DYNAMIC NULLING-AND-CANCELLING WITH NEAR-ML PERFORMANCE FOR MIMO COMMUNICATION SYSTEMS

Dominik Seethaler, Harold Artés, and Franz Hlawatsch

Institute of Communications and Radio-Frequency Engineering, Vienna University of Technology
Gusshausstrasse 25/389, A-1040 Vienna, Austria (Europe)
phone: +43 1 58801 38958, fax: +43 1 58801 38999, email: dominik.seethaler@tuwien.ac.at
web: http://www.nt.tuwien.ac.at/dspgroup/time.html

ABSTRACT

The conventional nulling-and-cancelling (NC) detection scheme for MIMO systems uses the layerwise post-detection mean-square errors (MSEs) as reliability measures for layer sorting. These MSEs are *average* measures that do not depend on the received vector. In this paper, we propose the novel *dynamic nulling-and-cancelling* (DNC) detector that performs "dynamic" layer sorting based on the current received vector. Approximate a-posteriori probabilities (constructed by means of a Gaussian approximation for the post-detection interference) are used as measures of layer reliability. This results in an MMSE nulling technique that uses a simple layer-sorting rule with significantly improved performance. Our simulation results show that the DNC scheme can yield near-ML performance for a wide range of system sizes and signal-to-noise ratios.

1. INTRODUCTION

It is well known that the nulling-and-cancelling (NC) detection scheme for MIMO systems (e.g., [1]) cannot exploit all of the available diversity, and thus its performance is inferior to the performance of maximum-likelihood (ML) detection. The NC scheme uses the layerwise post-detection mean-square errors (MSEs) [2] as a reliability criterion for layer sorting. However, these MSEs are just *average* measures that do not depend on the received vector.

Here, we propose the novel *dynamic nulling-and-cancelling* (DNC) detector that performs "dynamic" layer sorting based on the current received vector. At each decoding step, the DNC scheme detects the symbol and layer with maximum approximate a-posteriori probability (APP). The approximate APP is constructed by means of a Gaussian approximation for the post-detection interference; this approach is inspired by [3,4]. The DNC layer-sorting rule, although quite simple, can result in near-ML performance for a wide range of system sizes and signal-to-noise ratios (SNRs). We will here present the DNC scheme in a spatial multiplexing context; however, it can equally well be used for MIMO systems employing linear dispersion codes and for multiuser detection in CDMA systems.

Our paper is organized as follows. In the remainder of this section, we present the system model and briefly review existing detection schemes. In Section 2, we propose and discuss the novel DNC detector. Simulation results are finally presented in Section 3.

1.1. System Model

We consider a MIMO channel with M_{T} transmit antennas and $M_{\mathrm{R}} \geq M_{\mathrm{T}}$ receive antennas (briefly termed an $(M_{\mathrm{T}}, M_{\mathrm{R}})$ channel). We assume a spatial multiplexing system such as V-BLAST [1] where the ith data symbol (or *layer*) d_i is directly transmitted on the ith transmit antenna. For any given time instant, this leads to the well-known baseband model

$$\mathbf{r} = \mathbf{H}\mathbf{d} + \mathbf{w}, \tag{1}$$

with the transmitted data vector $\mathbf{d} \stackrel{\triangle}{=} (d_1 \cdots d_{M_{\mathrm{T}}})^T$, the $M_{\mathrm{R}} \times M_{\mathrm{T}}$ channel matrix $\mathbf{H}$, the received vector $\mathbf{r} \stackrel{\triangle}{=} (r_1 \cdots r_{M_{\mathrm{R}}})^T$, and the noise vector $\mathbf{w} \stackrel{\triangle}{=} (w_1 \cdots w_{M_{\mathrm{R}}})^T$. The data components d_i are drawn from a complex symbol alphabet $\mathscr{A}$ and are assumed zero-mean and independent with unit variance. The noise components w_i are assumed independent and circularly symmetric complex Gaussian with variance σ_w^2. The channel $\mathbf{H}$ is considered constant over a block of N time instants and perfectly known at the receiver.

Funding by FWF grant P15156-N02.

1.2. Review of Detection Schemes

As a background and for later reference, we briefly review major detection schemes for spatial multiplexing systems.

LINEAR SCHEMES. In linear equalization based schemes, the detected data vector is $\hat{\mathbf{d}} = Q\{\mathbf{y}\}$ with $\mathbf{y} = \mathbf{G}\mathbf{r}$, where $\mathbf{G}$ is the equalizer matrix and $Q\{\cdot\}$ denotes componentwise quantization according to

$$\hat{d}_i = \arg\min_{a \in \mathscr{A}} |y_i - a|^2. \tag{2}$$

The *zero-forcing* (ZF) equalizer is given by the pseudo-inverse [5] of $\mathbf{H}$. Thus, the result of ZF equalization (before quantization) is

$$\mathbf{y}_{\mathrm{ZF}} = \mathbf{H}^{\#}\mathbf{r} = (\mathbf{H}^H\mathbf{H})^{-1}\mathbf{H}^H\mathbf{r} = \mathbf{d} + \tilde{\mathbf{w}}, \tag{3}$$

which is the data vector $\mathbf{d}$ plus the transformed noise $\tilde{\mathbf{w}} = \mathbf{H}^{\#}\mathbf{w}$ with covariance matrix $\mathbf{R}_{\tilde{\mathbf{w}}} = \sigma_w^2(\mathbf{H}^H\mathbf{H})^{-1}$. The *minimum mean-square error* (MMSE) equalizer [6] minimizes the MSE $\mathrm{E}\{\|\mathbf{y} - \mathbf{d}\|^2\}$ and is given by $\mathbf{G}_{\mathrm{MMSE}} = (\mathbf{H}^H\mathbf{H} + \sigma_w^2\mathbf{I})^{-1}\mathbf{H}^H$, so that

$$\mathbf{y}_{\mathrm{MMSE}} = (\mathbf{H}^H\mathbf{H} + \sigma_w^2\mathbf{I})^{-1}\mathbf{H}^H\mathbf{r}. \tag{4}$$

NULLING-AND-CANCELLING. In contrast to linear detection, where all layers are detected jointly, NC uses a serial decision-feedback approach to detect each layer separately (e.g., [1]). At each decoding step, a single layer is detected and the corresponding contribution to the received vector $\mathbf{r}$ is then subtracted from $\mathbf{r}$; the other layers that have not yet been detected are "nulled out" (equalized) using a ZF or MMSE equalizer. NC thus attempts to progressively clean $\mathbf{r}$ from the interference corresponding to the layers already detected. To minimize error propagation effects, more reliable layers should be detected first. Commonly, the layerwise post-detection MSEs are used as measures of layer reliability [2]. The resulting performance is however still significantly inferior to that of ML detection (see Section 3).

OPTIMUM DETECTION. ML detection [7, 8] yields minimum vector error probability for equally likely data vectors. For our system model (1) and our assumptions, the ML detector is given by

$$\hat{\mathbf{d}}_{\mathrm{ML}} = \arg\min_{\mathbf{a} \in \mathscr{A}^{M_{\mathrm{T}}}} \|\mathbf{r} - \mathbf{H}\mathbf{a}\|^2.$$

The computational complexity of ML detection grows exponentially with M_{T}. The Fincke-Phost sphere-decoding algorithm for ML detection [8] has an *average* complexity of roughly $\mathscr{O}(M_{\mathrm{T}}^3)$ [9].

2. DYNAMIC NULLING-AND-CANCELLING

The novel *dynamic nulling-and-cancelling* (DNC) scheme uses approximate a-posteriori probabilities (APPs) to indicate layer reliability. Since the APPs depend on the current received vector, this results in *dynamic* layer sorting in contrast to the static (average) layer sorting of conventional NC. The approximate APPs are based on a Gaussian approximation for the post-detection interference. This is inspired by [3,4]; in particular, it is shown in [3] that for CDMA systems the post-detection multiple-access interference obtained with a linear filter becomes Gaussian for large system sizes. (However, we will see in Section 3 that DNC can yield excellent performance already for MIMO systems of moderate size.) We will discuss DNC just for the *first* layer detection step, without considering the subsequent interference cancellation step (see [1] for a discussion of interference cancellation). The subsequent layer detection steps are analogous, however with a reduced number of active layers.

2.1. Basic Approach

For the ith layer ($i \in \{1,\ldots,M_{\mathrm{T}}\}$), the optimum decision on the data symbol $d_i \in \mathscr{A}$ is given by the maximum a-posteriori (MAP) rule that maximizes the APP[1] $\mathrm{P}\{d_i{=}a|\mathbf{y}_{\mathrm{ZF}}\}$ [7]:

$$\hat{d}_i \stackrel{\triangle}{=} \arg\max_{a\in\mathscr{A}} \mathrm{P}\{d_i{=}a|\mathbf{y}_{\mathrm{ZF}}\}. \tag{5}$$

The resulting maximum APP $\mathrm{P}\{d_i{=}\hat{d}_i|\mathbf{y}_{\mathrm{ZF}}\}$ is a measure of the reliability of this optimum symbol decision. Our approach to layer sorting now is to first calculate the optimum symbol $\hat{d}_i$ for each layer i and then choose the layer $\hat{\imath}$ for which the reliability of this optimum symbol decision (APP for $d_i = \hat{d}_i$) is maximum:

$$\hat{\imath} \stackrel{\triangle}{=} \arg\max_{i\in\{1,\ldots,M_{\mathrm{T}}\}} \mathrm{P}\{d_i{=}\hat{d}_i|\mathbf{y}_{\mathrm{ZF}}\}. \tag{6}$$

Finally, layer $\hat{\imath}$ is decoded in favor of $\hat{d}_{\hat{\imath}}$; this result is subsequently used for interference cancellation.

Unfortunately, the computational complexity of APP calculation is exponential in M_{T}. Assuming that all data symbols are transmitted equally likely, i.e., $\mathrm{P}\{d_i{=}a\} = 1/|\mathscr{A}|$ for all $a \in \mathscr{A}$, the APP can be rewritten as

$$\mathrm{P}\{d_i{=}a|\mathbf{y}_{\mathrm{ZF}}\} = \frac{f(\mathbf{y}_{\mathrm{ZF}}|d_i{=}a)}{\sum_{a'\in\mathscr{A}} f(\mathbf{y}_{\mathrm{ZF}}|d_i{=}a')}, \tag{7}$$

where Bayes' rule has been used. Because of the Gaussianity of the noise, the conditional probability density functions (pdf's) $f(\mathbf{y}_{\mathrm{ZF}}|d_i{=}a)$ are multivariate multimodal Gaussian mixture pdf's. We now use a Gaussian approximation for the post-detection interference to obtain a computationally efficient approximation to (7). More specifically, we will approximate the Gaussian mixture pdf $f(\mathbf{y}_{\mathrm{ZF}}|d_i{=}a)$ by a Gaussian pdf $\tilde{f}(\mathbf{y}_{\mathrm{ZF}}|d_i{=}a)$ with mean $\boldsymbol{\mu}_i \stackrel{\triangle}{=} \mathrm{E}\{\mathbf{y}_{\mathrm{ZF}}|d_i{=}a\}$ and covariance $\mathbf{C}_i \stackrel{\triangle}{=} \mathrm{cov}\{\mathbf{y}_{\mathrm{ZF}}|d_i{=}a\}$. To find expressions of $\boldsymbol{\mu}_i$ and $\mathbf{C}_i$, we reformulate $\mathbf{y}_{\mathrm{ZF}} = \mathbf{d} + \tilde{\mathbf{w}}$ in (3) as

$$\mathbf{y}_{\mathrm{ZF}} = d_i\mathbf{e}_i + \sum_{\substack{j=1\\ j\neq i}}^{M_{\mathrm{T}}} d_j\mathbf{e}_j + \tilde{\mathbf{w}},$$

where $\mathbf{e}_i$ is the ith M_{T}-dimensional unit vector. We then obtain

$$\boldsymbol{\mu}_i = a\mathbf{e}_i, \qquad \mathbf{C}_i = \mathbf{I}_i + \mathbf{R}_{\tilde{\mathbf{w}}}, \tag{8}$$

in which $\mathbf{I}_i$ denotes the identity matrix of size M_{T} with the ith diagonal element replaced by zero. The Gaussian pdf $\tilde{f}(\mathbf{y}_{\mathrm{ZF}}|d_i{=}a)$ is now completely determined, and the APP in (7) is approximated by

$$\mathrm{P}\{d_i{=}a|\mathbf{y}_{\mathrm{ZF}}\} \approx \frac{\tilde{f}(\mathbf{y}_{\mathrm{ZF}}|d_i{=}a)}{\sum_{a'\in\mathscr{A}} \tilde{f}(\mathbf{y}_{\mathrm{ZF}}|d_i{=}a')}. \tag{9}$$

[1] Note that the APP can equivalently be conditioned on the result of ZF equalization $\mathbf{y}_{\mathrm{ZF}}$ rather than on the received vector $\mathbf{r}$ since ZF equalization without quantization does not imply any loss of information.

Using this Gaussian approximation, the maximization in (5) that yields the optimum symbol for the ith layer becomes

$$\hat{d}_i \approx \arg\max_{a\in\mathscr{A}} \tilde{f}(\mathbf{y}_{\mathrm{ZF}}|d_i{=}a), \quad i = 1,\ldots,M_{\mathrm{T}}. \tag{10}$$

Here we have used the fact that the denominator in (9) is nonnegative and independent of a. Furthermore, the maximization in (6) that is used to determine the optimum layer becomes

$$\hat{\imath} \approx \arg\max_{i\in\{1,\ldots,M_{\mathrm{T}}\}} \left\{ \frac{\tilde{f}(\mathbf{y}_{\mathrm{ZF}}|d_i{=}\hat{d}_i)}{\sum_{a\in\mathscr{A}} \tilde{f}(\mathbf{y}_{\mathrm{ZF}}|d_i{=}a)} \right\}. \tag{11}$$

(Note that whereas the denominator in (7), $\sum_{a\in\mathscr{A}} f(\mathbf{y}_{\mathrm{ZF}}|d_i{=}a) = |\mathscr{A}|\, f(\mathbf{y}_{\mathrm{ZF}})$, is independent of the layer index i, the denominator in (11) depends on i due to (8).) In what follows, $\hat{d}_i$ and $\hat{\imath}$ will be used to denote the right-hand sides in (10) and (11), respectively. We next discuss the computation of $\hat{d}_i$ and $\hat{\imath}$.

2.2. Stage 1: Calculation of $\hat{d}_i$

With (8), the maximization in (10) can be written as[2]

$$\begin{aligned}\hat{d}_i &= \arg\max_{a\in\mathscr{A}} \Big\{ \exp\big(-(\mathbf{y}_{\mathrm{ZF}} - a\mathbf{e}_i)^H \mathbf{C}_i^{-1} (\mathbf{y}_{\mathrm{ZF}} - a\mathbf{e}_i)\big) \Big\} \\ &= \arg\max_{a\in\mathscr{A}} \Big\{ 2\,\mathrm{Re}\big\{\mathbf{y}_{\mathrm{ZF}}^H \mathbf{C}_i^{-1}\mathbf{e}_i a\big\} - |a|^2 \mathbf{e}_i^T \mathbf{C}_i^{-1}\mathbf{e}_i \Big\}. \end{aligned} \tag{12}$$

The matrix inversion lemma [5] applied to $\mathbf{C}_i^{-1} = \big(\mathbf{I}_i + \mathbf{R}_{\tilde{\mathbf{w}}}\big)^{-1}$ yields

$$\mathbf{C}_i^{-1} = \mathbf{W}\left(\mathbf{I} + \frac{\mathbf{e}_i\mathbf{e}_i^T\mathbf{W}}{1 - W_{i,i}}\right), \tag{13}$$

with

$$\mathbf{W} \stackrel{\triangle}{=} (\mathbf{I} + \mathbf{R}_{\tilde{\mathbf{w}}})^{-1} = \big(\mathbf{I} + \sigma_w^2(\mathbf{H}^H\mathbf{H})^{-1}\big)^{-1} \tag{14}$$

(this is termed *Wiener estimator* in [10]) and with $W_{i,i}$ denoting the ith diagonal element of $\mathbf{W}$. The Wiener estimator converts ZF equalization (3) into MMSE equalization (4) [10]:

$$\mathbf{y}_{\mathrm{MMSE}} = \mathbf{W}\mathbf{y}_{\mathrm{ZF}}. \tag{15}$$

Using (13) and (15), (12) can be simplified to

$$\hat{d}_i = \arg\max_{a\in\mathscr{A}} \left\{ \frac{1}{1 - W_{i,i}} \Big(2\,\mathrm{Re}\big\{y_{\mathrm{MMSE},i}^* a\big\} - W_{i,i}|a|^2 \Big) \right\}.$$

It can be shown that $0 \le W_{i,i} < 1$, and thus $1 \le 1/(1 - W_{i,i}) < \infty$. We then have

$$\hat{d}_i = \arg\max_{a\in\mathscr{A}} \Big\{ 2\,\mathrm{Re}\big\{y_{\mathrm{MMSE},i}^* a\big\} - W_{i,i}|a|^2 \Big\} = \arg\min_{a\in\mathscr{A}} \psi_i^2(a), \tag{16}$$

with the "unbiased distance"

$$\psi_i^2(a) \stackrel{\triangle}{=} \left| \frac{y_{\mathrm{MMSE},i}}{W_{i,i}} - a \right|^2. \tag{17}$$

The result in (16) is known as *unbiased MMSE detection* [11]; it will hereafter be denoted as

$$\hat{d}_i = Q_{\mathrm{u}}\{y_{\mathrm{MMSE},i}\}. \tag{18}$$

In (17), the bias after MMSE equalization (defined as $\mathrm{E}\{y_{\mathrm{MMSE},i} - d_i|d_i\}$ [11]) is compensated via the division by $W_{i,i}$, i.e., $\mathrm{E}\big\{\frac{y_{\mathrm{MMSE},i}}{W_{i,i}}|d_i\big\} = d_i$. In general, this results in a slightly reduced error probability of (18) compared to conventional MMSE detection

[2] We assume $\sigma_w^2 \neq 0$ so that $\mathbf{C}_i = \mathbf{I}_i + \mathbf{R}_{\tilde{\mathbf{w}}}$ is nonsingular.

(where $y_{\mathrm{MMSE},i}$ is quantized according to (2)). For constant modulus signalling, i.e., when $|a|$ is equal for all $a \in \mathscr{A}$, unbiased MMSE detection is equivalent to conventional MMSE detection.

Thus, using the Gaussian approximation, MAP symbol detection for each layer is equivalent to unbiased MMSE detection, which is computationally simple.

2.3. Stage 2: Calculation of $\hat{i}$

By using the Gaussian approximation, and by applying (13) and (15), the maximization in (11) can be rewritten as

$$\hat{i} = \underset{i\in\{1,\ldots,M_{\mathrm{T}}\}}{\arg\min} \sum_{\substack{a\in\mathscr{A}\\ a\neq\hat{d}_i}} e^{g(a,i)} \tag{19}$$

with

$$\begin{aligned} g(a,i) &\triangleq \frac{2\mathrm{Re}\{y^*_{\mathrm{MMSE},i}(a-\hat{d}_i)\} - W_{i,i}(|a|^2-|\hat{d}_i|^2)}{1-W_{i,i}} \\ &= -\frac{W_{i,i}}{1-W_{i,i}}\left[\psi_i^2(a)-\psi_i^2(\hat{d}_i)\right]. \end{aligned}$$

Taking the logarithm of the function minimized in (19) and using the max-log approximation (e.g., [12]), we obtain

$$\begin{aligned} \hat{i} &= \underset{i\in\{1,\ldots,M_{\mathrm{T}}\}}{\arg\min} \left\{ \log\left(\sum_{\substack{a\in\mathscr{A}\\ a\neq\hat{d}_i}} e^{g(a,i)} \right) \right\} \\ &\approx \underset{i\in\{1,\ldots,M_{\mathrm{T}}\}}{\arg\min} \left\{ \max_{\substack{a\in\mathscr{A}\\ a\neq\hat{d}_i}} \{g(a,i)\} \right\} \\ &= \underset{i\in\{1,\ldots,M_{\mathrm{T}}\}}{\arg\max} \left\{ \frac{W_{i,i}}{1-W_{i,i}} \min_{\substack{a\in\mathscr{A}\\ a\neq\hat{d}_i}} \{\psi_i^2(a)-\psi_i^2(\hat{d}_i)\} \right\}. \end{aligned} \tag{20}$$

This simplifying approximation will be used in the following; we thus redefine $\hat{i}$ to be given by (20).

For an interpretation of (20), we consider the *MMSE post-detection SNR* of the ith layer defined as (e.g., [13])

$$\mathrm{SNR}_i \triangleq \frac{1}{\mathrm{MSE}_i} - 1, \tag{21}$$

where MSE_i is the minimum MSE of the ith layer (e.g., [2]):

$$\begin{aligned} \mathrm{MSE}_i &\triangleq \mathrm{E}\{|y_{\mathrm{MMSE},i}-d_i|^2\} = \sigma_w^2 \left(\left(\mathbf{H}^H\mathbf{H}+\sigma_w^2\mathbf{I}\right)^{-1} \right)_{i,i} \\ &= \sigma_w^2 \sum_{j=1}^{M_{\mathrm{T}}} \frac{1}{\sigma_j^2+\sigma_w^2} |(\mathbf{v}_j)_i|^2. \end{aligned} \tag{22}$$

Here, the σ_j and $\mathbf{v}_j$ denote, respectively, the eigenvalues and eigenvectors of $\mathbf{H}^H\mathbf{H}$. MSE_i can be related to $W_{i,i}$. From (14),

$$\begin{aligned} W_{i,i} &= \sum_{j=1}^{M_{\mathrm{T}}} \frac{\sigma_j^2}{\sigma_j^2+\sigma_w^2} |(\mathbf{v}_j)_i|^2 \\ &= \sum_{j=1}^{M_{\mathrm{T}}} |(\mathbf{v}_j)_i|^2 - \sigma_w^2 \sum_{j=1}^{M_{\mathrm{T}}} \frac{1}{\sigma_j^2+\sigma_w^2} |(\mathbf{v}_j)_i|^2 \\ &= 1-\mathrm{MSE}_i. \end{aligned}$$

Inserting this in (21) yields SNR_i in terms of $W_{i,i}$:

$$\mathrm{SNR}_i = \frac{W_{i,i}}{1-W_{i,i}}.$$

Thus, (20) can be written as

$$\hat{i} = \underset{i\in\{1,\ldots,M_{\mathrm{T}}\}}{\arg\max} \left\{ \mathrm{SNR}_i \min_{\substack{a\in\mathscr{A}\\ a\neq\hat{d}_i}} \{\psi_i^2(a)-\psi_i^2(\hat{d}_i)\} \right\}. \tag{23}$$

Note that due to (16), $\psi_i^2(a)-\psi_i^2(\hat{d}_i) \geq 0$ for all $a \in \mathscr{A}$.

For a real system with BPSK modulation (note that this is fully equivalent to a complex system with 4-QAM modulation [14]), this simplifies as follows. We have

$$\begin{aligned} \hat{i} &= \underset{i\in\{1,\ldots,M_{\mathrm{T}}\}}{\arg\max} \left\{ \mathrm{SNR}_i \left[\psi_i^2(-\hat{d}_i)-\psi_i^2(\hat{d}_i)\right] \right\} \\ &= \underset{i\in\{1,\ldots,M_{\mathrm{T}}\}}{\arg\max} \left\{ \frac{1}{\mathrm{MSE}_i} y_{\mathrm{MMSE},i}\hat{d}_i \right\}, \end{aligned}$$

where (17) has been used. Now

$$\hat{d}_i = \begin{cases} -1, & \text{for} \quad y_{\mathrm{MMSE},i} < 0 \\ 1, & \text{for} \quad y_{\mathrm{MMSE},i} > 0, \end{cases}$$

and thus we obtain

$$\hat{i} = \underset{i\in\{1,\ldots,M_{\mathrm{T}}\}}{\arg\max} \left\{ \frac{1}{\mathrm{MSE}_i} |y_{\mathrm{MMSE},i}| \right\}. \tag{24}$$

This is seen to be a simple extension of the layer sorting of conventional MMSE-based NC.

2.4. Discussion

The reliability of the detected symbol $\hat{d}_i$ corresponding to the ith layer—i.e., the function maximized in (23)—consists of two factors:

- The first factor, SNR_i, expresses the *average* reliability of the ith layer. This factor depends on the channel $\mathbf{H}$ and noise variance σ_w^2 but not on the current received vector $\mathbf{r}$.
- The second factor, $\min_{\substack{a\in\mathscr{A}\\ a\neq\hat{d}_i}} \{\psi_i^2(a)-\psi_i^2(\hat{d}_i)\}$, expresses the *instantaneous* reliability of the ith layer. This factor depends on the current received vector $\mathbf{r}$ via $y_{\mathrm{MMSE},i}$ (cf. (17)).

With conventional NC, the layers are sorted simply according to maximum SNR_i (equivalently, according to minimum MSE_i, see (21)). With DNC, layer sorting additionally takes into account the instantaneous-reliability factor $\min_{\substack{a\in\mathscr{A}\\ a\neq\hat{d}_i}} \{\psi_i^2(a)-\psi_i^2(\hat{d}_i)\}$. To appreciate the beneficial influence of this factor, suppose that SNR_i is large but, for a specific received vector $\mathbf{r}$, layer i is very unreliable in that $y_{\mathrm{MMSE},i}/W_{i,i}$ (cf. (17)) happens to be close to a boundary of the symbol decision regions. In that case, the unbiased distance for the detected symbol $\hat{d}_i$ will be close to the unbiased distance for some other symbol $a \neq \hat{d}_i$, i.e., $\psi_i^2(\hat{d}_i) \approx \psi_i^2(a)$. It follows that the instantaneous-reliability factor $\min_{\substack{a\in\mathscr{A}\\ a\neq\hat{d}_i}} \{\psi_i^2(a)-\psi_i^2(\hat{d}_i)\}$ is small, regardless of the average-reliability factor SNR_i, and thus DNC correctly treats this layer as unreliable. In contrast, conventional NC would erroneously treat this layer as reliable because of the large SNR_i. This explains the performance advantage of DNC over conventional NC.

2.5. Algorithm Summary

The first decoding step of the proposed DNC algorithm can be summarized a follows.

1. *Preparation*: Calculate $(\mathbf{H}^H\mathbf{H}+\sigma_w^2\mathbf{I})^{-1}$ and $\mathbf{y}_{\mathrm{MMSE}}$; determine SNR_i using (21) and (22).
2. *Stage 1*: Perform unbiased MMSE detection for all layers, i.e., calculate $\hat{d}_i = Q_{\mathrm{u}}\{y_{\mathrm{MMSE},i}\}$ for $i = 1,\ldots,M_{\mathrm{T}}$ according to (16).
3. *Stage 2*: Determine the most reliable layer index $\hat{i}$ according to (23) (or (24) in the case of an equivalent real system model with BSPK modulation).

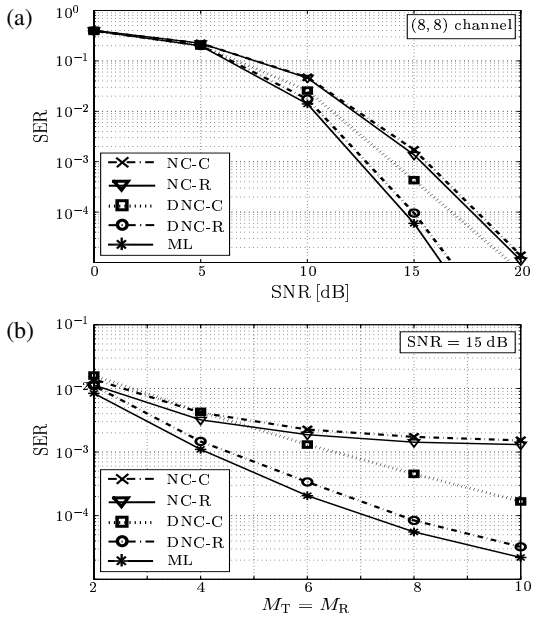

Figure 1: *SER performance of the proposed DNC schemes and of standard detectors for 4-QAM modulation: (a) SER versus SNR for an (8,8) system, (b) SER versus $M_T = M_R$ at an SNR of 15 dB.*

4. *Decoding and interference cancellation*: Finally, decode layer $\hat{i}$ in favor of $\hat{d}_{\hat{i}}$; then use this result for interference cancellation to obtain a reduced system model.

For the next decoding step, this procedure is applied to the reduced (interference-cleaned) system model, etc.

The computational complexity of the DNC detector is *a priori* higher than that of the NC detector. This is because DNC performs layer sorting anew for each received vector $\mathbf{r}$ whereas NC performs layer sorting only once for an entire data block during which the channel is constant. Nevertheless, in a direct implementation of DNC the dependence of the overall computational complexity on M_T is $\mathcal{O}(M_T^4)$ just as for NC (here, we have set $M_R = M_T$ for simplicity). An efficient implementation of DNC whose overall complexity is just $\mathcal{O}(M_T^3)$ will be presented in a future publication. Note that in contrast to DNC, the complexity of sphere-decoding [8] *strongly* depends on the SNR and on the specific channel realization and that it can be much larger than its average $\mathcal{O}(M_T^3)$ complexity [9] .

3. SIMULATION RESULTS

We will now assess the symbol-error rate (SER) performance of the proposed DNC scheme by means of simulation results. In our simulations, we used 4-QAM modulation and MIMO channels with iid Gaussian entries with unit variance. We applied DNC as well as conventional MMSE-based NC (with layer sorting according to minimum post-detection MSEs) to both the complex system model and the equivalent real system model. The corresponding schemes are denoted as DNC-C and DNC-R, respectively for the DNC scheme and NC-C and NC-R, respectively for the NC scheme. We also considered ML detection.

Fig. 1(a) shows the SER versus the SNR for the DNC-C and DNC-R schemes and the various standard detectors for an $(8,8)$ channel. (The SNR is defined as $\mathrm{E}\{\|\mathbf{Hd}\|^2\}/\mathrm{E}\{\|\mathbf{w}\|^2\} = M_T/\sigma_w^2$.) Fig. 1(b) shows the SER versus the number of transmit and receive antennas $M_T = M_R$ at an SNR of 15 dB. The following conclusions can be drawn from these results.

- DNC-R achieves near-ML performance over a wide range of system sizes and SNRs.
- DNC-C outperforms NC for $M_T = M_R \geq 6$; the performance gain is greater for larger system sizes.
- DNC-R performs significantly better than DNC-C. This is because for the complex system model, DNC may consider a layer as unreliable if just the real (or imaginary) part of that layer is unreliable, although possibly the imaginary (or real) part may be very reliable and should thus be used for interference cancellation. For the real system model, the real and imaginary parts are considered separately, which results in improved performance.
- NC-R performs slightly better than NC-C (cf. [14]).

4. SUMMARY AND CONCLUSIONS

We have presented the novel *dynamic nulling-and-cancelling* (DNC) scheme, which is a nulling-and-cancelling MIMO detection scheme with "dynamic" layer sorting depending on the current received vector. At each decoding step, the DNC scheme detects and cancels the symbol and layer with maximum approximate a-posteriori probability (APP). The approximate APPs are constructed via a Gaussian approximation for the post-detection interference. This results in an MMSE nulling scheme with a novel, simple layer-sorting rule that is superior to the conventional rule based on minimum post-detection MSEs. Our simulation results showed that the proposed DNC scheme can yield near-ML performance for a wide range of system sizes and signal-to-noise ratios. The performance of DNC was seen to be best if the real and imaginary parts can be decoded separately.

REFERENCES

[1] G. D. Golden, C. J. Foschini, R. A. Valenzuela, and P. W. Wolniansky, "Detection algorithm and initial laboratory results using V-BLAST space-time communications architecture," *Elect. Lett.*, vol. 35, pp. 14–16, Jan. 1999.

[2] B. Hassibi, "A fast square-root implementation for BLAST," in *Proc. 34th Asilomar Conf. Signals, Systems, Computers*, (Pacific Grove, CA), pp. 1255–1259, Nov./Dec. 2000.

[3] D. Guo, S. Verdú, and L. K. Rasmussen, "Asymptotic normality of linear multiuser outputs," *IEEE Trans. Inform. Theory*, vol. 48, pp. 3080–3095, Dec. 2002.

[4] J. Luo, K. Pattipati, P. Willett, and F. Hasegawa, "A PDA approach to CDMA multiuser detection," in *Proc. IEEE GLOBECOM-2001*, vol. 2, (San Antonio, TX), pp. 763–766, Nov. 2001.

[5] G. H. Golub and C. F. Van Loan, *Matrix Computations*. Baltimore: Johns Hopkins University Press, 3rd ed., 1996.

[6] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Englewood Cliffs (NJ): Prentice Hall, 1993.

[7] S. M. Kay, *Fundamentals of Statistical Signal Processing: Detection Theory*. Upper Saddle River (NJ): Prentice Hall, 1998.

[8] U. Fincke and M. Phost, "Improved methods for calculating vectors of short length in a lattice, including a complexity analysis," *Math. Comp.*, vol. 44, pp. 463–471, April 1985.

[9] B. Hassibi and H. Vikalo, "On the expected complexity of sphere decoding," in *Proc. 35th Asilomar Conf. Signals, Systems, Computers*, (Pacific Grove, CA), pp. 1051–1055, Nov. 2001.

[10] A. Klein, G. Kaleh, and P. W. Baier, "Zero forcing and minimum mean square error equalization for multiuser detection in code-division multiple-access channels," *IEEE Trans. Veh. Technol.*, vol. 45, pp. 276–287, May 1996.

[11] J. M. Cioffi, G. P. Dudevoir, M. V. Eyuboglu, and G. D. Forney, "MMSE decision-feedback equalizers and coding – Part I: Equalization results," *IEEE Trans. Commun.*, vol. 43, pp. 2582–2594, Oct. 1995.

[12] S. Bäro, J. Hagenauer, and M. Witzke, "Iterative detection of MIMO transmission using a list-sequential LISS detector," in *Proc. IEEE ICC 2003*, (Anchorage, AK), pp. 2653–2657, May 2003.

[13] R. W. Heath, S. Sandhu, and A. J. Paulraj, "Antenna selection for spatial mulitplexing systems with linear receivers," *IEEE Commun. Letters*, vol. 5, pp. 142–144, April 2001.

[14] R. F. H. Fischer and C. Windpassinger, "Real versus complex-valued equalisation in V-BLAST systems," *Elect. Lett.*, vol. 39, pp. 470–471, March 2003.