Achieving Near-Capacity in Multi-Antenna Multi-User Systems

Bertrand M. Hochwald, Christian B. Peel, and A. Lee Swindlehurst

Abstract

Recent theoretical results describing the sum-capacity when using multiple antennas to communicate with multiple users in a rich scattering environment have not yet been followed with practical transmission schemes that achieve this capacity. We introduce a simple encoding algorithm that achieves near-capacity at sum-rates of tens of bits/channel-use. The algorithm is a variation on channel inversion that regularizes the inverse and uses a “sphere encoder” to perturb the data to reduce the energy of the transmitted signal. The performance difference between channel inversion with and without this perturbation is shown to be dramatic. With the perturbation, we can achieve linear growth in the sum-rate with the number of users. The results of both uncoded and turbo-coded simulations are presented.

I. INTRODUCTION

Current information-theoretic interest in MIMO communications has shifted in part away from point-to-point links and into multi-user (or “broadcast”) links. Recent work by Caire and Shamai [1] and others [2]–[4] have shown that many of the advantages of using multiple antennas in a single-user scenario also translate to large gains in multi-user scenarios. We seek a simple scheme to achieve this multi-user gain.

It is well known that the point-to-point capacity of a $M$-transmit, $N$-receive antenna link grows linearly in a Rayleigh fading environment with the minimum of $M$ and $N$ when the receiver knows the channel [5]. It is also shown in [5] that $K$ users, each with one antenna, can transmit to a single receiver with $M$ antennas and the sum-capacity (total of transmission rates to all $K$ users) grows linearly with the minimum of $M$ and $K$. It has been more recently shown that this “uplink” transmission has a symmetric “downlink” where the $M$ antennas are used to transmit to the $K$ users; the sum-capacity grows linearly with $\min(M, K)$, provided the transmitter and receivers all know the channel [2]–[4].

This particular use of multiple antennas to communicate with many users simultaneously is especially appealing in wireless local area network (WLAN) environments such as IEEE 802.11, where channel conditions change slowly and there is sufficient time for all parties to learn their channel conditions. Some multi-antenna multi-user concepts have also been applied to digital subscriber line (DSL) services, where many twisted pairs of telephone lines are bundled together in one cable leading to interference between users. We are interested primarily in designing a coding technique for the downlink, where an access point (or base-station, or telephone switch) with $M$ antennas (or a bundle of $M$ wires) wants to communicate simultaneously with $K$ users.

To date, schemes to achieve the sum-capacity in these multi-antenna links are largely information-theoretic and rely on layered applications of “dirty-paper coding” and interference cancellation. Dirty-paper coding is first described for the Gaussian interference channel by Costa in [6], where he finds that the capacity of an interference channel where the interfering signal is known at the transmitter (but not necessarily under its control) is the same as the channel with no interference. Costa envisioned the interference as dirt and his signal as ink; his information-theoretic solution is not to oppose the dirt,
but to use a code that aligns itself as much as possible with the dirt. Costa builds on work of Gelfand and Pinsker [7] for the case where channel side-information is known non-causally at the transmitter.

Several researchers have investigated practical techniques to achieve the sum-capacity promised by dirty-paper coding. Nested lattices are used in [8] for the interference channel, as well as the general multi-user channel. Trellis coding for the broadcast channel is presented in [9], [10] as a practical technique for the multi-user channel. These techniques are generally in preliminary states of development.

Channel inversion is one of the simplest modulation techniques for the multi-user channel [11]. This technique multiplies the vector-signal to be transmitted by the inverse of the channel matrix; the result is an “equalized” channel to each user. We first show that the sum-rate for channel inversion (sometimes also referred to as “zero-forcing beamforming” [1]) in its raw form is poor. We develop a regularized form of inversion which overcomes the noise amplification problem of channel inversion at low SNR. We find the regularization parameter that maximizes the signal-to-interference-plus-noise ratio (SINR) at each receiver. However, regularization by itself still leaves a substantial gap to capacity at high SNR.

An important technique originally developed for the inter-symbol interference channel is Tomlinson-Harashima precoding [12], [13]. This technique applies a scalar integer offset at the transmitter that allows cancellation of the interference after application of a modulo function at the receiver.

We show that a technique related to both Tomlinson-Harashima precoding and channel inversion can achieve near-sum-capacity even at high SNR, with each user receiving $1/K$th of the sum-capacity. Our technique does not require explicit dirty paper techniques. In fact, while the technique requires the transmitter to know the channel, each receiver needs to know only a single scalar related to the SNR of the channel. We show that a properly chosen vector integer offset of the signal to be transmitted can make channel inversion work surprisingly well. In general, techniques such as the Fincke-Pohst algorithm [14], [15] (which in our context we label “sphere encoding”), can aid in selecting the desired vector perturbation. In all cases, however, the processing at the receiver is simple.

We show that in a channel having the same number of antennas as users ($M = K$) our vector perturbation method gives linear growth in the sum-rate with $K$. This is in contrast to plain channel inversion which has constant sum-rate with $K$. The perturbation algorithm has the simple interpretation of placing the largest signal components along the smallest singular values of the inverse channel, and the smallest signal components along the largest singular values of the inverse channel.

II. Model and Brief Synopsis

A. Model

A general model for the forward link of a multi-user system includes an access point with $M$ transmit antennas and $K$ users, each with one receive antenna. The received data at the $k$th user is

$$y_k = \sum_{i=1}^{M} h_{k,i} x_i + w_k ,$$

where $h_{k,i}$ is the zero-mean unit-variance complex-Gaussian fading gain between transmit antenna $i$ and user $k$, $x_i$ is the signal sent from the $i$th antenna, and $w_k$ is standard complex-Gaussian receiver noise seen at the $k$th user. The corresponding vector equation is

$$\mathbf{y} = H \mathbf{x} + \mathbf{w} ,$$

where $\mathbf{y} = [y_1, \ldots, y_K]^T$, with $\mathbf{x} = [x_1, \ldots, x_M]^T$ and $\mathbf{w} = [w_1, \ldots, w_K]^T$, and the $K \times M$ matrix $H$ has $h_{k,i}$ as elements. The power constraint $E \| \mathbf{x} \|^2 = 1$ is imposed, with $E \| \mathbf{w} \|^2 = K \sigma^2$ (we often impose the even stronger constraint $\| \mathbf{x} \|^2 = 1$).
It is often convenient to construct an unnormalized signal $s$, such that

$$x = \frac{s}{\sqrt{\gamma}},$$

(3)

where $\gamma = ||s||^2$. With this normalization, $x$ obeys $||x||^2 = 1$. We can, alternatively, let

$$x = \frac{s}{\sqrt{E \gamma}}.$$ 

(4)

In this case, $E ||x||^2 = 1$. Equation (3) has the advantage that $E \gamma$ does not need to exist (in simple channel inversion, $E \gamma = \infty$), but has the disadvantage that the receivers generally need to know $\gamma$, a data-dependent quantity, to decode their data properly. In the normalization (4), the receiver needs to know only $E \gamma$, which is not data dependent. In our analysis, we generally use the normalization (3), but we mention that our proposed algorithms tend to work equally well whether we use (3) or (4).

We concentrate on the scenario where all $K$ users are serviced at the same data rate $R_k = R$. We assume that $H$ is constant for some interval long enough for the transmitter to learn and use it until it changes to a new value. We are interested in the behavior of the system (2), its capacity, and algorithms to achieve capacity.

An important figure of merit for (2) is the ergodic sum-capacity [2]–[4]

$$C_{\text{sum}} = E \sup_{D \in \mathcal{A}} \log |I_M + \rho H^* DH|,$$ 

(5)

where $I_M$ is the $M \times M$ identity matrix, $\mathcal{A}$ is the set of $K \times K$ diagonal matrices $D \geq 0$ with $\text{tr} D = 1$, and we define $\rho = 1/\sigma^2$. (We measure capacity in bits/channel-use and assume that all logarithms are base-two). Although the total transmitted power is one, the quantity $\rho$ is directly related to, but is not necessarily the same as, the signal-to-noise ratio (SNR) at each receiver. By simply choosing $D = (1/K)I_K$, we can easily infer that $C_{\text{sum}}$ grows linearly with $\min(M, K)$. The expectation in equation (5) assumes that coding is done over multiple intervals with independent $H$. As needed, we compute the maximization in (5) numerically using a gradient-type method, but we omit the details from our discussion.

When $K < M$, the optimization over $D \in \mathcal{A}$ given in (5) gives nonzero energy to all $K$ users when $\rho$ is large enough. This occurs because omitting any user by setting any diagonal entry of $D$ to zero gains signal energy for the remaining users (which has a logarithmic effect) but loses a transmission degree of freedom (which has a more dramatic linear effect). On the other hand, when $K > M$, we know from the formula (5) that transmitting to at least $M$ out of the $K$ users simultaneously uses all of our available degrees of freedom; we may gain by judiciously choosing a subset of fewer than all $K$ users. We do not pursue the choice of subset here; in the interests of fairness to all users, we assume that a random choice of $M$ users is made. In this paper, we therefore generally consider the case $K = M$ to be most important.

We comment that the forward link problem we are considering needs a fundamentally different solution than the reverse link problem. In the reverse link, the $K$ users are transmitting simultaneously to the access point that is now acting as the receiver. The reverse link problem has readily available solutions: It is known that it is optimal for the $K$ users to use independent code books, subject to their own power constraints; the receiver can use many forms of decoding such as successive nulling/canceling or maximum likelihood with reduced complexity (using the sphere decoder [14]). We therefore omit considerations of the reverse link in this paper.

### B. Synopsis

We give a brief summary of our forward-link algorithm, present a performance curve, and leave the details for the remainder of the paper. Let $K$ bit streams be coded with $K$ separate channel (turbo)
codes whose outputs are mapped to symbols to form the vector \( \mathbf{u} = [u_1, \ldots, u_K]^T \). Let \( \alpha \) and \( \tau \) be fixed scalars that we leave unspecified for now. The transmitter solves the cost function
\[
I = \arg \min_I \| H^* (H H^* + \alpha I_K)^{-1} (\mathbf{u} + \tau I) \|^2
\]
for \( I \), and transmits
\[
x = \frac{1}{\sqrt{\gamma}} H^* (H H^* + \alpha I)^{-1} (\mathbf{u} + \tau I),
\]
where \( \gamma \) ensures that the norm of the transmitted signal is one. The role of \( \alpha \) is to help “regularize” the inverse of \( H \), and the vector \( \tau I \) is used to help remove the effect of the components of the data vector \( \mathbf{u} \) that lie along the eigenvectors of \( H^{-1} \) with large eigenvalues. The receivers know \( \tau \), but not \( I \). User \( k \) models its received signal as
\[
y_k = \frac{1}{\sqrt{\gamma}} (u_k + \tau l_k) + w_k',
\]
where the Gaussian \( w_k' \) contains not only the receiver noise \( w_k \), but also the crosstalk from other users introduced by \( \alpha \). Each user then passes this signal through a modulo function that removes the effects of the unknown \( l_k \) and uses a turbo decoder to decode its intended data \( u_k \). The curve closest to the capacity line in Figure 1 shows that this algorithm puts us roughly 4 dB from capacity when we are transmitting 3 bits/user with \( M = K = 10 \). Note that this represents the very high sum-rate of 30 bits/channel-use. The receivers need to know either \( \gamma \) or \( \sqrt{\gamma} \) (depending on the transmitter normalization) but do not need to know \( H \).

The remainder of the papers briefly describes the operating parameters such as \( \tau \) and \( \alpha \).

**C. Channel Inversion when \( K = M \)**

When \( K = M \), channel inversion is simply
\[
\mathbf{s} = H^{-1} \mathbf{u},
\]
This equation can obviously be problematic when $H$ is poorly conditioned, and this problem manifests itself in the normalization constant $(3) \gamma = ||s||^2 = u^*(HH^*)^{-1}u$. Let the entries of $u$ be zero-mean unit-variance independent complex-Gaussian random variables. Then $\gamma$ has density [16] 

$$p(\gamma) = K \frac{\gamma^{K-1}}{(1+\gamma)^{K+1}}. \quad (10)$$

A preview of the poor performance of channel inversion can be gleaned by observing that this density has infinite mean, $\mathbb{E} \gamma = \infty$.

The received data at the $k$th user is

$$y_k = u_k \sqrt{\gamma} + w_k. \quad (11)$$

The receivers all know $\gamma$ and we assume that $K$ is large enough so that any user’s data does not significantly affect the value of $\gamma$. It is shown in [17] that the sum-rate with channel inversion is

$$\lim_{K \to \infty} \frac{C_{\text{CI}}}{K} = \rho \log e \text{ bits per channel use.} \quad (12)$$

The unfortunate conclusion is that the sum-rate for $K = M$ users with channel inversion is constant as a function of $K$, as $K \to \infty$. This is in contrast to (5), which grows linearly with $K$.

The eigenvalues of $(HH^*)^{-1}$ are to blame for this poor capacity. The smallest eigenvalue of $HH^*$ has distribution $p(\lambda) = Ke^{-K\lambda}$, which is an exponential distribution. The largest eigenvalue of $(HH^*)^{-1}$ therefore has the distribution

$$p(\mu) = (\mu^2) e^{-K/\mu}, \quad (13)$$

which is sometimes called the inverse-gamma distribution with parameter one. This density is zero at $\mu = 0$ but decays as $1/\mu^2$ as $\mu \to \infty$ for any $K$. Hence it is a long-tailed distribution with infinite mean. Although the remaining $K-1$ eigenvalues of $(HH^*)^{-1}$ are significantly better behaved, any component of $u$ along the eigenvector corresponding to this large eigenvalue yields a large $\gamma$.

### III. Regularizing the Inverse

One technique often used to “regularize” an inverse is to add a multiple of the identity matrix before inverting. For example, instead of forming $s$ using (9), we use

$$s = H^*(HH^* + \alpha I_N)^{-1}u. \quad (14)$$

After going through the channel, the unnormalized signal $s$ becomes

$$Hs = HH^*(HH^* + \alpha I)^{-1}u. \quad (15)$$

The signal received at user $k$ is no longer simply a scaled version of $u_k$, but also includes some “crosstalk” interference from the remaining users. The amount of interference is determined by $\alpha > 0$; when $\alpha = 0$, we return to (9). It is clear that, no matter how poorly conditioned $H$ is, the inverse in (14) can be made to behave as well as desired by choosing $\alpha$ large enough.

It is shown in [17] that the $k$th user can model its (normalized) received signal as

$$y_k = (1/\sqrt{\gamma}) \left( \sum_{l=1}^{K} \frac{\lambda_l}{\lambda_l + \alpha} |q_{kl}|^2 \right) u_k + w_k', \quad (16)$$

where $\lambda_1, \ldots, \lambda_K$ are the eigenvalues of $HH^*$, $q_{kl}$ are elements of the matrix of eigenvectors of $HH^*$, and $w_k'$ combines the additive receiver noise $w_k$ and the interference. It is also shown that the signal-to-interference plus noise is approximately (for large $K$)

$$\text{SINR} \approx \frac{\left( \sum_{k=1}^{K} \frac{\lambda_k}{\lambda_k + \alpha} \right)^2} {\sigma^2 K^2 \gamma + K \sum_{k=1}^{K} \left( \frac{\lambda_k}{\lambda_k + \alpha} \right)^2 - \left( \sum_{k=1}^{K} \frac{\lambda_k}{\lambda_k + \alpha} \right)^2}. \quad (17)$$
Remarkably, (17) is maximized for $\alpha \geq 0$ at $\alpha_{\text{opt}} = K\sigma^2 = K/\rho$, independently of $\lambda_1, \ldots, \lambda_K$. We see that $\alpha_{\text{opt}}$ is proportional to $K$ and the noise variance. As we decrease the noise variance at each receiver, thereby increasing the signal-to-noise ratio, $\alpha_{\text{opt}} \to 0$.

The sum-rate for regularized channel inversion is obtained using a numerical estimate of the SINR with $\alpha = K/\rho$:

$$C_{\text{reg}} \approx KE \log(1 + \text{SINR}).$$

Unlike channel inversion, the sum-rate of regularized inversion has linear growth with $K$, although its slope is different from the sum-capacity.

![Fig. 2. Comparison of sum-capacity (5) (dashed line) as a function of $\rho$ for $K = M = 10$ with the regularized channel inversion sum-rate (18) (solid line) and the plain channel inversion sum-rate (dash-dotted line). At low power regularized inversion approaches $C_{\text{sum}}$, while for high $\rho$ it approaches $C_{\text{ci}}$.](image)

Figure 2 shows that for a fixed $K$, as $\rho \to \infty$ ($\sigma^2 \to 0$), the sum-rate of regularized inversion $C_{\text{reg}} \to C_{\text{ci}}$. Thus we still do not have a modulation technique which is close to capacity for all $\rho$ and $K$. The next section proposes a vector perturbation method that works for a wide range of $\rho$, and in Section V we combine regularization with perturbation.

## IV. Perturbing the Data

The previous sections argue that many of the problems with inverting the channel are due to the normalization constant $\gamma$, which is often very large because of the large eigenvalues in the inverse of the channel matrix $H$. One way to help $H$ is to regularize its inverse, as described in the previous section. Another way is to make sure the transmitted data does not lie along the eigenvectors associated with the large eigenvalues of $H^{-1}$. In this section, we present a way to “perturb” the data in a data-dependent way (unknown to the receivers) so that the data vector is approximately orthogonal to the eigenvectors associated with the large eigenvalues. Our goal therefore is to form a $\tilde{\mathbf{u}}$ from the data vector $\mathbf{u}$ such that

$$\mathbf{s} = H^{-1}\tilde{\mathbf{u}}$$

has norm (much) smaller than $H^{-1}\mathbf{u}$, but the entries of $\tilde{\mathbf{u}}$ can still be decoded individually at the receivers.

We cannot, in general, perturb $\mathbf{u}$ by an arbitrary complex vector because this perturbation is not known to the receivers and would therefore cause decoding errors. We can, however, use an idea derived from Tomlinson-Harashima precoding [12], [13] where we allow each element of $\mathbf{u}$ to be perturbed by an integer. In the simplest case, we set $\tilde{\mathbf{u}} = \mathbf{u} + \tau I$ where $\tau$ is a positive real number and $I$ is a
$K$-dimensional complex vector $a + ib$, where $a$ and $b$ are integers. The scalar $\gamma = \|s\|^2$ is computed as before and the transmitted signal is

$$x = \frac{1}{\sqrt{\gamma}}H^{-1}u.$$  \hfill (20)

The scalar $\tau$ is chosen large enough so that the receivers may apply the modulo function

$$f_\tau(y) = y - \left\lfloor \frac{y + \tau/2}{\tau} \right\rfloor \tau,$$  \hfill (21)

where the function $\lfloor \cdot \rfloor$ is the largest integer less than or equal to its argument. The function (21) removes the effect of the integer multiple of $\tau$. (The function $f_\tau(y)$ is applied separately to the real and imaginary components of a complex $y$.) We have more to say about the choice of $\tau$ shortly. After passing through the channel $H$, the transmitted signal $x$ in (20) appears at receiver $k$ as

$$y_k = \frac{1}{\sqrt{\gamma}}\bar{u}_k + w_k.$$

If we ignore for the moment the effect of $w_k$, and assume that $\gamma = 1$, then

$$f_\tau(y_k) = f_\tau(u_k + \tau l_k) = u_k,$$

and we recover the transmitted symbol. The receivers know $\gamma$ and therefore may compensate for $\gamma \neq 1$ by dividing $\tau$ by $\sqrt{\gamma}$. As we note in Section II, the transmitter may instead divide by $\sqrt{E\gamma}$; our simulations show that the performance difference is not significant. Nevertheless, our figures assume that the transmitter divides by $\sqrt{\gamma}$. An error is made at the receiver if the additive channel noise pushes the received signal across the standard symbol decoding boundaries or across the nonlinear boundaries of $f_\tau(y)$ at $\pm \tau/2$.

A. Choice of $l$

An obvious choice of $l$ at the transmitter minimizes $\gamma = \|s\|^2$,

$$l = \arg \min_l (u + \tau l)^*(HH^*)^{-1}(u + \tau l).$$  \hfill (22)

This is a $K$-dimensional integer-lattice least-squares problem, for which there are many choices of algorithms based on techniques by Fincke and Pohst [14] as well as Kannan [18]. See [15] for a review of algorithms that solve this problem. The Fincke-Pohst algorithm was used for space-time demodulation in [19]; it is labeled the sphere decoder in this context. Because we are using this algorithm for encoding data to be transmitted, we refer to it as the sphere encoder. We leave the details of this algorithm to the references, but mention that the algorithm avoids an exhaustive search over all possible integers in the lattice by limiting the search space to a sphere of some given radius centered around a starting point. In our case the center is the vector $u$. Generally, the sphere encoder works on real lattices, so we assume that a complex version is used [20], or that the equation (22) has been converted to a $2K$-dimensional real lattice problem.

The scalar $\tau > 0$ is a design parameter that may be chosen to provide a symmetric decoding region around (the real or imaginary part of) every signal constellation point, whence

$$\tau = 2(|c|_{\text{max}} + \Delta/2)$$  \hfill (23)

where $|c|_{\text{max}}$ is the absolute value of the constellation symbol(s) with largest magnitude and $\Delta$ is the spacing between constellation points. If we want to reduce the effects of the perturbation vector $l$, we may increase $\tau$, thereby increasing the decoding region at the upper and lower extremes of the constellation. While this improves error performance in these decoding regions, the $\gamma$ that results is typically also larger, possibly reducing total error performance. If $\tau$ is made too large, the minimization
in (22) yields \( l = 0 \) independently of \( u \), and the perturbation technique reduces to simple channel inversion. If \( \tau \) is made smaller than \( 2|c|_{\text{max}} \) then error-free decoding becomes impossible even in the absence of channel noise. We find that choosing \( \tau \) as in (23) often works well. Table I shows that in this case that the maximum value for \( \tau \) is 4 when a BPSK constellation is used and the minimum is \( \sqrt{6} \) when the signal is uniformly distributed.

\[
\begin{array}{|c|c|c|c|c|}
\hline
\text{Constellation} & \text{BPSK} & \text{QPSK} & \text{16-QAM} & \text{64-QAM} & \text{Uniform} \\
\hline
\tau & 4 & 2.8284 & 2.5298 & 2.4689 & 2.4495 \\
\hline
\end{array}
\]

**TABLE I**

VALUES FOR \( \tau \) AS IN (23) FOR UNIT-ENERGY CONSTellATIONS.

V. REGULARIZED PERTURBATION

We can marry the methods from Sections III and IV to reduce \( \gamma \) more than either method could alone over a wide range of \( \rho \). The choice of the integer vector \( l \) that minimizes \( \gamma \) is made with the modified cost function given in (6). Unfortunately, the analysis of the combined method appears to be difficult. In Section III, \( \alpha \) is chosen to maximize an approximation to the SINR. We do not know how to compute the average SINR after the minimization (6), and \( \alpha_{\text{opt}} \) generally no longer \( K/\rho \) when regularization is combined with perturbation. Because \( \gamma \) is significantly smaller in (6) than with regularization alone, \( \alpha = K/\rho \) is too large and gives too much crosstalk from the other users. The optimum \( \alpha \) is generally significantly smaller. For example, probability-of-error simulations show that \( \alpha_{\text{opt}} \approx 1/(5\rho) \) for \( K = 4 \), and \( \alpha_{\text{opt}} \approx 1/\rho \) for \( K = 10 \). We do not have a good explanation for these choices of \( \alpha \) and leave this as an open problem. We choose \( \tau \) from Table I.

A. Simulation of a Complete System

To check our distance from capacity, we simulated a complete system for \( M = K = 4 \) antennas/users and \( M = K = 10 \) antennas/users. The transmission and reception methods are summarized in Section II-B. To compare our results with the sum-capacity, we first present our operating points. We examine both \( M = K = 4 \) and \( M = K = 10 \), using 16-QAM constellations with either rate \( r = 1/2 \) (2 bits/user) and rate \( r = 3/4 \) (3 bits/user) codes. The sum rate is therefore

\[
R_{\text{sum}} = 4rK \text{ bits/channel use.} \quad (24)
\]

The possible sum-rates for \( r = 1/2 \) are therefore \( R_{\text{sum}} = 8 \) bits/channel-use and \( R_{\text{sum}} = 20 \) for \( K = 4 \) and \( K = 10 \), respectively. The sum-rates for \( r = 3/4 \) are \( R_{\text{sum}} = 12 \) and \( R_{\text{sum}} = 30 \). To find the receiver operating points that correspond to these sum rates, we turn to Figure 3 which shows the sum capacity for \( M = K = 4 \) and \( M = K = 10 \) systems as a function of \( \rho = 1/\sigma^2 \). These sum-capacity curves are computed by evaluating (5) numerically (we omit the details). The operating point for \( r = 1/2 \) is approximately \( \rho = 7 \) dB for either \( K = 4 \) or \( K = 10 \), and the operating point for \( r = 3/4 \) is approximately \( \rho = 11.2 \) dB for either \( K \). Since our total transmitted signal is normalized to unit power and our additive noise variance at each receiver is \( 1/\rho \), we define \( \frac{E_b}{N_0} = \frac{E_b}{r b} \), where \( b \) is the number of bits per constellation symbol (\( b = 4 \) for a 16-QAM constellation) and \( r \) is the channel code rate. This measure of bit-energy is included in our performance curves.

The 16-QAM constellation is mapped to bits using a standard Gray mapping. Successful bit-level turbo decoding at the receiver requires accurate knowledge of the likelihood function of the transmitted bits. The signal received by user \( k \) is \( y_k = (u_k + \tau l_k)/\sqrt{\gamma} + w_k' \), where \( w_k' \) contains the receiver noise and crosstalk from the other users. Then \( y_k \) is passed through the function \( f_{r/\sqrt{\gamma}}(\cdot) \), (see Section IV), the output of which is fed to the turbo decoder. Even though the additive noise \( w_k' \) is approximately Gaussian, the modulo function \( f_{r/\sqrt{\gamma}}(\cdot) \) operates on \( w_k' \), and the resulting likelihood function is therefore no longer Gaussian.
Fig. 3. Sum-capacity for $M = K = 4$ (lower curve) and $M = K = 10$ (upper curve) as a function of the receiver additive noise variance. The marker lines show that to achieve $C = 8$ ($K = 4$) or $C = 20$ ($K = 10$) the (reciprocal) noise variance must be $\rho = 1/\sigma^2 = 7$ dB. For $C = 12$ ($K = 4$) or $C = 30$ ($K = 10$) the noise variance must be $\rho = 11.2$ dB.

Figures 1 and 4 show the results of our trials. The turbo coders are formed from a UMTS standard parallel concatenated code with systematic component, feedforward polynomial $1 + D + D^3$, feedback polynomial $1 + D^2 + D^3$, and block length of 4000 bits. We assumed that the channel $H$ is interleaved and is therefore independent but known from channel-use to channel-use, so as to make the comparison with the ergodic capacities in Figure 3 meaningful. We see that the combination of regularization and vector perturbation performs to within approximately 4 dB from capacity and is significantly better than regularization alone. The $K = 10$ user system performs better than $K = 4$. This is perhaps surprising since the total system sum rate for $K = 10$ is 2.5 times the sum rate for $K = 4$. Perhaps this is because the larger system showed less variability in the transmission scale factor $\gamma$.

Fig. 4. Bit probability of error for rate $r = 1/2$ turbo-encoded data using 16-QAM symbols, for $M = K = 4$ and $M = K = 10$. The performance of combining regularization and perturbation is much better than regularization alone; the best performance is obtained for $M = K = 10$.

We are therefore transmitting at very high sum-rates (tens of bits/channel use) and we are reasonably close to capacity. There are some possible ways to get closer that we have yet to explore: i) Match the turbo code carefully to the channel or increase its block length; ii) Transmit at higher rates to the users whose channels happen to be best, since the sum-capacity is not necessarily attained by transmitting at equal rates to all of the users; iii) Computing and overcoming the penalty for using the mod-operation...
at the receiver.

Other possible applications of our technique include (single or multiple-antenna) CDMA systems, where a transmitter communicates with multiples users through spreading sequences that are nonorthogonal because of intersymbol interference. Our algorithm provides a method for compensating for the cross-coupling in the matrix of effective spreading codes $H$. We leave the analysis of this possible application for future work.

REFERENCES


