26}
\]

where (26) is derived since \( x = \| g_t \|_F^2 \) is Gamma distributed with degree \( R (R \geq 2) \). Its PDF is \( p_{\|g\|_F^2} (x) = \frac{x^{R-1} e^{-x}}{\Gamma(R) R^R} \).

So, we have that the lower bound on MSE(\( \hat{f} \)) scales as \( 1/P \).

Now, we upper bound \( R_{\Delta f} \). Since \( S^*_p S_p = I_{N_p} \), we have \( S^*_p S_p \preceq I_{N_p} \) and thus \( I_R \preceq (S^*_p S_p) \preceq I_{RN_p} \). Combined with \( \hat{G}^t \hat{G} \preceq \| \hat{G} \|_F^2 I_R \), we can upper bound as

\[
R_{\Delta f} = \frac{R \alpha^2}{k + \alpha^2 \| \hat{G} \|_F^2} \preceq \frac{R \alpha^2}{k + \alpha^2 \| \hat{G} \|_F^2} \frac{R \alpha^2}{k + \alpha^2 \| \hat{G} \|_F^2} = \frac{N_p + R}{1 \! \! \! \! \! /} + O \left( \frac{1}{P} \right). \tag{27}
\]

Applying the upper bound to \( R_{\Delta f} \) given in (7), we have

\[
R_{\Delta f} \preceq \text{diag} \left\{ \left( 1 + \frac{\beta^2 P}{k_1 + \alpha^2 \| g_t \|_F^2} \right)^{-1}, \cdots, \left( 1 + \frac{\beta^2 P}{k_1 + \alpha^2 \| g_t \|_F^2} \right)^{-1} \right\}. \tag{28}
\]

MSE(\( \hat{f} \)) can thus be upper bounded as

\[
\text{MSE}(\hat{f}) \leq \mathbb{E}_G \left[ \sum_{i=1}^{R} \left( 1 + \frac{\beta^2 P}{(k_1 + \alpha^2 \| g_t \|_F^2) (P + 1)} \| g_t \|_F^2 \right)^{-1} \right]
\]

This has the same form as the pairwise-error-probability (PEP) upper bound formula (30) in [19]. Following the technique in [19], we can show that it scales as \( 1/P \) when \( P \gg 1 \). Thus, we conclude that the upper bound on MSE(\( \hat{f} \)) also scales as \( 1/P \), which completes the proof.

**B. Proof of Theorem 3**

Take the singular value decomposition (SVD) of \( \hat{G}^t \) as \( \hat{G}^t = U \Sigma V^t \) where \( U \) and \( V \) are \( R \times R \) unitary matrices, and \( \Sigma = \text{diag} \{ \lambda_1^{1/2}, \cdots, \lambda_R^{1/2} \} \) with \( \lambda_1 \geq \cdots \geq \lambda_R \). We first prove two lemmas to help the diversity analysis.

**Lemma 1:** When \( P \gg 1 \), given that \( \lambda_R \leq 1/P \), we have \( R_{\Delta f} \succeq C_{\text{R}_f} \) where

\[
C_{\text{R}_f} = V^t \text{diag} \left\{ 0, \cdots, 0, \frac{2P}{2R + 1} \right\} V.
\]

**Proof:** Replace \( \hat{G}^t \) with its SVD in (19),

\[
R_{\Delta f} = V^t \text{diag} \left\{ \frac{k + \alpha^2 \lambda_1}{k + \alpha^2 \lambda_R}, \cdots, \frac{k + \alpha^2 \lambda_R}{k + \alpha^2 \lambda_R} \right\} V.
\]

When \( \lambda_R \leq 1/P \), we have \( \frac{k + \alpha^2 \lambda_R}{k + \alpha^2 \lambda_R} \geq \frac{k + \alpha^2 \lambda_R}{k + \alpha^2 \lambda_R} = \frac{k + \alpha^2 \lambda_R}{k + \alpha^2 \lambda_R} = \frac{k + \alpha^2 \lambda_R}{k + \alpha^2 \lambda_R} \), where the first inequality follows from the fact that \( \frac{k + \alpha^2 \lambda_R}{k + \alpha^2 \lambda_R} \) is a decreasing function. Thus, we have \( R_{\Delta f} \succeq C_{\text{R}_f} \).

**Lemma 2:** The PEP of the network given \( \hat{\lambda}_R \leq 1/P \) is no smaller than a positive constant \( c_{\text{R}_f} \).

**Proof:** Now we calculate the PEP of mistaking the codeword \( S_1 \) with \( S_2 \). The two codewords correspond to the information vectors \( s_1 \) and \( s_2 \) respectively. Given that \( S_1 \) is sent from the approximate training-based system equation (16), we have \( \hat{X}_1 = \beta Z_1 + \hat{W}_1 + \hat{W}_1^T \), where \( Z_1 \equiv (G^t \otimes I_T) S_1 \) and the equivalent noise \( \hat{W}_1 \). Following the two codewords \( (G^t \otimes I_T) S_2 \). Using the mismatched decoding DECI in (14), the conditional PEP can be calculated as follows.

\[
\text{PEP}(S_1 \rightarrow S_2 \mid G, \hat{f}) \leq \frac{1}{P} \left( \frac{\hat{X}_1 - \beta Z_1 \hat{f}}{\hat{W}} \right), \tag{29}
\]

where \( \hat{f} \equiv (Z_1 - \hat{Z}_2) \hat{f} \equiv \Delta \hat{f} \). Let \( \beta \equiv \left( Z_1 - \hat{Z}_2 \right) \hat{f} \equiv \Delta \hat{f} \).

\[
\text{PEP}(S_1 \rightarrow S_2 \mid G, \hat{f}) \leq \frac{1}{P} \left( \frac{\hat{X}_1 - \beta Z_1 \hat{f}}{\hat{W}} \right), \tag{30}
\]

where \( Q(x) \) is the \( Q \)-function defined as \( Q(x) = \frac{1}{\sqrt{2\pi}} \int_{x}^{\infty} e^{-t^2/2} dt \), and the equality comes from the fact that \( R_{\Delta f} \) is decreasing. For the tractability of analysis, we make another approximation \( \hat{G}^t \hat{G}^t \approx \alpha^2 R^2 \| \hat{G} \|_F^2 \) for large \( R \). That is, we replace \( \hat{G}^t \hat{G}^t \) with its average. This approximation is expected to be tight especially for large \( R \), since \( \hat{G}^t \hat{G}^t \rightarrow \alpha^2 R^2 \| \hat{G} \|_F^2 \) as \( R \rightarrow \infty \). With this approximation, we have

\[
\text{PEP}(S_1 \rightarrow S_2 \mid G, \hat{f}) \leq \frac{1}{P} \left( \frac{\hat{X}_1 - \beta Z_1 \hat{f}}{\hat{W}} \right), \tag{31}
\]

where \( Q(x) \) is the \( Q \)-function defined as \( Q(x) = \frac{1}{\sqrt{2\pi}} \int_{x}^{\infty} e^{-t^2/2} dt \), and the equality comes from the fact that \( R_{\Delta f} \) is decreasing. For the tractability of analysis, we make another approximation \( \hat{G}^t \hat{G}^t \approx \alpha^2 R^2 \| \hat{G} \|_F^2 \) for large \( R \). That is, we replace \( \hat{G}^t \hat{G}^t \) with its average. This approximation is expected to be tight especially for large \( R \), since \( \hat{G}^t \hat{G}^t \rightarrow \alpha^2 R^2 \| \hat{G} \|_F^2 \) as \( R \rightarrow \infty \). With this approximation, we have
The argument of the $Q$-function in (32) is irrelative of $P$ and the denominator of the argument is non-zero. Thus, after performing the expectation over $\mathbf{f}$, we have

$$
\text{PEP}(\mathbf{S}_i \rightarrow \mathbf{S}_2(\mathbf{G}, \mathbf{f})) \geq \frac{\|s_1 - s_2\|^2}{\sqrt{\mathbb{E}((s_1 - s_2)^T s_1)}} \cdot \text{tr}(\mathbf{f}^* \mathbf{C}_{R_s})
$$

(32)

With the result in Lemma 2, we can lower bound the overall PEP. When $P \gg 1$, $P(\lambda_R \leq 1/P) = 1 - e^{-R_1(1+P)/P^2} \approx RP^{-1} + O(P^{-2})$. Together with Lemma 2, we have

$$
\text{PEP}(\mathbf{S}_i \rightarrow \mathbf{S}_2(\mathbf{G}, \mathbf{f})) \geq \frac{\|s_1 - s_2\|^2}{\sqrt{\mathbb{E}((s_1 - s_2)^T s_1)}} \cdot \text{tr}(\mathbf{f}^* \mathbf{C}_{R_s})
$$

which means that the diversity order is upper bounded by 1.

REFERENCES


Sun Sun received the B.S. degree in Electrical Engineering and Automation from Tongji University, Shanghai, China, in 2005. From 2006 to 2008, she was a software engineer in the Department of GSM Base Transceiver Station of Huawei Technologies Co. Ltd. She received the M.Sc. degree in Electrical and Computer Engineering from University of Alberta, Edmonton, Canada, in 2011. Currently, she is pursuing her Ph.D. degree in the Department of Electrical and Computer Engineering of University of Toronto, Toronto, Canada. Her research interest lies in wireless communications in relay networks, focusing on limited feedback, beamforming, and network performance analysis.

Yindi Jing received the B.Eng. and M.Eng. degrees in automatic control from the University of Science and Technology of China, Hefei, China, in 1996 and 1999, respectively. She received the M.Sc. degree and the Ph.D. in electrical engineering from California Institute of Technology, Pasadena, CA, in 2000 and 2004, respectively. From October 2004 to August 2005, she served as an Associate Editor for the IEEE Transactions on Wireless Communications. Her research interests are in wireless communications, focusing on cooperative relay networks, training and channel estimation, distributed beamforming, and the analysis on network fundamental limits.