Stochastic Gradient Algorithms for Design of Minimum Error-Rate Linear Dispersion Codes in MIMO Wireless Systems

Xiaodong Wang, *Senior Member, IEEE*, Vikram Krishnamurthy, *Fellow, IEEE*, and Jibing Wang, *Member, IEEE*

***Abstract*—Linear dispersion (LD) codes are a good candidate for high-data-rate multiple-input multple-ouput (MIMO) signaling. Traditionally LD codes were designed by maximizing the average mutual information, which cannot guarantee good error performance. This paper presents a new design scheme for LD codes that directly minimizes the block error rate (BLER) in MIMO channels with arbitrary fading statistics and various detection algorithms. For MIMO systems employing LD codes, the error rate does not admit an explicit form. Therefore, we cannot use deterministic optimization methods to design the minimum-error-rate LD codes. In this paper, we propose a simulation-based optimization methodology for the design of LD codes through stochastic approximation and simulation-based gradient estimation. The gradient estimation is done using the score function method originally developed in the discrete-event-system community. The proposed method can be applied to design the minimum-error-rate LD codes for a variety of detector structures including the maximum-likelihood (ML) detector and several suboptimal detectors. It can also design optimal codes under arbitrary fading channel statistics; in particular, it can take into account the knowledge of spatial fading correlation at the transmitter and receiver ends. Simulation results show that codes generated by the proposed new design paradigm generally outperform the codes designed based on algebraic number theory.**

***Index Terms*—Gradient estimation, linear dispersion codes, multiple-input multiple-output (MIMO), score function, stochastic approximation.**

I. Introduction

MULTIPLE-INPUT multiple-output (MIMO) technology for wireless communications is currently an active research area. Linear dispersion (LD) codes introduced in [1] are a good candidate for high-data-rate MIMO signaling over wireless channels. LD codes use a linear modulation framework and the transmitted codeword is a linear combination over space and time of certain "dispersion matrices" with the transmitted symbols as combining coefficients (see also [2]). The LD codes are simple to encode. Furthermore, LD codes can be decoded very efficiently either by the polynomial-complexity maximum-likelihood (ML) detector, i.e., the sphere decoder [3], or by a suboptimal detector, such as the nulling-and-cancellation detector [5] or the linear detector (see, e.g., [6]). Traditionally LD codes (e.g., [1]) only optimize the average mutual information; and therefore cannot guarantee good error performance [7]. More recently, full-rate full-diversity linear space-time codes have been introduced in the literature (see, e.g., [16]–[19]). In [16], a scheme called threaded algebraic space–time (TAST) coding is proposed. This scheme falls into the general framework of LD codes. TAST codes guarantee full diversity and full rate with arbitrary number of transmit and receive antennas. The design of the TAST focuses on the worst-case pairwise error probability (PEP). The PEP, however, may not be the best performance metric, since it is not true in general that the codes optimized with respect to the worst case PEP will end up with the minimum block error rate (BLER). In [18], linear space–time codes are designed via a deterministic optimization of the Chernoff union bound. However, the Chernoff bound is not an asymptotic tight upper bound of the BLER.

This paper has two main ideas.

1) *Average BLER minimization via a gradient estimation based stochastic approximation algorithm*: We design LD codes with minimum BLER based on stochastic approximation together with gradient estimation. Stochastic approximation (SA) algorithms with gradient estimation have been used extensively for optimizing the performance of discrete-event systems (DES) [8]. Examples of DES include computer-communication networks, traffic systems and flexible manufacturing systems, queueing systems, product lines, etc. For most DES, analytical expressions are not available for the performance metrics—these metrics need to be optimized via *simulation-based optimization* [8], [10], [26]. In this paper, we show how these gradient estimation methods together with stochastic approximation can be successfully used to design LD codes such that the BLER is minimized. For MIMO systems employing LD codes, we do not have an explicit form of BLER. Therefore, we cannot use deterministic optimization method to design the optimal (i.e., minimum error rate) LD codes. The stochastic approximation algorithm with gradient estimation is ideally suited for such situation. In this paper, we employ the score function method to obtain an

Manuscript received May 10, 2004; revised April 27, 2005. The work of X. Wang was supported in part by the U.S. National Science Foundation (NSF) under grants DMS-0225692, and by the U.S. Office of Naval Research (ONR) under grant N00014-03-1-0039. The work of V. Krishnamurthy was supported in part by NSERC and the British Columbia Advanced Systems Institute. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Gregori Vazquez.

X. Wang is with the Department of Electrical Engineering, Columbia University, New York, NY 10027 USA (e-mail: wangx@ee.columbia.edu).

V. Krishnamurthy is with the Department of Electrical and Computer Engineering University of British Columbia, Vancouver, BC V6T 1Z4, Canada (e-mail: vikramk@ece.ubc.ca).

J. Wang is with the Qualcomm, Inc., San Diego, CA 92121-1714 USA (e-mail: jibingw@qualcomm.com).

Digital Object Identifier 10.1109/TSP.2005.863122

unbiased estimate of the gradient of BLER with respect to the dispersion matrices. We then use this gradient estimator to optimize the LD codes via the well-known Robbins–Monro (R–M) stochastic gradient algorithm [25]. Section III-A gives more perspective on gradient estimation in stochastic approximation.

2) *Spherical parameterization of energy constraint*: Our LD code design problem is a stochastic optimization problem with an energy constraint that requires the dispersion matrix coefficients to lie on the surface of a hyper-sphere. We show that by re-expressing the constrained optimization problem in spherical coordinates, it can be converted into an equivalent unconstrained optimization problem. That is, the LD codes obtained at each iteration in terms of spherical coordinates automatically satisfy the energy constraints. Actually this formulation is equivalent to using differential geometry to project the derivative on the tangent space of the sphere [23] and is inspired by our recent work in reinforcement learning [24]. More important, this spherical coordinate formulation allows us to use a straightforward proof from the stochastic approximation literature to show that the algorithm yields strongly consistent estimates. Also, since the spherical coordinate formulation exploits the structure of the energy constraint, it is much simpler than the usual generic approaches for constrained stochastic optimization: e.g., a primal dual stochastic approximation algorithm to deal with the constraint (which requires convexity while our cost function is not convex in general).

Most work on space–time code design assumes ML detection. For very high data rate signaling, even the sphere decoder might be too complicated to implement in practice. It is difficult to design space–time codes where a suboptimal detector e.g., the nulling-and-cancellation detector as the performance analysis seems intractable (see, e.g., [27]). One advantage of the proposed method is that we can optimize the LD codes for both the ML detector as well as suboptimal detectors such as the nulling-and-cancellation detector. On the other hand, in MIMO wireless systems, the individual antennas could be correlated due to insufficient antenna spacing and lack of scattering [14], [15]. Moreover, the fading channel statistics may deviate from the common Rayleigh assumption, due to, e.g., line-of-sight component. We demonstrate how to design optimal LD codes if these long-term statistics can be measured beforehand. Finally, we present simulation results to demonstrate that the LD codes obtained using the proposed design procedure generally outperform the codes designed based on the algebraic number theory, especially when a suboptimal detector is employed or when the MIMO channels are spatially correlated.

For the simulation-based code design, one needs to know the MIMO configurations as well as the fading statistics. We note that other analytical space–time code design methods (such as the mutual information criterion [1] as well as the rank-and-determinant criterion [12]) also require such knowledge. While the rank-and-determinant criterion does not require the number of receive antennas, it turns out that the optimal codes do depend on the number of receive antennas (see [1] and [16]). In our design, we also need to know the operating signal-to-noise ratio (SNR) (see also [1]). As we demonstrate through examples, the codes generated by the proposed method under some fixed SNR value perform well for a range of SNR of interest.

The remainder of this paper is organized as follows. We formulate the LD code design problem as a stochastic optimization problem in Section II. We discuss the design procedure and the proposed stochastic optimization algorithm in Section III. In Section IV, we provide some simulation results. The paper is concluded in Section V. We defer some detailed computations and proofs to the Appendices.

II. Design of Linear Dispersion Codes

In this section, we present the signal model for MIMO systems employing LD codes, and formulate the LD design problem as a constrained stochastic optimization problem.

A. Signal Model

Consider a MIMO system with M_T transmit antennas and M_R receive antennas. Assume that the channel is flat fading and remains constant for τ symbol intervals, and the fading coefficient from the ith transmit antenna to the jth receive antenna is denoted by $h_{i,j}$. The signal transmitted from the ith transmit antenna at time index t is denoted by $x_{t,i}$, while the signal received at the jth receive antenna at time t is denoted by $y_{t,j}$. The input–output relationship is given by

$$y_{t,j} = \sqrt{\frac{\rho}{M_T}} \sum_{i=1}^{M_T} h_{i,j} x_{t,i} + w_{t,j},\ t = 1, \cdots, \tau,\ j = 1, \cdots, M_R \tag{1}$$

where $w_{t,j}$ is independent zero-mean complex Gaussian noise with unit variance. The transmitted energy on all M_T antennas at any given time is normalized to unity; therefore, ρ is the SNR at each receive antenna regardless of the number of transmit antennas. Equation (1) can be written in matrix form as

$$\boldsymbol{Y} = \sqrt{\frac{\rho}{M_T}} \boldsymbol{X}\boldsymbol{H} + \boldsymbol{W} \tag{2}$$

where $\boldsymbol{Y}$ is the $\tau \times M_R$ matrix of the received signal, $\boldsymbol{X}$ is the $\tau \times M_T$ matrix of the transmitted signal, $\boldsymbol{W}$ is the $\tau \times M_R$ matrix of the additive white Gaussian noise, and $\boldsymbol{H}$ is the $M_T \times M_R$ MIMO channel matrix.

Assume we transmit Q r-QAM symbols $\{s_q\}_{q=1}^{Q}$ with unit average energy over τ symbol intervals, the LD transmission matrix $\boldsymbol{X}$ is given by [1]

$$\boldsymbol{X} = \sum_{q=1}^{Q} \alpha_q \boldsymbol{A}_q + j\beta_q \boldsymbol{B}_q \tag{3}$$

where we have decomposed the transmitted symbols s_q into their real and imaginary parts, i.e., $s_q = \alpha_q + j\beta_q, q = 1, \cdots, Q$, and $\{\boldsymbol{A}_q, \boldsymbol{B}_q\}_{q=1}^{Q}$ are complex-valued dispersion matrices of dimension $\tau \times M_T$ that specify the code. The rate of the LD code is

$$R = \frac{Q \log_2 r}{\tau}.$$

We also assume that the dispersion matrices $\{\boldsymbol{A}_q, \boldsymbol{B}_q\}_{q=1}^{Q}$ satisfy the following energy constraint:

$$\sum_{q=1}^{Q} \mathrm{Tr}\left(\boldsymbol{A}_q^H \boldsymbol{A}_q + \boldsymbol{B}_q^H \boldsymbol{B}_q\right) = 2\tau M_T. \tag{4}$$

As in [1], we denote $\boldsymbol{Y}_R = \mathrm{Re}\{\boldsymbol{Y}\}$ and $\boldsymbol{Y}_I = \mathrm{Im}\{\boldsymbol{Y}\}$. Denote the columns of $\boldsymbol{Y}_R$, $\boldsymbol{Y}_I$, $\boldsymbol{H}_R$, $\boldsymbol{H}_I$, $\boldsymbol{W}_R$ and $\boldsymbol{W}_I$, respectively, by $\boldsymbol{y}_{R,n}$, $\boldsymbol{y}_{I,n}$, $\boldsymbol{h}_{R,n}$, $\boldsymbol{h}_{I,n}$, $\boldsymbol{w}_{R,n}$ and $\boldsymbol{w}_{I,n}$ and define

$$\mathcal{A}_q = \begin{bmatrix} \boldsymbol{A}_{R,q} & -\boldsymbol{A}_{I,q} \\ \boldsymbol{A}_{I,q} & \boldsymbol{A}_{R,q} \end{bmatrix}$$

$$\mathcal{B}_q = \begin{bmatrix} -\boldsymbol{B}_{I,q} & -\boldsymbol{B}_{R,q} \\ \boldsymbol{B}_{R,q} & -\boldsymbol{B}_{I,q} \end{bmatrix}, \quad q = 1, \ldots, Q \tag{5}$$

$$\underline{\boldsymbol{h}}_i = \begin{bmatrix} \boldsymbol{h}_{R,i} \\ \boldsymbol{h}_{I,i} \end{bmatrix}, \quad i = 1, \ldots, M_R. \tag{6}$$

Then we gather equations in $\boldsymbol{Y}_R$ and $\boldsymbol{Y}_I$ to form the single real-valued system of equations [1]

$$\underbrace{\begin{bmatrix} \boldsymbol{y}_{R,1} \\ \boldsymbol{y}_{I,1} \\ \vdots \\ \boldsymbol{y}_{R,M_R} \\ \boldsymbol{y}_{I,M_R} \end{bmatrix}}_{\boldsymbol{y}} = \sqrt{\frac{\rho}{M_T}} \boldsymbol{\mathcal{H}} \underbrace{\begin{bmatrix} \alpha_1 \\ \beta_1 \\ \vdots \\ \alpha_Q \\ \beta_Q \end{bmatrix}}_{\boldsymbol{x}} + \underbrace{\begin{bmatrix} \boldsymbol{w}_{R,1} \\ \boldsymbol{w}_{I,1} \\ \vdots \\ \boldsymbol{w}_{R,M_R} \\ \boldsymbol{w}_{I,M_R} \end{bmatrix}}_{\boldsymbol{w}} \tag{7}$$

where the equivalent $2M_R\tau \times 2Q$ real-valued channel matrix $\boldsymbol{\mathcal{H}}$ is given by

$$\boldsymbol{\mathcal{H}} = \begin{bmatrix} \mathcal{A}_1 \underline{\boldsymbol{h}}_1 & \mathcal{B}_1 \underline{\boldsymbol{h}}_1 & \ldots & \mathcal{A}_Q \underline{\boldsymbol{h}}_1 & \mathcal{B}_Q \underline{\boldsymbol{h}}_1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathcal{A}_1 \underline{\boldsymbol{h}}_{M_R} & \mathcal{B}_1 \underline{\boldsymbol{h}}_{M_R} & \ldots & \mathcal{A}_Q \underline{\boldsymbol{h}}_{M_R} & \mathcal{B}_Q \underline{\boldsymbol{h}}_{M_R} \end{bmatrix}. \tag{8}$$

The LD codes subsume, as special cases, both V-BLAST [5] and orthogonal STBC [13]. From (3), we can see that LD codes are very simple to encode. Furthermore, LD codes can be decoded very efficiently by several well-known MIMO demodulation algorithms, such as the sphere decoder [3], the sequential Monte Carlo (SMC)-based detector [4], the nulling-and-cancellation detector [5], as well as the simple linear detectors.

B. Problem Formulation

The LD codes introduced in [1] are designed to maximize the average mutual information between the input and output. As pointed out by [7] and [16], maximizing the average mutual information does not necessarily lead to better performance in terms of error rate. Unfortunately, the BLER is difficult to analyze for arbitrary LD codes. Simulation optimization turns out to be useful for this scenario. In this paper, we demonstrate how to optimize the average error rate for LD codes through simulation optimization with gradient estimation. First, we denote

$$\boldsymbol{h} = \left[\underline{\boldsymbol{h}}_1^T, \ldots, \underline{\boldsymbol{h}}_{M_R}^T\right]^T \tag{9}$$

and denote the set of dispersion matrices as $\boldsymbol{\theta} \triangleq \{\boldsymbol{A}_q, \boldsymbol{B}_q, q = 1, \ldots, Q\}$. With a slight abuse of notation, we also use $\boldsymbol{\theta}$ to denote the column vector that stacks all the columns of $\boldsymbol{A}_{R,q}$, $\boldsymbol{A}_{I,q}$, $\boldsymbol{B}_{R,q}$, and $\boldsymbol{B}_{I,q}$, for $q = 1, \ldots, Q$, i.e.,

$$\boldsymbol{\theta} = \Big[(\mathrm{vec}(\boldsymbol{A}_{R,1}))^T \ (\mathrm{vec}(\boldsymbol{A}_{I,1}))^T \ \cdots \ (\mathrm{vec}(\boldsymbol{B}_{R,Q}))^T \ (\mathrm{vec}(\boldsymbol{B}_{I,q}))^T\Big]^T.$$

Note that $\boldsymbol{\theta}$ is a $(4\tau M_T Q)$-dimensional vector.

Let $\boldsymbol{y}_n$ denote the nth block of the received signal corresponding to the nth block of transmitted signal $\boldsymbol{x}_n$ and channel $\boldsymbol{h}_n$, $n = 1, 2, \ldots$. Let $\gamma_{\mathrm{BL}}(\boldsymbol{y}_n, \boldsymbol{x}_n, \boldsymbol{h}_n, \boldsymbol{\theta})$ denote the empirical BLER for a given set of dispersion matrices $\boldsymbol{\theta}$. That is,

$$\gamma_{\mathrm{BL}}(\boldsymbol{y}_n, \boldsymbol{x}_n, \boldsymbol{h}_n, \boldsymbol{\theta}) = \mathbb{I}(\hat{\boldsymbol{x}}_n \neq \boldsymbol{x}_n | \boldsymbol{y}_n, \boldsymbol{h}_n, \boldsymbol{\theta}) \tag{10}$$

where $\mathbb{I}(\cdot)$ is an indicator function, $\hat{\boldsymbol{x}}_n$ is the detector output. Note that for fixed $\boldsymbol{\theta}$, since $\{\boldsymbol{y}_n, \boldsymbol{x}_n, \boldsymbol{h}_n\}$ is an i.i.d. sequence, $\{\gamma_{\mathrm{BL}}(\boldsymbol{y}_n, \boldsymbol{x}_n, \boldsymbol{h}, \boldsymbol{\theta})\}$ is also an independent identically distributed (i.i.d.) sequence of random variables. For a given set of dispersion matrices $\boldsymbol{\theta}$, the *average BLER* denoted by $\Upsilon_{\mathrm{BL}}(\boldsymbol{\theta})$ is given by

$$\begin{aligned} \Upsilon_{\mathrm{BL}}(\boldsymbol{\theta}) &= \mathbb{E}\left\{\gamma_{\mathrm{BL}}(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{h}, \boldsymbol{\theta})\right\} \\ &= \iiint \gamma_{\mathrm{BL}}(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{h}, \boldsymbol{\theta}) p(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{h} | \boldsymbol{\theta}) d\boldsymbol{y} d\boldsymbol{x} d\boldsymbol{h}. \end{aligned} \tag{11}$$

The integrals in (11) are over the space $\mathbb{R}^{2M_R\tau}$ (for $\boldsymbol{y}$), $\mathcal{A}^{2Q}$ (for $\boldsymbol{x}$), where $\mathcal{A}$ is the discrete set of real-valued constellation symbols that elements of $\boldsymbol{x}$ take value from, and $\mathbb{R}^{2M_R M_T}$ (for $\boldsymbol{h}$), respectively. For notational simplicity we subsequently omit the space over which these integrals are defined.

Aim: The design goal is to solve the following constrained stochastic optimization problem: Given the sequence of empirical BLER measurements $\gamma_{\mathrm{BL}}(\boldsymbol{y}_n, \boldsymbol{x}_n, \boldsymbol{h}_n, \boldsymbol{\theta})$ for any choice of $\boldsymbol{\theta}$, find the optimal set of dispersion matrices to minimize the average BLER, i.e., compute

$$\min_{\boldsymbol{\theta} \in \Theta} \Upsilon_{\mathrm{BL}}(\boldsymbol{\theta}) \tag{12}$$

with the energy constraint set Θ given by

$$\Theta = \left\{ \boldsymbol{\theta} \in \mathbb{R}^{4\tau M_T Q} : \sum_{q=1}^{Q} \mathrm{Tr}\left(\boldsymbol{A}_q^H \boldsymbol{A}_q + \boldsymbol{B}_q^H \boldsymbol{B}_q\right) = \boldsymbol{\theta}^T \boldsymbol{\theta} = 2\tau M_T. \right\} \tag{13}$$

From (11), we have

$$\Upsilon_{\mathrm{BL}}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{x}} \mathbb{E}_{\boldsymbol{h}} \mathbb{E}_{\boldsymbol{y}|\boldsymbol{x},\boldsymbol{h},\boldsymbol{\theta}} \left\{\gamma_{\mathrm{BL}}(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{h}, \boldsymbol{\theta})\right\} \tag{14}$$

where

$$\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x},\boldsymbol{h},\boldsymbol{\theta}} \left\{\gamma_{\mathrm{BL}}(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{h}, \boldsymbol{\theta})\right\} = \int \gamma_{\mathrm{BL}}(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{h}, \boldsymbol{\theta}) p(\boldsymbol{y} | \boldsymbol{x}, \boldsymbol{h}, \boldsymbol{\theta}) d\boldsymbol{y}. \tag{15}$$

Remark 1: Because $p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{h}, \boldsymbol{\theta})$ is Gaussian (as we will show later), and it is continuously differentiable in $\boldsymbol{\theta}$, it follows that $\Upsilon_{\mathrm{BL}}(\boldsymbol{\theta})$ is continuously differentiable in $\boldsymbol{\theta}$. (This point will be clear in the next section.) Hence, $\Upsilon_{\mathrm{BL}}(\boldsymbol{\theta})$ attains a minimum

on the compact set $\boldsymbol{\theta}$ and the optimization problem (12), (13) is well posed.

Remark 2: Note that an explicit closed-form expression for the average BLER $\Upsilon_{\mathrm{BL}}(\boldsymbol{\theta})$ is usually not available. Indeed, $\Upsilon_{\mathrm{BL}}(\boldsymbol{\theta})$ also depends on the particular detector employed (e.g., ML detector or suboptimal detector). We will use a stochastic gradient algorithm that uses measurements of the empirical BLER $\gamma_{\mathrm{BL}}(\boldsymbol{y}_n, \boldsymbol{x}_n, \boldsymbol{h}_n, \boldsymbol{\theta})$ to compute the optimal LD code $\boldsymbol{\theta}^*$. On the other hand, $\nabla_{\boldsymbol{\theta}}\Upsilon_{\mathrm{BL}}(\boldsymbol{\theta})$ cannot be computed analytically. We therefore need to devise a scheme that estimates the gradient $\nabla_{\boldsymbol{\theta}}\Upsilon_{\mathrm{BL}}(\boldsymbol{\theta})$ using the empirical BLER measurements $\gamma_{\mathrm{BL}}(\boldsymbol{y}_n, \boldsymbol{x}_n, \boldsymbol{h}_n, \boldsymbol{\theta})$.

Remark 3: Note that the reason we formulate the code design problem based on the minimum BLER criterion is that it leads to algorithms that can be proven to converge to the optimum codes. On the other hand, it remains an open problem to devise a provably convergent code design algorithm that minimizes the bit error rate (BER). We discuss more on this in Section III-E.

C. Spherical Coordinate Parameterization

Recall that the dimension of $\boldsymbol{\theta}$ is $4\tau M_T Q$. Denote $d \triangleq 4\tau M_T Q$. In this subsection, we parameterize the dispersion matrices $\boldsymbol{\theta}$ by spherical coordinates $\boldsymbol{\psi} \in \mathbb{R}^{d-1}$. We show that under these spherical coordinates $\boldsymbol{\psi}$, the constrained optimization (12), (13) transforms to the equivalent unconstrained optimization problem. This in turn implies that we can present a simple convergence proof without requiring any form of projection of the estimates which typically makes convergence proofs for stochastic approximation algorithms very difficult.

Consider $\boldsymbol{\theta}(\boldsymbol{\psi})$ parameterized by $\boldsymbol{\psi} \in \mathbb{R}^{d-1}$. Here the spherical coordinate $\psi = [\Psi_1, \ldots, \Psi_{d-1}]^T \in \mathbb{R}^{d-1}$ such that

$$\theta_p(\boldsymbol{\psi}) = \sqrt{2\tau M_T}\cos\Psi_p \prod_{k=1}^{p-1}\sin\Psi_k, \quad p = 1, 2, \cdots d-1 \tag{16}$$

$$\theta_d(\boldsymbol{\psi}) = \sqrt{2\tau M_T}\sin\Psi_{d-1} \prod_{k=1}^{d-2}\sin\Psi_k. \tag{17}$$

Remark: The above transformation is merely a higher dimensional version of the well-known two-dimensional spherical coordinate transformation $\theta_1 = \cos\Psi_1$, $\theta_2 = \sin\Psi_1$—so that $\boldsymbol{\theta}$ automatically satisfies the constraint $\theta_1^2 + \theta_2^2 = 1$.

There are three important properties of $\boldsymbol{\theta}(\psi)$ in (16), (17) that we will use.

i) $\boldsymbol{\theta}(\boldsymbol{\psi})$ involves $\sin(\cdot)$ and $\cos(\cdot)$ and hence is periodic in $\boldsymbol{\psi}$. Thus, it suffices to consider $\boldsymbol{\psi} \in [0, 2\pi]^{d-1}$ instead of $\boldsymbol{\psi} \in \mathbb{R}^{d-1}$. This boundedness of $\boldsymbol{\psi}$ makes proving convergence of a stochastic gradient algorithm very simple. Indeed one of the main reasons for the complexity in the convergence proofs of stochastic gradient algorithms is demonstrating "tightness" [25] of the estimates—i.e., that they are bounded in probability. Here $\boldsymbol{\psi}$ is bounded almost surely which is a much stronger condition!

ii) It is easy to check that the vector $\boldsymbol{\theta}(\boldsymbol{\psi}) = [\theta_1, \ldots, \theta_d]^T$ satisfies the constraint $\mathrm{Tr}(\boldsymbol{\theta}(\boldsymbol{\psi})^T\boldsymbol{\theta}(\boldsymbol{\psi})) = 2\tau M_T$. Thus, in terms of $\boldsymbol{\psi}$, $\boldsymbol{\theta}(\boldsymbol{\psi})$ automatically satisfies the constraint (13).

iii) Finally, the transformation from $\boldsymbol{\theta}$ to $\boldsymbol{\psi}$ is invertible; see (47) for an explicit formula for converting from $\boldsymbol{\psi}$ to $\boldsymbol{\theta}$. Therefore, the constrained optimization (12), (13) in $\boldsymbol{\theta}$ is equivalent to an unconstrained one in $\boldsymbol{\psi}$, as follows:

$$\min_{\boldsymbol{\theta}\in\Theta} \Upsilon_{\mathrm{BL}}(\boldsymbol{\theta}) = \min_{\boldsymbol{\psi}\in\mathbb{R}^{d-1}} \Upsilon_{\mathrm{BL}}\left(\boldsymbol{\theta}(\boldsymbol{\psi})\right). \tag{18}$$

Let $\boldsymbol{\psi}^*$ denote the local minimum of (18), and thus $\boldsymbol{\theta}^* = \boldsymbol{\theta}(\boldsymbol{\psi}^*)$ is the corresponding local minimum of (12), (13). Note that by the chain rule of differentiation

$$\nabla_{\Psi_k}\Upsilon_{\mathrm{BL}}\left(\boldsymbol{\theta}(\boldsymbol{\psi})\right) = \left(\frac{\partial\boldsymbol{\theta}}{\partial\Psi_k}\right)^T \nabla_{\boldsymbol{\theta}}\Upsilon_{\mathrm{BL}}(\boldsymbol{\theta}), \quad k = 1, 2, \ldots, d-1. \tag{19}$$

In Section III-B, estimates of $\nabla_{\boldsymbol{\theta}}\Upsilon_{\mathrm{BL}}(\boldsymbol{\theta})$ will be computed by a simulation-based gradient estimator. Here we focus on computing $\partial\boldsymbol{\theta}/\partial\Psi_p$ which is needed in the gradient estimation Algorithm 1 below. Define

$$S_{p-1} = \prod_{k=1}^{p-1}\sin\Psi_k. \tag{20}$$

For $p = 1, 2, \cdots d-2$, we have

$$\frac{\partial\boldsymbol{\theta}}{\partial\Psi_p} = \sqrt{2\tau M_T} \times \begin{bmatrix} \mathbf{0}_{p-1} \\ -\sin\Psi_p \cdot S_{p-1} \\ \cos\Psi_{p+1}\cdot\cos\Psi_p \cdot S_{p-1} \\ \cos\Psi_{p+2}\cdot\sin\Psi_{p+1}\cdot\cos\Psi_p \cdot S_{p-1} \\ \cos\Psi_{p+3}\cdot\sin\Psi_{p+2}\cdot\sin\Psi_{p+1}\cdot\cos\Psi_p \cdot S_{p-1} \\ \vdots \\ \cos\Psi_{d-1}\cdot\sin\Psi_{d-2}\cdot\cdots\cdot\sin\Psi_{p+1}\cdot\cos\Psi_p \cdot S_{p-1} \\ \sin\Psi_{d-1}\cdot\sin\Psi_{d-2}\cdot\cdots\cdot\sin\Psi_{p+1}\cdot\cos\Psi_p \cdot S_{p-1} \end{bmatrix}. \tag{21}$$

For $p = d-1$, we have

$$\frac{\partial\boldsymbol{\theta}}{\partial\Psi_p} = \sqrt{2\tau M_T}\begin{bmatrix} \mathbf{0}_{d-2} \\ -\sin\Psi_{d-1}\cdot\prod_{k=1}^{d-2}\sin\Psi_k \\ \cos\Psi_{d-1}\cdot\prod_{k=1}^{d-2}\sin\Psi_k \end{bmatrix}. \tag{22}$$

It is apparent that elements of the above vectors can be computed recursively to save computations.

III. Optimal Code Design

A. Rationale for Simulation-Based Gradient Estimation and Optimization

In this subsection, we provide a brief overview of the stochastic optimization and simulation-based gradient estimation aspects of the linear dispersion code design problem addressed in later sections of this paper. As we just described, our goal is to compute the optimal dispersion matrices $\boldsymbol{\theta}$ or equivalently spherical coordinates $\boldsymbol{\psi}$ so as to minimize the average BLER

$\Upsilon_{\mathrm{BL}}(\boldsymbol{\theta}(\boldsymbol{\psi}))$. Since it is difficult if not impossible to get a closed-form expression for the average BLER $\Upsilon_{\mathrm{BL}}(\boldsymbol{\theta}(\boldsymbol{\psi}))$ for an arbitrary dispersion matrices $\boldsymbol{\theta}$, we need to resort to a stochastic gradient algorithm to optimize $\Upsilon_{\mathrm{BL}}(\boldsymbol{\theta}(\boldsymbol{\psi}))$.

Stochastic gradient algorithms are well studied in signal processing. One of the best known stochastic gradient algorithms in adaptive filtering is the least mean-squares (LMS) algorithm. However, there is a key difference between the LMS algorithm and the stochastic gradient scheme required to minimize $\Upsilon_{\mathrm{BL}}(\boldsymbol{\theta})$. In the LMS algorithm, the gradient estimate can be computed analytically—it is simply the derivative of a quadratic function. In our case, due to the dependence of the density function on $\boldsymbol{\theta}$, we need to introduce an additional step of estimating the gradient.

The aim of gradient estimation is to compute an unbiased estimate of the true gradient. Let $\widehat{\boldsymbol{g}}(\boldsymbol{\psi})$ denote an estimate of $\nabla_{\boldsymbol{\psi}}\Upsilon_{\mathrm{BL}}(\boldsymbol{\theta}(\boldsymbol{\psi}))$, we require $\mathbb{E}\{\widehat{\boldsymbol{g}}(\boldsymbol{\psi})\} = \nabla_{\boldsymbol{\psi}}\Upsilon_{\mathrm{BL}}(\boldsymbol{\theta}(\boldsymbol{\psi}))$. Unbiased gradient estimates (or more generally asymptotically unbiased estimates) are necessary for the stochastic gradient algorithm to converge to the optimal value $\boldsymbol{\theta}^*$. A naive approach of estimating the gradient is to implement a numerical difference method for computing the gradient. Such an approach, called the Kiefer–Wolfowtiz approach [25], was developed during the 1950s. It is numerically ill-conditioned (since the definition of the gradient requires the denominator to be as small as possible), does not make use of the available information on the explicit form of the density functions, and generates biased estimates [25] with large variance.

Modern gradient estimation methods have been developed during the last ten years in the stochastic discrete-event systems community [8], [26]. These recent methods do not resort to numerical differences and they exploit the known structure of the underlying probability density function to generate unbiased estimates of the gradient. Broadly speaking there are three classes of gradient estimation algorithms that deal with density functions which are dependent on the parameter $\boldsymbol{\theta}$, namely, the score function method (also known as the likelihood ratio method) [26] which we use in this paper, the measure valued differentiation (also known as the weak derivative approach), and the process derivative approach which includes infinitesimal perturbation analysis (IPA). In the case when the random samples (e.g., $\boldsymbol{y}$, $\boldsymbol{x}$, $\boldsymbol{h}$) are timewise independent, as in this paper, the score function and measure valued differentiation perform similarly. In addition the score function method is intuitively simple to explain and implement. That is why we have used the score function method for gradient estimation in this paper. The process derivative method requires a larger computational complexity at each step and is only applicable to the ML detector where the decision boundaries can be analytically determined—see the last remark in Section III-E.

The stochastic gradient algorithm (which is known as the Robbins–Monro (R–M) algorithm [25] when $\boldsymbol{y}$, $\boldsymbol{x}$, $\boldsymbol{h}$ are timewise independent) is of the form

$$\boldsymbol{\psi}_{n+1} = \boldsymbol{\psi}_n - a_n\widehat{\boldsymbol{g}}(\boldsymbol{\psi}_n). \tag{23}$$

Here, $\boldsymbol{\psi}_n$ is the parameter value at the beginning of iteration n, $\widehat{\boldsymbol{g}}(\boldsymbol{\psi}_n)$ is the simulation-based gradient estimate of $\nabla_{\boldsymbol{\psi}}\Upsilon_{\mathrm{BL}}(\boldsymbol{\theta}(\boldsymbol{\psi}))|_{\boldsymbol{\psi}=\boldsymbol{\psi}_n}$ at iteration n, $\{a_n\}$ is a decreasing step-size sequence of positive real numbers such that

$$\sum_{n=1}^{\infty} a_n = \infty, \quad \sum_{n=1}^{\infty} a_n^2 < \infty. \tag{24}$$

The step sequence $\{a_n\}$ is usually chosen as a harmonic series $a_n = a/n$ for all n, where a is a positive scalar. It is well known that the asymptotic convergence rate of the above stochastic gradient algorithm can be made asymptotically optimal by using the ingenious procedure of iterate averaging [25]. Such a procedure uses $a_n = n^{-\nu}$, $0.5 < \nu < 1$, and a moving average of the resulting parameters. The above stochastic gradient algorithm will converge (under the conditions given in Section III-D) with probability one to a local stationary point of $\Upsilon_{\mathrm{BL}}(\boldsymbol{\theta}(\boldsymbol{\psi}))$.

B. Gradient Estimation Algorithm for ML Detector

In this subsection, we present a simulation-based gradient estimation algorithm for generating unbiased random samples $\widehat{\boldsymbol{g}}(\boldsymbol{\psi})$ of the gradient $\nabla_{\boldsymbol{\psi}}\Upsilon_{\mathrm{BL}}(\boldsymbol{\theta}(\boldsymbol{\psi}))$ for the case when the ML detector is employed. The unbiased estimates will be used in the stochastic gradient algorithm [cf. (23)] to compute the optimum $\boldsymbol{\psi}^*$ or equivalently $\boldsymbol{\theta}^* = \boldsymbol{\theta}(\boldsymbol{\psi}^*)$. For a given set of dispersion matrices $\boldsymbol{\theta}(\boldsymbol{\psi})$, a given information symbol vector $\boldsymbol{x}$, and a given channel realization $\boldsymbol{h}$, from (7) we obtain that $\boldsymbol{y}$ is Gaussian with mean $\sqrt{\frac{\rho}{M_T}}\boldsymbol{\mathcal{H}}\boldsymbol{x}$ and covariance matrix 1/2 $\bar{\boldsymbol{I}}_{2\tau M_R}$, namely

$$p\left(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{h},\boldsymbol{\theta}(\boldsymbol{\psi})\right) = \pi^{-\tau M_R}\exp\left(-\left(\boldsymbol{y}-\sqrt{\frac{\rho}{M_T}}\boldsymbol{\mathcal{H}}\boldsymbol{x}\right)^T \left(\boldsymbol{y}-\sqrt{\frac{\rho}{M_T}}\boldsymbol{\mathcal{H}}\boldsymbol{x}\right)\right). \tag{25}$$

We propose the following two-stage simulation algorithm to generate the gradient estimate $\widehat{\boldsymbol{g}}(\boldsymbol{\psi})$.

Algorithm 1 *[Composite-score function algorithm]*

Given the set of dispersion matrices $\boldsymbol{\theta}(\boldsymbol{\psi}_k)$ at the kth iteration, perform the following simulation steps:

Step 1) Composition method to generate mixture sample:

a) Draw M symbol vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_M$ uniformly from the constellation set $\mathcal{A}^{2Q}$.

b) Simulate M observations $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_M$ where each $\boldsymbol{y}_i$ is generated according to (7) using symbol vector $\boldsymbol{x}_i$, i.e.,

$$\boldsymbol{y}_i = \sqrt{\frac{\rho}{M_T}}\boldsymbol{\mathcal{H}}_i\boldsymbol{x}_i + \boldsymbol{w}_i, \quad i = 1,2,\ldots,M. \tag{26}$$

c) Using the ML detection algorithm, decode $\boldsymbol{x}_i$ based on the observations $\boldsymbol{y}_i$ and the channel value $\boldsymbol{\mathcal{H}}_i$, $i = 1,2,\ldots,M$.

Compute the empirical BLER $\gamma_{\mathrm{BL}}(\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{h}_i, \boldsymbol{\theta}(\boldsymbol{\psi}_k))$.

Step 2) Score function method for gradient estimation: Using the empirical BLER $\gamma_{\mathrm{BL}}(\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{h}_i, \boldsymbol{\theta}(\boldsymbol{\psi}_k))$, compute the gradient estimate as

$$\widehat{\boldsymbol{g}}(\boldsymbol{\psi}_k)=\frac{1}{M}\sum_{i=1}^{M}\gamma_{\mathrm{BL}}(\boldsymbol{y}_i,\boldsymbol{x}_i,\boldsymbol{h}_i,\boldsymbol{\theta}_k)[\nabla_{\boldsymbol{\psi}}\log p(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{h}_i,\boldsymbol{\theta})|_{\boldsymbol{\psi}=\boldsymbol{\psi}_k}] \tag{27}$$

where the element of $\nabla_{\boldsymbol{\psi}}\log p(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{h}_i,\boldsymbol{\theta})$ is given by

$$\frac{\partial \log p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{h},\boldsymbol{\theta})}{\partial \Psi_p} = [\nabla_{\boldsymbol{\theta}}\log p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{h},\boldsymbol{\theta})]^T\left(\frac{\partial \boldsymbol{\theta}}{\partial \Psi_p}\right), \quad \text{for} \quad p=1,\cdots,d-1. \tag{28}$$

An explicit formula for $\nabla_{\boldsymbol{\theta}}\log p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{h},\boldsymbol{\theta})$ is given in Appendix I. $\partial\boldsymbol{\theta}/\partial\Psi_p$ is given by (21) and (22).

The following theorem shows that the simulation-based gradient estimate $\widehat{\boldsymbol{g}}(\boldsymbol{\psi})$ generated by the above algorithm is indeed an unbiased estimator.

Theorem 1: Under the conditions: 1) $\nabla_{\boldsymbol{\psi}}p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{h},\boldsymbol{\theta}(\boldsymbol{\psi}))$ exists for all ψ; 2)$\nabla_{\boldsymbol{\psi}}p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{h},\boldsymbol{\theta}(\boldsymbol{\psi}))/p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{h},\boldsymbol{\theta}(\boldsymbol{\psi}))$ is uniformly bounded for all $\boldsymbol{\psi}\in\mathbb{R}^{d-1}$; then Algorithm 1 generates unbiased samples, i.e.,

$$\mathbb{E}\left\{\widehat{\boldsymbol{g}}(\boldsymbol{\psi}_k)\right\} = \nabla_{\boldsymbol{\psi}}\Upsilon_{\mathrm{BL}}\left(\boldsymbol{\theta}(\boldsymbol{\psi})\right)|_{\boldsymbol{\psi}=\boldsymbol{\psi}_k}, \quad k=1,2,\cdots. \tag{29}$$

Proof: Conditions (1) and (2) hold, since $p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{h},\boldsymbol{\theta}(\boldsymbol{\psi}))$ is continuously differentiable with respect to $\boldsymbol{\psi}$, and $\boldsymbol{\theta}(\boldsymbol{\psi})$ is periodic and bounded in $\boldsymbol{\psi}$. From (19)

$$\nabla_{\Psi_k}\Upsilon_{\mathrm{BL}}\left(\boldsymbol{\theta}(\boldsymbol{\psi})\right) = \left(\frac{\partial\boldsymbol{\theta}}{\partial\Psi_k}\right)^T\nabla_{\boldsymbol{\theta}}\Upsilon_{\mathrm{BL}}(\boldsymbol{\theta}). \tag{30}$$

Consider the term $\nabla_{\boldsymbol{\theta}}\Upsilon_{\mathrm{BL}}(\boldsymbol{\theta})$ in the RHS of the above equation. From (14) and (15), we have

$$\begin{aligned}
&\nabla_{\boldsymbol{\theta}}\Upsilon_{\mathrm{BL}}(\boldsymbol{\theta})\\
&\quad=\mathbb{E}_{\boldsymbol{x}}\mathbb{E}_{\boldsymbol{h}}\left\{\nabla_{\boldsymbol{\theta}}\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x},\boldsymbol{h},\boldsymbol{\theta}}\left\{\gamma_{\mathrm{BL}}(\boldsymbol{y},\boldsymbol{x},\boldsymbol{h},\boldsymbol{\theta})\right\}\right\}\\
&\quad=\mathbb{E}_{\boldsymbol{x}}\mathbb{E}_{\boldsymbol{h}}\int\nabla_{\boldsymbol{\theta}}\left[\gamma_{\mathrm{BL}}(\boldsymbol{y},\boldsymbol{x},\boldsymbol{h},\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{h},\boldsymbol{\theta})\right]d\boldsymbol{y}\\
&\quad=\mathbb{E}_{\boldsymbol{x}}\mathbb{E}_{\boldsymbol{h}}\int\left[\left(\nabla_{\boldsymbol{\theta}}\gamma_{\mathrm{BL}}(\boldsymbol{y},\boldsymbol{x},\boldsymbol{h},\boldsymbol{\theta})\right)p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{h},\boldsymbol{\theta})\right.\\
&\qquad\left.+\gamma_{\mathrm{BL}}(\boldsymbol{y},\boldsymbol{x},\boldsymbol{h},\boldsymbol{\theta})\nabla_{\boldsymbol{\theta}}p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{h},\boldsymbol{\theta})\right]d\boldsymbol{y}.
\end{aligned} \tag{31}$$

We now use the following proposition whose proof appears in Appendix II.

Proposition 1: For the ML detector, we have

$$\mathbb{E}_{\boldsymbol{x}}\int\nabla_{\boldsymbol{\theta}}\gamma_{\mathrm{BL}}(\boldsymbol{y},\boldsymbol{x},\boldsymbol{h},\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{h},\boldsymbol{\theta})d\boldsymbol{y}=0. \tag{32}$$

Using (31) and (32) in (30), we have

$$\begin{aligned}
&\nabla_{\Psi_k}\Upsilon_{\mathrm{BL}}\left(\boldsymbol{\theta}(\boldsymbol{\psi})\right)\\
&\quad=\left(\frac{\partial\boldsymbol{\theta}}{\partial\Psi_k}\right)^T\mathbb{E}_{\boldsymbol{x}}\mathbb{E}_{\boldsymbol{h}}\\
&\qquad\times\int\gamma_{\mathrm{BL}}(\boldsymbol{y},\boldsymbol{x},\boldsymbol{h},\boldsymbol{\theta})\nabla_{\boldsymbol{\theta}}p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{h},\boldsymbol{\theta})d\boldsymbol{y}\\
&\quad=\left(\frac{\partial\boldsymbol{\theta}}{\partial\Psi_k}\right)^T\mathbb{E}_{\boldsymbol{x}}\mathbb{E}_{\boldsymbol{h}}\\
&\qquad\times\int\gamma_{\mathrm{BL}}(\boldsymbol{y},\boldsymbol{x},\boldsymbol{h},\boldsymbol{\theta})\frac{\nabla_{\boldsymbol{\theta}}p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{h},\boldsymbol{\theta})}{p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{h},\boldsymbol{\theta})}p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{h},\boldsymbol{\theta})d\boldsymbol{y}\\
&\quad=\left(\frac{\partial\boldsymbol{\theta}}{\partial\Psi_k}\right)^T\mathbb{E}_{\boldsymbol{x}}\mathbb{E}_{\boldsymbol{h}}\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x},\boldsymbol{h},\boldsymbol{\theta}}\\
&\qquad\times\left\{\gamma_{\mathrm{BL}}(\boldsymbol{y},\boldsymbol{x},\boldsymbol{h},\boldsymbol{\theta})\nabla_{\boldsymbol{\theta}}\log p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{h},\boldsymbol{\theta})\right\}.
\end{aligned} \tag{33}$$

The above equation shows that $\gamma_{\mathrm{BL}}(\boldsymbol{y},\boldsymbol{x},\boldsymbol{h},\boldsymbol{\theta})(\partial\boldsymbol{\theta}/\partial\Psi_k)^T\nabla_{\boldsymbol{\theta}}\log p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{h},\boldsymbol{\theta})$ is an unbiased estimate of $\nabla_{\Psi_k}\Upsilon_{\mathrm{BL}}(\boldsymbol{\theta}(\boldsymbol{\psi}))$. Theorem 1 follows from the definition of $\widehat{\boldsymbol{g}}(\boldsymbol{\psi}_k)$ in (27). ■

C. Stochastic Approximation Algorithm

In this subsection, we use the above gradient estimator in a stochastic approximation algorithm to solve the optimization problem given in (18).

Algorithm 2 *[Optimal LD codes design]*

Assume at the kth iteration the current set of dispersion matrices is $\boldsymbol{\theta}(\boldsymbol{\psi}_k)$ with the coordinate parameterization $\boldsymbol{\psi}_k$, perform the following steps during the next iteration to generate $\boldsymbol{\psi}_{k+1}$ and the corresponding $\boldsymbol{\theta}(\boldsymbol{\psi}_{k+1})$:

Steps 1 & 2) Same as Steps 1) and 2) of Algorithm 1.

Step 3) Update new set of dispersion matrices: Generate

$$\boldsymbol{\psi}_{k+1}=\boldsymbol{\psi}_k - a_k\widehat{\boldsymbol{g}}(\boldsymbol{\psi}_k) \tag{34}$$

then update $\boldsymbol{\theta}(\boldsymbol{\psi}_{k+1})$ according to (16) and (17).

The convergence of the above algorithm is given by the following theorem:

Theorem 2: Under the conditions: 1) the step size $\{a_n\}$ satisfies (24); 2) $\widehat{\boldsymbol{g}}(\boldsymbol{\psi}_k)$ is uniformly integrable; 3) $\nabla_{\boldsymbol{\psi}}\Upsilon_{\mathrm{BL}}(\boldsymbol{\theta}(\boldsymbol{\psi}))$ is Lipschitz for all $\boldsymbol{\psi}$ in $[0,2\pi]^{d-1}$; then the sequence of estimate $\{\boldsymbol{\psi}_k\}$ generated by the stochastic approximation Algorithm 2 converges almost surely to a local stationary point $\boldsymbol{\psi}^*$, or equivalently $\boldsymbol{\theta}(\boldsymbol{\psi}_k)$ converges almost surely to $\boldsymbol{\theta}^*$.

We refer the reader to of [25, Th. 2.5, Ch. 5, p. 105] for a general proof with martingale difference noise and to [25, Ch. 6] for Markovian noise. Actually the proof in our case is considerably simpler since the noise samples are i.i.d. The uniform integrability of $\widehat{\boldsymbol{g}}(\boldsymbol{\psi}_k)$ follows straightforwardly as following: In (27), the first term on the RHS is uniformly bounded since it is an indicator. From (28) the second term of (27) comprises

of the product of two subterms. The second subterm $\partial\boldsymbol{\theta}/\partial\Psi_p$ is uniformly bounded since it is the product of $\sin(\cdot)$ and $\cos(\cdot)$ functions, see (21), (22). The first subterm of (28) is uniformly integrable since from Appendix I it involves $\left(\boldsymbol{y}-\sqrt{\frac{\rho}{M_T}}\boldsymbol{\mathcal{H}}\boldsymbol{x}\right)$ which is a finite variance Gaussian by assumption (see also (2)).

D. Gradient Estimation for Suboptimal Detectors

Linear Detector: A linear detector applies a linear transformation $\tilde{\boldsymbol{W}}$—e.g., linear zero-forcing detector $\tilde{\boldsymbol{W}} = \boldsymbol{\mathcal{H}}(\boldsymbol{\mathcal{H}}^T\boldsymbol{\mathcal{H}})^{-1}$, or linear minimum mean square error (MMSE) detector $\tilde{\boldsymbol{W}} = \boldsymbol{\mathcal{H}}(\frac{\rho}{M_T}\boldsymbol{\mathcal{H}}^T\boldsymbol{\mathcal{H}} + 1/2)^{-1}$—on the received signal $\boldsymbol{y}$ in (7), and then performs symbol-by-symbol threshold detection. After the linear transformation $\tilde{\boldsymbol{z}} = \tilde{\boldsymbol{W}}^T\boldsymbol{y}$, we have

$$\tilde{z}_\ell = a_\ell x_\ell + \tilde{v}_\ell + \tilde{\mu}_\ell, \text{ with } x_\ell \in \mathcal{A},\ v_\ell \sim \mathcal{N}\left(0,\sigma_\ell^2\right), \quad \ell = 1,\ldots,2Q. \tag{35}$$

where $\tilde{\mu}_\ell$ represents the residual multiple-access interference (MAI) for the MMSE detector, and $\tilde{\mu}_\ell = 0$ for the ZF detector. We then perform the following normalization

$$z_\ell \triangleq \frac{\tilde{z}_\ell}{a_\ell} = x_\ell + v_\ell + \mu_\ell, \text{ with } v_\ell \sim \mathcal{N}\left(0,\frac{\sigma_\ell^2}{a_\ell^2}\right), \quad \ell = 1,\ldots,2Q. \tag{36}$$

Note that for either detector, the decision region for x_ℓ is fixed and given z_ℓ the decision rule is independent of $\boldsymbol{h},\boldsymbol{\theta}$ MAI and noise variance. Denote $\boldsymbol{A} \triangleq \operatorname{diag}(a_1,\ldots,a_{2Q}) = \operatorname{diag}\left(\sqrt{\frac{\rho}{M_T}}\tilde{\boldsymbol{W}}^T\boldsymbol{\mathcal{H}}\right)$, and $\boldsymbol{W} \triangleq \tilde{\boldsymbol{W}}\boldsymbol{A}^{-1}$, then we have

$$\boldsymbol{z} \triangleq [z_1,\ldots,z_{2Q}]^T = \boldsymbol{W}^T\boldsymbol{y} \sim \mathcal{N}\left(\sqrt{\frac{\rho}{M_T}}\boldsymbol{W}^T\boldsymbol{\mathcal{H}}\boldsymbol{x},\frac{1}{2}\boldsymbol{W}^T\boldsymbol{W}\right). \tag{37}$$

Note that in the ZF case, we have $\sqrt{\frac{\rho}{M_T}}\boldsymbol{W}^T\boldsymbol{\mathcal{H}} = \boldsymbol{I}$. The empirical BLER is then denoted as

$$\gamma_{\mathrm{BL}}(\boldsymbol{z},\boldsymbol{x}) = 1 - \prod_{\ell=1}^{2Q}\mathbb{I}(\hat{x}_\ell = x_\ell|\boldsymbol{z}). \tag{38}$$

Hence the average BLER is given by

$$\Upsilon_{\mathrm{BL}}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{x}}\mathbb{E}_{\boldsymbol{h}}\mathbb{E}_{\boldsymbol{z}|\boldsymbol{x},\boldsymbol{h},\boldsymbol{\theta}}\left\{\gamma_{\mathrm{BL}}(\boldsymbol{z},\boldsymbol{x})\right\}. \tag{39}$$

Due to the above normalization and the symbol-by-symbol threshold decision rule, γ_{BL} in (38) is independent of θ. The gradient is then

$$\nabla_{\boldsymbol{\theta}}\Upsilon_{\mathrm{BL}}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{x}}\mathbb{E}_{\boldsymbol{h}}\mathbb{E}_{\boldsymbol{z}|\boldsymbol{x},\boldsymbol{h},\boldsymbol{\theta}}\{\gamma_{\mathrm{BL}}(\boldsymbol{z},\boldsymbol{x})\nabla_{\boldsymbol{\theta}}\log p(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{h},\boldsymbol{\theta})\}. \tag{40}$$

The above equation shows that $\gamma_{\mathrm{BL}}(\boldsymbol{z},\boldsymbol{x})\ \nabla_{\boldsymbol{\theta}}\log p(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{h},\boldsymbol{\theta})$ is an unbiased estimate of $\nabla_{\boldsymbol{\theta}}\Upsilon_{\mathrm{BL}}(\boldsymbol{\theta})$. Note that by (37) $p(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{h},\boldsymbol{\theta})$ is a multivariate Gaussian density whose covariance is parameterized by $\boldsymbol{\mathcal{H}}$ and $\boldsymbol{W}$ (which is a function of $\boldsymbol{\mathcal{H}}$), both of which are explicit functions of $\boldsymbol{\theta}$. Hence, $\nabla_{\boldsymbol{\theta}}\log p(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{h},\boldsymbol{\theta})$ can be obtained analytically.

Nulling-and-Cancellation Detector: In the nulling-and-cancellation detector, we first perform a QR decomposition on the channel matrix to obtain $\sqrt{\frac{\rho}{M_T}}\boldsymbol{\mathcal{H}} = \boldsymbol{Q}\bar{\boldsymbol{R}}$, where $\boldsymbol{Q}$ is a unitary matrix, and $\bar{\boldsymbol{R}}$ is an upper triangular matrix. We then perform the following nulling operation:

$$\tilde{\boldsymbol{z}} \triangleq \boldsymbol{Q}^T\boldsymbol{y} = \bar{\boldsymbol{R}}\boldsymbol{x} + \tilde{\boldsymbol{v}} \Longrightarrow \tilde{z}_\ell = \sum_{j=\ell}^{2Q} r_{\ell,j}x_j + \tilde{v}_\ell, \quad \ell = 1,\ldots,2Q \tag{41}$$

where $\tilde{\boldsymbol{v}} \sim \mathcal{N}(\boldsymbol{0},1/2\boldsymbol{I}_{2Q})$. We first detect symbol x_{2Q} based on the following normalized $\tilde{z}_{2Q}$:

$$z_{2Q} \triangleq \frac{z_{2Q}}{r_{2Q,2Q}} = x_{2Q} + v_{2Q}, \text{ with } v_{2Q} \sim \mathcal{N}\left(0,\frac{1}{2r_{2Q,2Q}^2}\right). \tag{42}$$

Subsequently the detection of x_ℓ is based on

$$z_\ell \triangleq \frac{z_\ell - \sum_{j=\ell+1}^{2Q} r_{\ell,j}\hat{x}_j}{r_{\ell,\ell}}, \quad \ell = 2Q-1,\ldots,1. \tag{43}$$

Note that in terms of BLER performance, the above decision-feedback detection is exactly the same as a "genie-aided" detector where *perfect feedback* is assumed in (43) [28]. Hence, as far as BLER is concerned, we can write (43) equivalently as

$$z_\ell \triangleq \frac{z_\ell - \sum_{j=\ell+1}^{2Q} r_{\ell,j}x_j}{r_{\ell,\ell}} = x_\ell + v_\ell, \quad \text{with } v_\ell \sim \mathcal{N}\left(0,\frac{1}{2r_{\ell,\ell}^2}\right), \quad \ell = 2Q-1,\ldots,1. \tag{44}$$

Denote the decision statistics

$$\boldsymbol{z} \triangleq [z_1,\ldots,z_{2Q}]^T \sim \mathcal{N}\left(\boldsymbol{x},\frac{1}{2}\operatorname{diag}\left(\frac{1}{r_{11}^2},\ldots,\frac{1}{r_{2Q,2Q}^2}\right)\right). \tag{45}$$

Then as in the above linear detector case, the empirical BLER and the average BLER are given respectively by (38) and (39). Moreover, again since the symbol decision regions are independent of $\boldsymbol{\theta}$, the gradient is given by (40), with $p(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{h},\boldsymbol{\theta})$ replaced by (45). Now in order to find an analytical expression for $\nabla_{\boldsymbol{\theta}}\log p(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{h},\boldsymbol{\theta})$, we need to find an explicit expression for $1/r_{\ell,\ell}^2$, $\ell = 1,\ldots,2Q$ in terms of $\boldsymbol{\mathcal{H}}$.

To that end, denote $\boldsymbol{g}_\ell$ as the ℓth column of $\sqrt{\frac{\rho}{M_T}}\boldsymbol{\mathcal{H}}$, i.e., $\sqrt{\frac{\rho}{M_T}}\boldsymbol{\mathcal{H}} = [\boldsymbol{g}_1,\ldots,\boldsymbol{g}_{2Q}] \triangleq \boldsymbol{G}_{2Q}$. Denote further $\boldsymbol{G}_\ell \triangleq [\boldsymbol{g}_1,\ldots,\boldsymbol{g}_\ell]$, $\ell = 1,\ldots,2Q$. Denote $\boldsymbol{Q}_\ell = \boldsymbol{G}_\ell^T\boldsymbol{G}_\ell$, $\ell = 1,\ldots,2Q$. Then it can be shown that

$$\frac{1}{r_{\ell,\ell}^2} = \left[\boldsymbol{Q}_\ell^{-1}\right]_{\ell,\ell}, \quad \ell = 1,\ldots,2Q. \tag{46}$$

Hence, indeed $1/r_{\ell,\ell}^2, \ell = 1,\ldots,2Q$ can be expressed explicitly in terms of $\boldsymbol{\mathcal{H}}$. Therefore, $\nabla_{\boldsymbol{\theta}}\log p(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{h},\boldsymbol{\theta})$ can be obtained analytically.

E. Implementation Details and Discussions

Variance Reduction: In simulation-based optimization (e.g., Algorithm 2), reducing the variance of the gradient estimates leads to improved convergence rate. Since our LD code design is offline, variance reduction to improve convergence rate is not of great importance. We simply remark that the variance can be reduced by choosing large M in Step 2) of Algorithm 2 (recall that the gradient estimate is unbiased for any integer M). From our experience, we find that choosing M to be (around) 1000 works quite well.

SNR Dependence of Convergence Rate: From the previous discussions, the optimal code $\boldsymbol{\theta}^*$ is a function of the number of transmit antennas M_T, the number of receive antennas M_R, and the signal constellation constellation $\mathcal{A}$. Moreover $\boldsymbol{\theta}^*$ is also a function of the operating SNR ρ as both the empirical BLER $\gamma_{\mathrm{BL}}(\boldsymbol{y},\boldsymbol{x},\boldsymbol{h},\boldsymbol{\theta})$ and the gradient $\nabla_{\boldsymbol{\theta}}\log p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{h},\boldsymbol{\theta})$ depend on ρ (see Appendix). Therefore, the optimal code we obtain is SNR-dependent (see also [1]). However, as we will demonstrate through examples in Section V, the codes optimized for a particular SNR work fine for a wide range of SNR of interest.

Initialization of Algorithm 2: To initialize the LD code design Algorithm 2, we need to specify an initial $\boldsymbol{\psi}_0$ (and hence $\boldsymbol{\theta}(\boldsymbol{\psi}_0)$). In numerical studies, we found that if we randomly generated $\boldsymbol{\psi}_0$, then sometimes many of the components of $\boldsymbol{\theta}(\boldsymbol{\psi}_0)$ would be close to zero, leading to numerical problems in the detection algorithm. To overcome this difficulty, we first randomly generate the vector $\boldsymbol{\theta}_0$ satisfying the energy constraint (13), and then compute the corresponding $\boldsymbol{\psi}_0$ by inverting the spherical coordinate transformation (16), (17) as follows:

$$\begin{aligned}\Psi_1 &= \arccos\left(\frac{\theta_1}{\sqrt{2\tau M_T}}\right)\\ \Psi_p &= \arccos\left(\frac{\theta_p}{\sqrt{2\tau M_T}\prod_{k=1}^{p-1}\sin\Psi_k}\right),\\ &\qquad \text{for } p=2,\cdots,d-1.\end{aligned} \tag{47}$$

Computation Complexity of the Algorithm: The complexity of the proposed method comes from the Monte Carlo simulations as well as the gradient computations. From (27), we only need to compute the gradient of the log-likelihood function when the empirical BLER is not zero. The overall complexity can still be high especially for large number of antennas and/or with high data rates as the simulation can be slow and also we have to compute gradients for more parameters. Fortunately, since the design is a purely offline design, once the codes are designed the implementation complexity (encoding and decoding) is the same as the general LD codes.

Local Convergence: Since Algorithm 2 is a stochastic gradient algorithm, it converges to a local minimum $\boldsymbol{\psi}^*$. By trying different initial conditions, and picking the best solution, we can obtain better codes. Another possibility is to use a simulated annealing based stochastic gradient scheme [11] to obtain the global optimum solution.

Performance-Complexity Trade-off: In general, the parameters τ and Q are also the design variables. Usually τ is chosen to be equal to the number of transmit antennas to guarantee full spatial diversity, and Q is chosen to be $\tau\min(M_T,M_R)$ to facilitate the polynomial-time ML detection [16]. In some scenarios, we can choose τ and Q to obtain the best performance–complexity tradeoff. For example, we can choose $\tau < M_T$ to reduce detection complexity. One of the advantages of Algorithm 2 is that we can design optimal LD codes with arbitrary τ and Q.

Approximate Score Function Method for Suboptimal Detectors: As seen in Section III-D, the *exact* implementation of the score function method calls for the use of $\nabla_{\boldsymbol{\theta}}\log p(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{h},\boldsymbol{\theta})$, which is detector-dependent. Fortunately, numerical studies indicate that if we simply employ (27), (28) for gradient estimation for suboptimal detectors, and replace γ_{BL} by the empirical BLER of the corresponding detector, we can still obtain very good codes. Such an *approximate* implementation of the score function method provides a universal code design algorithm as we can design optimal LD codes for various types of detector structures, i.e., exactly the same stochastic gradient Algorithm 2 can be used with gradient estimator given in Algorithm 1 to minimize the average BLER given empirical estimates of the BLER from the suboptimal detector. Of course, in general the optimal LD codes for different detection structures may be different.

LD Code Design for Arbitrary Fading Statistics: Most work on space–time codes assumes the idealistic case of i.i.d. Rayleigh MIMO channels. In reality, the individual antennas could be correlated due to insufficient antenna spacing and lack of scattering [14], [15]. It is very difficult (if not impossible) to optimize the design of space–time codes analytically for a specific transmit and receive correlation structure. In [20], the authors propose to combine beamforming with space–time coding for the case of transmit correlation only. However, only orthogonal STBC is considered which incurs a significant rate loss. Our algorithm turns out to be useful in this scenario as well. Assume there is correlation at both the transmitter side and the receiver side. We employ the spatial fading correlation model in [14], wherein the channel matrix $\bar{\boldsymbol{H}}$ can be decomposed into three parts, namely

$$\bar{\boldsymbol{H}} = \boldsymbol{S}^{\frac{1}{2}}\boldsymbol{H}_{\omega}\bar{\boldsymbol{R}}^{\frac{1}{2}} \tag{48}$$

where $\boldsymbol{H}_{\omega}$ is an $M_T \times M_R$ matrix composed of i.i.d. complex Gaussian entries with zero mean and unit variance, and $\boldsymbol{S}=\boldsymbol{S}^{1/2}(\boldsymbol{S}^{1/2})^H$ and $\boldsymbol{R}=\boldsymbol{R}^{1/2}(\boldsymbol{R}^{1/2})^H$ are the transmit and receive correlation matrices, respectively. When the long-term correlation, i.e., $\boldsymbol{S}$ and $\bar{\boldsymbol{R}}$, can be measured in advance, such knowledge can easily be taken into account in our LD code design framework. The only modification in the algorithm for this case is that during the first step of the algorithm the channel matrix $\boldsymbol{H}$ is randomly generated according to (48). All the other steps remain the same. Algorithm 2 will "automatically" generate the optimal LD codes adapting to the specific correlation structure. Moreover, the proposed design framework can also be employed to design codes for MIMO channels that exhibit more complicated fading statistics, e.g., Rician fading, or Nakagami fading, a task traditional approaches are unable to accomplish.

Minimum BER LD Code Design: So far, we have assumed that the performance metric in our code design is the BLER.

Another performance metric of practical interest is the BER. The empirical bit error function is defined as

$$\gamma_b(\boldsymbol{y}_n, \boldsymbol{x}_n, \boldsymbol{h}_n, \boldsymbol{\theta}) = \left(\frac{\#\ \text{of bit errors in}\ \boldsymbol{x}_n \to \hat{\boldsymbol{x}}_n}{Q \log_2 r} \right) \times \mathbb{I}(\boldsymbol{x}_n \to \hat{\boldsymbol{x}}_n | \boldsymbol{y}_n, \boldsymbol{h}_n, \boldsymbol{\theta}). \quad (49)$$

Note that for a particular detector, $\hat{\boldsymbol{x}}_n$ is a deterministic function of $\boldsymbol{y}_n, \boldsymbol{h}_n, \boldsymbol{\theta}$. The objective function (14) is then replaced by

$$\Upsilon_b(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{x}} \mathbb{E}_{\boldsymbol{h}} \mathbb{E}_{\boldsymbol{y}|\boldsymbol{x},\boldsymbol{h},\boldsymbol{\theta}} \left\{ \gamma_b(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{h}, \boldsymbol{\theta}) \right\}. \quad (50)$$

As in (31) its gradient $\nabla_{\boldsymbol{\theta}} \Upsilon_b(\boldsymbol{\theta})$ can be decomposed into two terms. An unbiased estimate of the second term, i.e., $\mathbb{E}_{\boldsymbol{x}} \mathbb{E}_{\boldsymbol{h}} \int \gamma_b(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{h}, \boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{h}, \boldsymbol{\theta}) d\boldsymbol{y}$, can be obtained using Algorithm 1, with $\gamma_{\mathrm{BL}}(\cdot)$ replaced by $\gamma_b(\cdot)$. On the other hand, the first term of the gradient, i.e., $\mathbb{E}_{\boldsymbol{x}} \mathbb{E}_{\boldsymbol{h}} \int \nabla_{\boldsymbol{\theta}} \gamma_{\mathrm{BL}}(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{h}, \boldsymbol{\theta}) p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{h}, \boldsymbol{\theta}) d\boldsymbol{y}$, is no longer zero. That is, Proposition 1 does *not* hold under the BER metric. The *exact* design algorithm for the minimum-BER LD code remains an open problem. Nevertheless, numerical studies indicate that if we simply employ Algorithms 1 and 2, and replace the empirical BLER γ_{BL} by the empirical BER γ_b, we can obtain codes with very good BER performance.

Pathwise Derivative Estimate for ML Detector: It can be shown similar to the proof of Prop 1 in Appendix II that

$$\nabla_{\boldsymbol{\theta}} [1 - \Upsilon_{\mathrm{BL}}(\boldsymbol{\theta})] = - \sum_{i=1}^{|\mathcal{A}|^{2Q}} \nabla_{\boldsymbol{\theta}} P(\text{ no error} | \boldsymbol{s}_i) P(\boldsymbol{s}_i) \quad (51)$$

with

$$\nabla_{\boldsymbol{\theta}} P(\text{ no error} | \boldsymbol{s}_i) = -\mathbb{E} \left\{ \sum_{j \neq i} \boldsymbol{a}_{j|i}^T \nabla_{\boldsymbol{\theta}} \boldsymbol{\mathcal{H}}^T \boldsymbol{w} \times \prod_{\ell \neq j} \mathbb{I} \left(\boldsymbol{a}_{\ell|i}^T \boldsymbol{\theta}^T \boldsymbol{w} \leq b_{\ell|i} | \boldsymbol{a}_{j|i}^T \boldsymbol{\theta}^T \boldsymbol{w} = b_{j|i} \right) \right\}. \quad (52)$$

where $\boldsymbol{a}_{j|i}$ and $b_{j|i}$ are defined in (71). Just as for the score function algorithm, the composition method can be used to efficiently simulate (51). However, simulating unbiased estimates from (52) requires computing the summation over all the symbols which is numerically expensive. Also note that for other detectors such as the nulling-and-cancellation detector, an expression for $\nabla_{\boldsymbol{\theta}}[1 - \Upsilon_{\mathrm{BL}}(\boldsymbol{\theta})]$ is not available.

IV. Simulation Results

In this section, we give three examples that illustrate the performance of the LD codes obtained by Algorithm 2. As we have mentioned in Section III-E, we first randomly generate the dispersion matrices with proper scaling such that the energy constraint is satisfied. We then obtain the initial spherical coordinates via (47). Note that different random initializations might lead to different LD codes. However, we have found that the codes generated with different initial conditions usually end up with very similar error performance. In the examples given in this section, only one random initial point is used to feed into Algorithm 2. Note also that our code design depends on the operating SNR. In the following examples, we design the codes by choosing the SNR so that the BLER is around 10^{-2}.[1] We will see that the codes optimized for a particular SNR work fine for a wide range of SNR of interest.

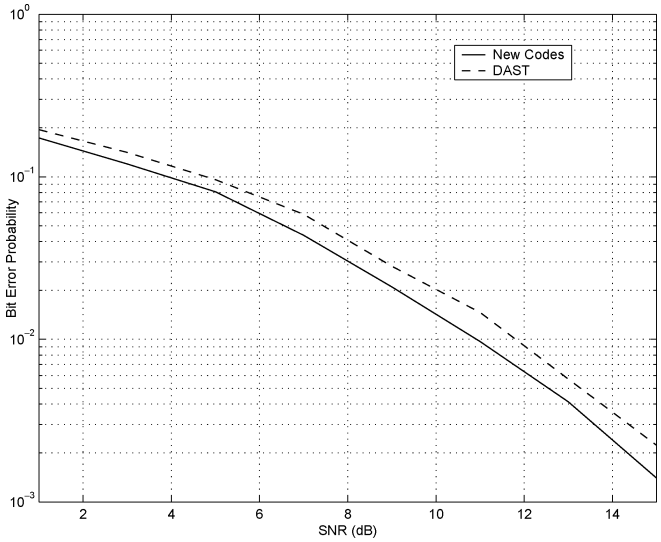

Fig. 1. BER performance of the new LD code and the DAST code in 3×1 uncorrelated MIMO channel with QPSK constellation.

All codes are designed under the BLER criterion. To illustrate the performance of the optimized codes, we show their BER performance together with the BER performance of some known good codes in the literature. Here instead of using the exact score function algorithm to design codes for the nulling-and-cancellation detector, we simply employed Algorithm 2 to design the corresponding codes.

Example 1—New LD Codes With ML Detector: We first present simulation results for i.i.d. fading channels using the sphere decoder. Fig. 1 compares the BER performance of the new LD code (obtained using Algorithm 2) with that of the TAST code for a system employing three transmit antennas and one receive antenna, and QPSK constellation. The data rate $R = 2$ b/s/Hz. In this case, the TAST code is actually the DAST code proposed in [21]. At the BER of 10^{-2}, the gain of the new LD code over the DAST code is about 1 dB. Next we consider a system with three transmit antennas and two receive antennas. Figs. 2 and 3 show the BER performance of the new LD codes and the TAST codes employing QPSK and 16QAM constellations, respectively. The rate R is 4 b/s/Hz for the codes in Fig. 2 and 8 b/s/Hz for the codes in Fig. 3. In Fig. 2, we also plot the BER of a randomly generated LD code. It is seen that the new codes perform better than the TAST codes for a wide range of SNR. Note that the new LD codes have the same encoding and decoding complexity as the TAST codes. Note also that the new LD codes in Fig. 2 and Fig. 3 are different as the proposed design depends on the specific symbol constellation employed. Similar results are given in Fig. 4 and Fig. 5

[1]For example, the LD code in Fig. 4 is searched with SNR = 8 dB, while the LD code in Fig. 5 is searched with SNR = 14 dB.

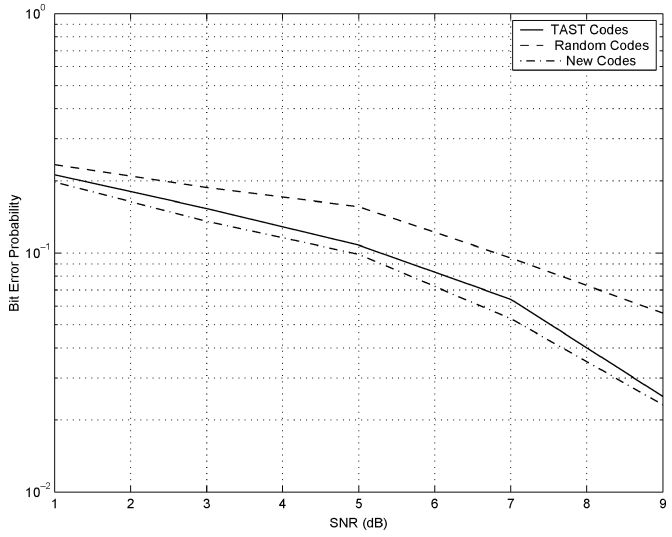

Fig. 2. BER performance of the new LD code and the TAST code in 3×2 uncorrelated MIMO channel with QPSK constellation.

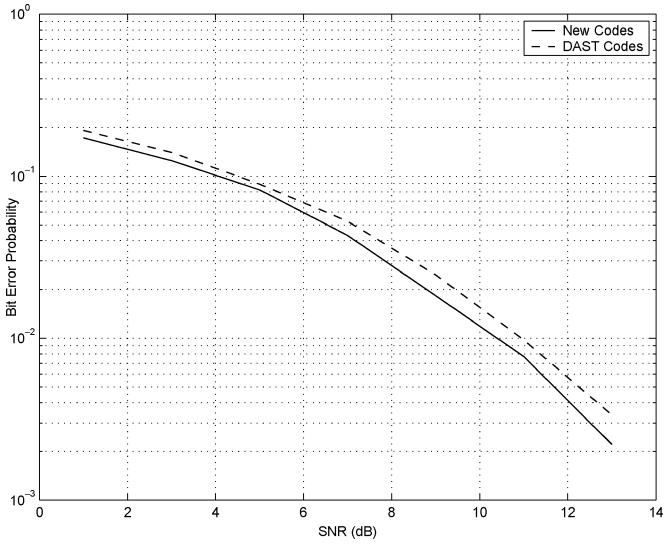

Fig. 4. BER performance of the new LD code and the DAST code in 4×1 uncorrelated MIMO channel with QPSK constellation.

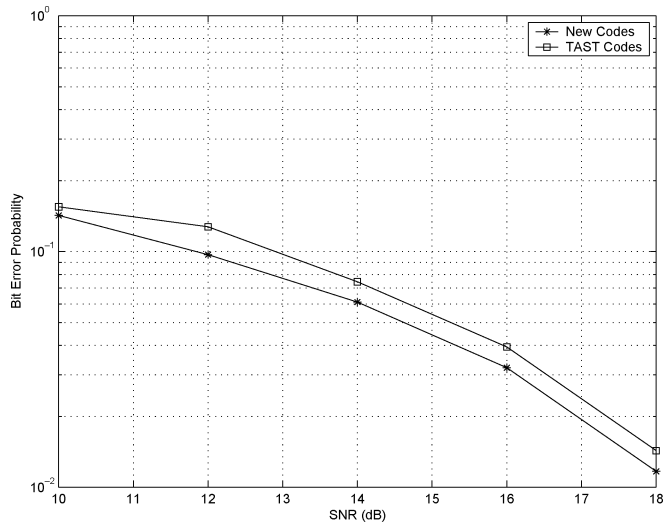

Fig. 3. BER performance of the new LD code and the TAST code in 3×2 uncorrelated MIMO channel with 16QAM constellation.

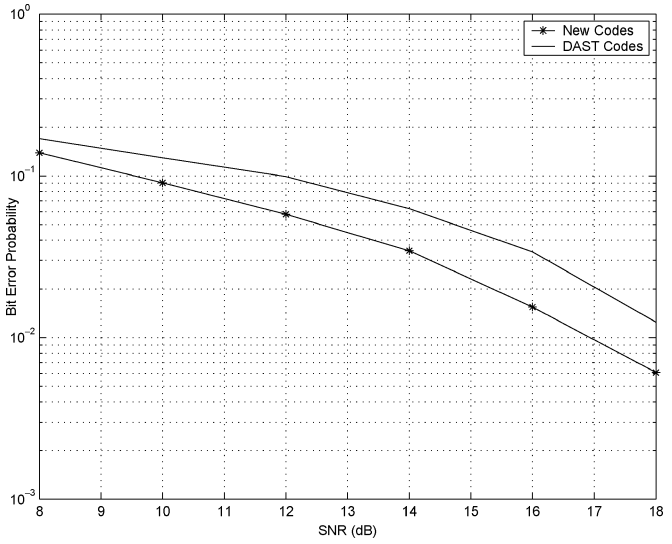

Fig. 5. BER performance of the new LD code and the DAST code in 4×1 uncorrelated MIMO channel with 16QAM constellation.

for a system employing four transmit antennas and one receive antenna, with QPSK and 16QAM, respectively. The gain of the new LD codes over DAST codes is more pronounced for the second case, which is more than 1 dB. In Fig. 6 we show the performance of the new LD code for the case of four transmit antennas and two receive antennas with QPSK constellations. The rate R is 4 b/s/Hz for this case. Again the new code performs uniformly better (though slightly) than the TAST code. In the same figure we also plot the performance of a new LD code designed by choosing $\tau = 2$. Note that by reducing τ from 4 to 2, we only need to jointly decode four instead of eight QPSK symbols, resulting in reduced detection complexity. The performance loss is less than 1 dB at the BLER of 10^{-2}.

Example 2—New LD Codes With Nulling-and-Cancellation Detector: In this example, we consider designing LD codes for suboptimal detectors, in particular, the zero-forcing nulling-and-cancellation detector. We also assume i.i.d. fading channels. In Fig. 7 we present the BER performance of the new code optimized for four transmit antennas and one receive antenna with QPSK constellation. We also show the performance of the DAST code and a randomly generated LD codes. The gain of the new LD code can be clearly seen. Fig. 8 shows the performance of the new LD code for three transmit antennas and two receive antennas with 16QAM or QPSK constellations. The rate R is kept to be 4 b/s/Hz. Therefore, for 16QAM constellation only one symbol is transmitted per channel use, whereas for QPSK constellation two symbols are transmitted per channel use. We also show the performance of the TAST code with the same detector as a reference. It is seen that using a larger constellation leads to better performance. The results in Fig. 8 verify the fact, as pointed out in [22], that reducing the transmission rate, in terms of number of symbols per channel use, will lead

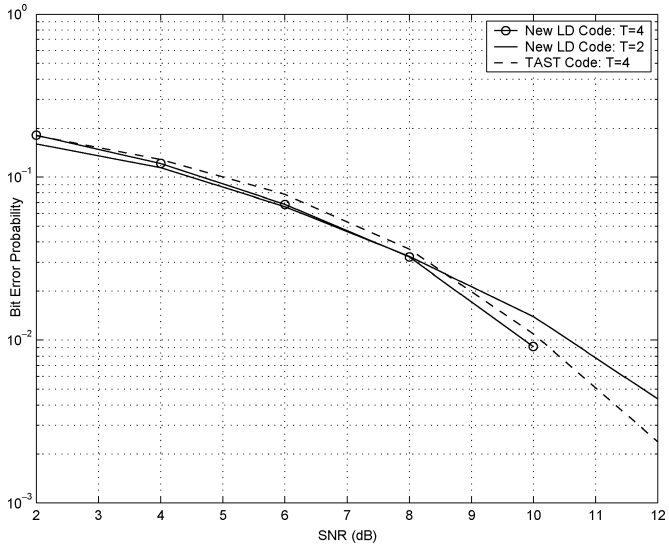

Fig. 6. BER performance of the new LD code and the TAST code in 4 × 2 uncorrelated MIMO channel with QPSK constellation.

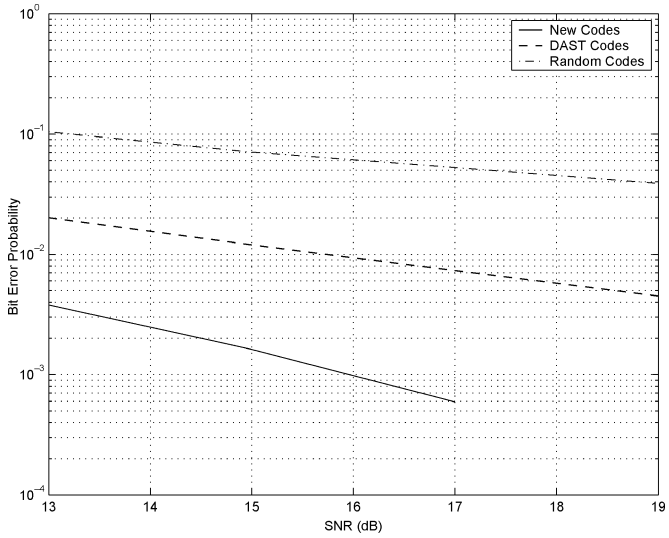

Fig. 7. BER performance of the new LD code and the DAST code in 4 × 1 uncorrelated MIMO channel with QPSK constellation, nulling-and-cancellation detector.

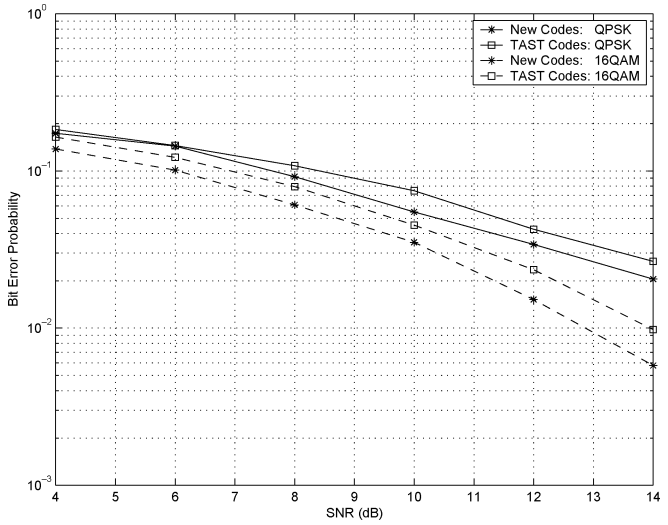

Fig. 8. BER performance of the new LD code and the TAST code in 3 × 2 uncorrelated MIMO channel with QPSK or 16QAM constellation, nulling-and-cancellation detector.

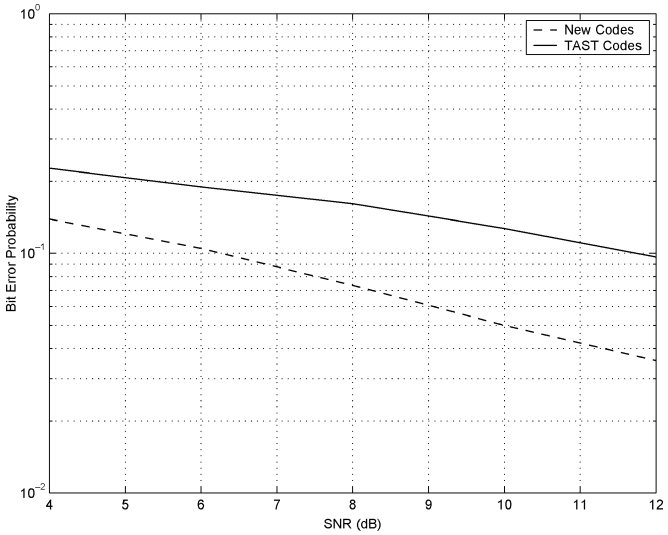

Fig. 9. BER performance of the new LD code and the TAST code in 2 × 2 correlated MIMO channel with QPSK constellation.

to improved performance using the nulling-and-cancellation detector.

Example 3—New LD Codes for Spatially Correlated Rayleigh Fading Channels: Finally we give an example of the optimal LD codes for spatially correlated Rayleigh fading channels. We consider a 2×2 MIMO channel with spatial correlation at both the transmitter and the receiver. The correlation matrices are given by

$$\boldsymbol{S} = \boldsymbol{R} = \begin{bmatrix} 1 & 0.7+0.7j \\ 0.7-0.7j & 1 \end{bmatrix}. \tag{53}$$

We assume QPSK constellation is employed so the data rate is $R = 4$ b/s/Hz. Fig. 9 shows the BER performance of the LD code obtained by Algorithm 2 over this correlated MIMO channel. In the same figure we also show the performance of the TAST code. Note that although for 2×2 i.i.d. fading channels using QPSK constellation, the LD codes obtained by Algorithm 2 perform roughly the same as the TAST code; however, in the presence of spatial correlation, the new LD codes outperforms the TAST code considerably.

V. Conclusion

In this paper, we have proposed a simulation-based optimization approach for the design of optimal (i.e., minimum error rate) LD codes—the resulting algorithm uses stochastic approximation and simulation-based gradient estimation.

The proposed algorithm (Algorithm 2) turns out to be a universal algorithm in the sense that it can be applied to a wide range of detector structures in either i.i.d. fading or spatially correlated fading wireless MIMO channels, with arbitrary fading distributions. Simulation results show that codes generated by the new algorithm generally outperform the codes designed based on algebraic number theory. We have also showed that the amount of improvement obtained by the new

codes depends on the scenario, in particular, the number of transmit and receive antennas, the symbol constellation, the detection algorithm, and the availability of the knowledge on the spatial channel correlation structure.

APPENDIX I
CALCULATION OF $\nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{h},\boldsymbol{\theta})$

Here, we explicitly compute $\nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{h},\boldsymbol{\theta})$ which is required in the simulation-based gradient estimation Algorithm 1. Note from (25) we simply need to compute the gradient of the following function:

$$f \triangleq -(\boldsymbol{y} - \sqrt{\frac{\rho}{M_T}}\boldsymbol{\mathcal{H}}\boldsymbol{x})^T(\boldsymbol{y} - \sqrt{\frac{\rho}{M_T}}\boldsymbol{\mathcal{H}}\boldsymbol{x}). \tag{54}$$

We first compute the gradient of f with respect to $\boldsymbol{A}_{R,q}$. The gradient with respect to $\boldsymbol{A}_{I,q}$, $\boldsymbol{B}_{R,q}$, and $\boldsymbol{B}_{I,q}$ follow similar arguments and are given at the end of this section. The (n,l)th entry of the gradient of $f(\boldsymbol{A}_{R,q})$ is

$$\left[\frac{\partial f(\boldsymbol{A}_{R,q})}{\partial \boldsymbol{A}_{R,q}}\right]_{n,l} = \lim_{\delta\to 0} \frac{f\left(\boldsymbol{A}_{R,q} + \delta\boldsymbol{\varsigma}_n\boldsymbol{\eta}_l^T\right) - f(\boldsymbol{A}_{R,q})}{\delta} \tag{55}$$

where $\boldsymbol{\varsigma}_n$ and $\boldsymbol{\beta}_l$ are τ-dimensional and M_T-dimensional unit column vectors with one in the nth and lth entries, respectively, and zeros elsewhere. From (8), we have

$$\boldsymbol{\mathcal{H}}_{\boldsymbol{A}_{R,q}+\delta\boldsymbol{\varsigma}_n\boldsymbol{\eta}_l^T} = \boldsymbol{\mathcal{H}} + \delta\Xi_{n,l}^{\boldsymbol{A}_{R,q}} \tag{56}$$

where

$$\Xi_{n,l}^{\boldsymbol{A}_{R,q}} = \begin{bmatrix} 0 & 0 & \dots & \Lambda_{n,l}^{\boldsymbol{A}_{R,q}}\underline{\boldsymbol{h}}_1 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & \Lambda_{n,l}^{\boldsymbol{A}_{R,q}}\underline{\boldsymbol{h}}_{M_R} & 0 & \dots & 0 & 0 \end{bmatrix} \tag{57}$$

with

$$\Lambda_{n,l}^{\boldsymbol{A}_{R,q}} = \begin{bmatrix} \boldsymbol{\varsigma}_n\boldsymbol{\eta}_l^T & 0 \\ 0 & \boldsymbol{\varsigma}_n\boldsymbol{\eta}_l^T \end{bmatrix}. \tag{58}$$

We obtain

$$\begin{aligned} &f\left(\boldsymbol{A}_{R,q}+\delta\boldsymbol{\varsigma}_n\boldsymbol{\eta}_l^T\right) \\ &= -\left(\boldsymbol{y}-\sqrt{\frac{\rho}{M_T}}\boldsymbol{\mathcal{H}}_{\boldsymbol{A}_{R,q}+\delta\boldsymbol{\varsigma}_n\boldsymbol{\eta}_l^T}\boldsymbol{x}\right)^T\left(\boldsymbol{y}-\sqrt{\frac{\rho}{M_T}}\boldsymbol{\mathcal{H}}_{\boldsymbol{A}_{R,q}+\delta\boldsymbol{\varsigma}_n\boldsymbol{\eta}_l^T}\boldsymbol{x}\right) \\ &= -\left(\boldsymbol{y} - \sqrt{\frac{\rho}{M_T}}\boldsymbol{\mathcal{H}}\boldsymbol{x} - \sqrt{\frac{\rho}{M_T}}\delta\Xi_{n,l}^{\boldsymbol{A}_{R,q}}\boldsymbol{x}\right)^T \\ &\quad\times\left(\boldsymbol{y} - \sqrt{\frac{\rho}{M_T}}\boldsymbol{\mathcal{H}}\boldsymbol{x} - \sqrt{\frac{\rho}{M_T}}\delta\Xi_{n,l}^{\boldsymbol{A}_{R,q}}\boldsymbol{x}\right) \\ &= f + \sqrt{\frac{\rho}{M_T}}\delta\boldsymbol{x}^T\left(\Xi_{n,l}^{\boldsymbol{A}_{R,q}}\right)^T\left(\boldsymbol{y} - \sqrt{\frac{\rho}{M_T}}\boldsymbol{\mathcal{H}}\boldsymbol{x}\right) \\ &\quad + \sqrt{\frac{\rho}{M_T}}\delta\left(\boldsymbol{y} - \sqrt{\frac{\rho}{M_T}}\boldsymbol{\mathcal{H}}\boldsymbol{x}\right)^T\Xi_{n,l}^{\boldsymbol{A}_{R,q}}\boldsymbol{x} + o(\delta). \end{aligned} \tag{59}$$

Therefore, we have

$$\begin{aligned} \left[\frac{\partial f(\boldsymbol{A}_{R,q})}{\partial \boldsymbol{A}_{R,q}}\right]_{n,l} &= \sqrt{\frac{\rho}{M_T}}\boldsymbol{x}^T\left(\Xi_{n,l}^{\boldsymbol{A}_{R,q}}\right)^T\left(\boldsymbol{y} - \sqrt{\frac{\rho}{M_T}}\boldsymbol{\mathcal{H}}\boldsymbol{x}\right) \\ &\quad + \sqrt{\frac{\rho}{M_T}}\left(\boldsymbol{y} - \sqrt{\frac{\rho}{M_T}}\boldsymbol{\mathcal{H}}\boldsymbol{x}\right)^T\Xi_{n,l}^{\boldsymbol{A}_{R,q}}\boldsymbol{x}. \end{aligned} \tag{60}$$

For the gradients with respect to $\boldsymbol{A}_{I,q}$, $\boldsymbol{B}_{R,q}$, and $\boldsymbol{B}_{I,q}$, similar expressions can be given as

$$\begin{aligned} \left[\frac{\partial f(\boldsymbol{A}_{I,q})}{\partial \boldsymbol{A}_{I,q}}\right]_{n,l} &= \sqrt{\frac{\rho}{M_T}}\boldsymbol{x}^T\left(\Xi_{n,l}^{\boldsymbol{A}_{I,q}}\right)^T\left(\boldsymbol{y} - \sqrt{\frac{\rho}{M_T}}\boldsymbol{\mathcal{H}}\boldsymbol{x}\right) \\ &\quad + \sqrt{\frac{\rho}{M_T}}\left(\boldsymbol{y} - \sqrt{\frac{\rho}{M_T}}\boldsymbol{\mathcal{H}}\boldsymbol{x}\right)^T\Xi_{n,l}^{\boldsymbol{A}_{I,q}}\boldsymbol{x} \end{aligned} \tag{61}$$

$$\begin{aligned} \left[\frac{\partial f(\boldsymbol{B}_{R,q})}{\partial \boldsymbol{B}_{R,q}}\right]_{n,l} &= \sqrt{\frac{\rho}{M_T}}\boldsymbol{x}^T\left(\Xi_{n,l}^{\boldsymbol{B}_{R,q}}\right)^T\left(\boldsymbol{y} - \sqrt{\frac{\rho}{M_T}}\boldsymbol{\mathcal{H}}\boldsymbol{x}\right) \\ &\quad + \sqrt{\frac{\rho}{M_T}}\left(\boldsymbol{y} - \sqrt{\frac{\rho}{M_T}}\boldsymbol{\mathcal{H}}\boldsymbol{x}\right)^T\Xi_{n,l}^{\boldsymbol{B}_{R,q}}\boldsymbol{x} \end{aligned} \tag{62}$$

$$\begin{aligned} \left[\frac{\partial f(\boldsymbol{B}_{I,q})}{\partial \boldsymbol{B}_{I,q}}\right]_{n,l} &= \sqrt{\frac{\rho}{M_T}}\boldsymbol{x}^T\left(\Xi_{n,l}^{\boldsymbol{B}_{R,q}}\right)^T\left(\boldsymbol{y} - \sqrt{\frac{\rho}{M_T}}\boldsymbol{\mathcal{H}}\boldsymbol{x}\right) \\ &\quad + \sqrt{\frac{\rho}{M_T}}\left(\boldsymbol{y} - \sqrt{\frac{\rho}{M_T}}\boldsymbol{\mathcal{H}}\boldsymbol{x}\right)^T\Xi_{n,l}^{\boldsymbol{B}_{R,q}}\boldsymbol{x} \end{aligned} \tag{63}$$

where

$$\Xi_{n,l}^{\boldsymbol{A}_{I,q}} = \begin{bmatrix} 0 & 0 & \dots & \Lambda_{n,l}^{\boldsymbol{A}_{I,q}}\underline{\boldsymbol{h}}_1 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & \Lambda_{n,l}^{\boldsymbol{A}_{I,q}}\underline{\boldsymbol{h}}_{M_R} & 0 & \dots & 0 & 0 \end{bmatrix} \tag{64}$$

with

$$\Lambda_{n,l}^{\boldsymbol{A}_{I,q}} = \begin{bmatrix} 0 & -\boldsymbol{\varsigma}_n\boldsymbol{\eta}_l^T \\ \boldsymbol{\varsigma}_n\boldsymbol{\eta}_l^T & 0 \end{bmatrix} \tag{65}$$

$$\Xi_{n,l}^{\boldsymbol{B}_{R,q}} = \begin{bmatrix} 0 & 0 & \dots & 0 & \Lambda_{n,l}^{\boldsymbol{B}_{R,q}}\underline{\boldsymbol{h}}_1 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & \Lambda_{n,l}^{\boldsymbol{B}_{R,q}}\underline{\boldsymbol{h}}_{M_R} & \dots & 0 & 0 \end{bmatrix} \tag{66}$$

with

$$\Lambda_{n,l}^{\boldsymbol{B}_{R,q}} = \begin{bmatrix} 0 & -\boldsymbol{\varsigma}_n\boldsymbol{\eta}_l^T \\ \boldsymbol{\varsigma}_n\boldsymbol{\eta}_l^T & 0 \end{bmatrix} \tag{67}$$

$$\Xi_{n,l}^{\boldsymbol{B}_{I,q}} = \begin{bmatrix} 0 & 0 & \dots & 0 & \Lambda_{n,l}^{\boldsymbol{B}_{I,q}}\underline{\boldsymbol{h}}_1 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & \Lambda_{n,l}^{\boldsymbol{B}_{I,q}}\underline{\boldsymbol{h}}_{M_R} & \dots & 0 & 0 \end{bmatrix} \tag{68}$$

with

$$\Lambda_{n,l}^{\boldsymbol{B}_{I,q}} = \begin{bmatrix} -\boldsymbol{\varsigma}_n\boldsymbol{\eta}_l^T & 0 \\ 0 & -\boldsymbol{\varsigma}_n\boldsymbol{\eta}_l^T \end{bmatrix}. \tag{69}$$

APPENDIX II
PROOF OF PROPOSITION 1

Denote the set of possible symbol block vector as $\mathcal{A}^{2Q} = \{\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_{|\mathcal{A}|^{2Q}}\}$. Since all block vectors are equiprobable, it is sufficient to show that

$$\sum_i \int \nabla_{\boldsymbol{\theta}} \left[1 - \gamma_{\mathrm{BL}}(\boldsymbol{y}, \boldsymbol{s}_i, \boldsymbol{h}, \boldsymbol{\theta})\right] p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{h}, \boldsymbol{\theta}) d\boldsymbol{y} = 0. \quad (70)$$

By the definition of $\gamma_{\mathrm{BL}}(\cdot)$ in (10), the signal model in (7), and the ML detection rule, we have

$$\begin{aligned} 1 - \gamma_{\mathrm{BL}}(\boldsymbol{y}, \boldsymbol{s}_i, \boldsymbol{h}, \boldsymbol{\theta}) &= \mathbb{I}(\hat{\boldsymbol{x}} = \boldsymbol{s}_i | \boldsymbol{y}, \boldsymbol{x} = \boldsymbol{s}_i, \boldsymbol{h}, \boldsymbol{\theta}) \\ &= \prod_{j \neq i} \mathbb{I}\left(\|\boldsymbol{y} - \sqrt{\frac{\rho}{M_T}} \boldsymbol{\mathcal{H}} \boldsymbol{s}_j\|^2 \geq \|\boldsymbol{y} - \sqrt{\frac{\rho}{M_T}} \boldsymbol{\mathcal{H}} \boldsymbol{s}_i\|^2 \right) \\ &= \prod_{j \neq i} \mathbb{I}\left[\underbrace{2\sqrt{\frac{\rho}{M_T}} (\boldsymbol{s}_j - \boldsymbol{s}_i)^T}_{\boldsymbol{a}_{j|i}^T} \boldsymbol{\mathcal{H}}^T \boldsymbol{y} \right. \\ &\qquad \left. \geq \underbrace{\sqrt{\frac{\rho}{M_T}} \left(\boldsymbol{s}_j^T \boldsymbol{\mathcal{H}}^T \boldsymbol{\mathcal{H}} \boldsymbol{s}_j - \boldsymbol{s}_i^T \boldsymbol{\mathcal{H}}^T \boldsymbol{\mathcal{H}} \boldsymbol{s}_i \right)}_{b_{j|i}} \right]. \quad (71) \end{aligned}$$

Using (25) and (32) we have

$$\begin{aligned} &\sum_i \int \nabla_{\boldsymbol{\theta}} \left[1 - \gamma_{\mathrm{BL}}(\boldsymbol{y}, \boldsymbol{s}_i, \boldsymbol{h}, \boldsymbol{\theta})\right] p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{h}, \boldsymbol{\theta}) d\boldsymbol{y} \\ &= -\pi^{-\tau M_R} \int \sum_i \sum_{j \neq i} \left(\boldsymbol{a}_{j|i}^T \nabla_{\boldsymbol{\theta}} \boldsymbol{\mathcal{H}}^T \boldsymbol{y} - \Delta_{\boldsymbol{\theta}} b_{j|i} \right) \\ &\quad \times \left(\|\boldsymbol{y} - \sqrt{\frac{\rho}{M_T}} \boldsymbol{\mathcal{H}} \boldsymbol{s}_j\| - \|\boldsymbol{y} - \sqrt{\frac{\rho}{M_T}} \boldsymbol{\mathcal{H}} \boldsymbol{s}_i\| \right) \\ &\quad \times \prod_{\ell \neq j} \mathbb{I}\left(\|\boldsymbol{y} - \sqrt{\frac{\rho}{M_T}} \boldsymbol{\mathcal{H}} \boldsymbol{s}_\ell\|^2 \geq \|\boldsymbol{y} - \sqrt{\frac{\rho}{M_T}} \boldsymbol{\mathcal{H}} \boldsymbol{s}_i\|^2 \right) \\ &\quad \times \exp\left(-\|\boldsymbol{y} - \sqrt{\frac{\rho}{M_T}} \boldsymbol{\mathcal{H}} \boldsymbol{s}_i\|^2 \right) d\boldsymbol{y}. \quad (72) \end{aligned}$$

Note that since $\boldsymbol{a}_{i|j} = -\boldsymbol{a}_{j|i}$, then in (72) the summand pair corresponding to indexes (i, j) and (j, i) will cancel each other, $\forall i \neq j$. Hence, the final result is zero.

ACKNOWLEDGMENT

The authors would like to thank Dr. N. Prasad for enlightening discussions.

REFERENCES

[1] B. Hassibi and B. Hochwald, "High-rate codes that are linear in space and time," *IEEE Trans. Inf. Theory*, vol. 48, no. 7, pp. 1804–1824, Jul. 2002.

[2] M. Kuhn, I. Hammerstroem, and A. Wittneben, "Linear scalable space-time codes: Trade-off between spatial multiplexing and transmit diversity," in *Proc. IEEE SPAWC*, 2003, pp. 21–25.

[3] M. O. Damen, A. Chkeif, and J. C. Belfiore, "Lattice code decoder for space–time codes," *IEEE Commun. Lett.*, vol. 4, pp. 166–169, May 2000.

[4] B. Dong, X. Wang, and A. Doucet, "A new class of soft MIMO demodulation algorithms," *IEEE Trans. Signal Process.*, vol. 51, no. 11, pp. 2752–2763, Nov. 2003.

[5] G. J. Foschini, G. Golden, R. Valenzuela, and P. Wolniansky, "Simplified processing for high spectral efficiency wireless communication employing multi-element arrays," *IEEE J. Sel. Areas Commun.*, vol. 17, no. 11, pp. 1841–1852, Nov. 1999.

[6] D. Gesbert, "Minimum-error linear receivers for ill-conditioned MIMO channels," in *Proc. IEEE Workshop Signal Processing Advances in Wireless Communications (SPAWC)*, 2003, pp. 462–466.

[7] R. W. Heath. Jr and A. J. Paulraj, "Linear dispersion codes for MIMO systems based on frame theory," *IEEE Trans. Signal Process.*, vol. 50, no. 10, pp. 2429–2441, Oct. 2002.

[8] C. G. Cassandras and S. Lafortune, *Introduction to Discrete Event Systems*. Norwell, MA: Kluwer, 1999.

[9] M. Fu and J. Q. Hu, *Conditional Monte Carlo: Gradient Estimation and Optimization Applications*. Norwell, MA: Kluwer, 1991.

[10] J. C. Spall, *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*. New York: Wiley, 2003.

[11] H. Fang, F. Gong, and M. Qian, "Annealing of iterative stochastic schemes," *SIAM J. Control Optim.*, vol. 35, pp. 1886–1907, 1997.

[12] V. Tarokh, N. Seshadri, and A. R. Calderbank, "Space–time codes for high data rate wireless communication: Performance criterion and code construction," *IEEE Trans. Inf. Theory*, vol. 44, no. 2, pp. 744–765, Mar. 1998.

[13] V. Tarokh, H. Jafarkhani, and A. R. Calderbank, "Space–time block codes from orthogonal designs," *IEEE Trans. Inf. Theory*, vol. 45, no. 5, pp. 1456–1467, Jul. 1999.

[14] C. Chuah, D. Tse, J. M. Kahn, and R. A. Valenzuela, "Capacity scaling in MIMO wireless systems under correlated fading," *IEEE Trans. Inf. Theory*, vol. 48, no. 3, pp. 637–650, Mar. 2002.

[15] D. Shiu, G. J. Foschini, M. J. Gans, and J. M. Kahn, "Fading correlation and its effect on the capacity of multielement antenna systems," *IEEE Trans. Commun.*, vol. 48, no. 3, pp. 502–513, Mar. 2000.

[16] H. E. Gamal and M. O. Damen, "Universal space–time coding," *IEEE Trans. Inf. Theory*, vol. 48, no. 5, pp. 1097–1119, May 2003.

[17] X. Ma and G. B. Giannakis, "Full-diversity full rate complex field space–time coding," *IEEE Trans. Signal Process.*, vol. 51, no. 11, pp. 2917–2930, Nov. 2003.

[18] B. Varadarajan and J. R. Barry, "Optimization of full-rate full diversity linear space–time codes using the union bound," in *Proc. IEEE Information Theory Workshop*, Mar.–Apr. 2003, pp. 210–213.

[19] J. K. Zhang, K. M. Wong, and T. N. Davidson, "Information lossless full rate full diversity cyclotomic linear dispersion codes," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, vol. 4, May 2004, pp. 465–468.

[20] S. Zhou and G. B. Giannakis, "Optimal transmitter eigen-beamforming and space–time block coding based on channel correlations," *IEEE Trans. Inf. Theory*, vol. 49, no. 7, pp. 1673–1690, Jul. 2003.

[21] M. O. Damen, K. Abed-Meraim, and J. C. Belfiore, "Diagonal algebraic space–time block codes," *IEEE Trans. Inf. Theory*, vol. 48, no. 3, Mar. 2002.

[22] M. O. Damen, H. E. Gamal, and N. C. Beaulieu, "Linear threaded algebraic space–time constellations," *IEEE Trans. Inf. Theory*, vol. 49, no. 10, pp. 2372–2388, Oct. 2003.

[23] V. Krishnamuthy and A. Logothetis, "Iterative and recursive estimators for hidden Markov errors-in-variables models," *IEEE Trans. Signal Process.*, vol. 44, no. 3, pp. 629–639, Mar. 1996.

[24] F. Vazquez Abad and V. Krishnamurthy, "Constrained stochastic approximation algorithms for adaptive control of constrained Markov decision processes," in *Proc. 42nd IEEE Conf. Decision Control*, Maui, HI, 2003, pp. 2823–2828.

[25] H. J. Kushner and G. Yin, *Stochastic Approximation and Recursive Algorithms and Applications*. New York: Springer, 2003.

[26] G. Pflug, *Optimization of Stochastic Models: The Interface Between Simulation and Optimization*. Boston, MA: Kluwer, 1996.

[27] B. Varadarajan and J. R. Barry, "Linear space-time codes optimized for successive cancellation decoding," in *Proc. Int. Zurich Seminar on Communications*, 2004, pp. 132–135.

[28] M. K. Varanasi, "Decision feedback multiuser detection: a systematic approach," *IEEE Trans. Inf. Theory*, vol. 45, no. 1, pp. 219–240, Jan. 1999.

Xiaodong Wang (S'90–M'92–SM'04) received the B.S. degree (with the highest honors) in electrical engineering and applied mathematics from Shanghai Jiao Tong University, Shanghai, China, in 1992; the M.S. degree in electrical and computer engineering from Purdue University, West Lafayette, IN, in 1995; and the Ph.D. degree in electrical engineering from Princeton University, Princeton, NJ, in 1998.

From July 1998 to December 2001, he was on the faculty of the Department of Electrical Engineering, Texas A&M University, College Station. In January 2002, he joined the faculty of the Department of Electrical Engineering, Columbia University, New York. Among his publications is a recent book entitled *Wireless Communication Systems: Advanced Techniques for Signal Reception* (Englewood Cliffs, NJ: Prentice-Hall, 2004. His research interests fall in the general areas of computing, signal processing, and communications, and his current research interests include wireless communications, statistical signal processing, and genomic signal processing.

Dr. Wang received the 1999 NSF CAREER Award, and the 2001 IEEE Communications Society and Information Theory Society Joint Paper Award. He currently serves as an Associate Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS, the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, the IEEE TRANSACTIONS ON SIGNAL PROCESSING, and the IEEE TRANSACTIONS ON INFORMATION THEORY.

Vikram Krishnamurthy (S'90–M'91–SM'99–F'05) was born in 1966. He received the Bachelor's degree from the University of Auckland, New Zealand, in 1988 and the Ph.D. degree from the Australian National University, Canberra, in 1992.

Prior to 2002, he was a chaired Professor at the Department of Electrical and Electronic Engineering, University of Melbourne, Australia, where he also served as Deputy Head of the department. Currently, he is a Professor and Canada Research Chair at the Department of Electrical Engineering, University of British Columbia, Vancouver, BC, Canada. He is coeditor of the research monograph *Handbook on Biological Membrane Ion Channels: Dynamics, Structure and Applications* (New York: Springer-Verlag, 2005). His current research interests include ion channels and biological nanotubes, networked sensor scheduling and control, statistical signal processing, Bayesian filtering, and cross-layer optimization of wireless networks.

Dr. Krishnamurthy currently serves on the Signal Processing Theory and Methods (SPTM) Technical Committee of the IEEE Signal Processing Society and the International Federation of Automatic Control (IFAC) Technical Committee on Modeling, Identification and Signal Processing. He served as Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING from 1999 to 2005 and is currently an Associate Editor for the IEEE TRANSACTIONS ON AEROSPACE AND ELECTRONIC SYSTEMS, and *Systems and Control Letters*. In March 2005, he was a Guest Editor of a special issue of IEEE TRANSACTIONS ON NANOBIOSCIENCE on biological nanotubes.

Jibing Wang (S'98–M'03) received the Ph.D. degree in electrical engineering from the University of California, Los Angeles, in 2003.

From September 2003 to September 2004, he was a Research Staff Member at NEC Laboratories America, Inc., Princeton, NJ. Since September 2004, he has been with Qualcomm, Inc., San Diego, CA. His research interest lies in the area of wireless communications, including space–time coding, multiple-input multiple-output (MIMO) OFDM and multiuser communications.

Dr. Wang is a recipient of a Microsoft Research Fellowship.