A Vector-Perturbation Technique for Near-Capacity Multiantenna Multiuser Communication—Part I: Channel Inversion and Regularization

Christian B. Peel, *Member, IEEE*, Bertrand M. Hochwald, *Member, IEEE*, and A. Lee Swindlehurst, *Fellow, IEEE*

***Abstract*—Recent theoretical results describing the sum capacity when using multiple antennas to communicate with multiple users in a known 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 power of the transmitted signal. This paper is comprised of two parts. In this first part, we show that while the sum capacity grows linearly with the minimum of the number of antennas and users, the sum rate of channel inversion does not. This poor performance is due to the large spread in the singular values of the channel matrix. We introduce regularization to improve the condition of the inverse and maximize the signal-to-interference-plus-noise ratio at the receivers. Regularization enables linear growth and works especially well at low signal-to-noise ratios (SNRs), but as we show in the second part, an additional step is needed to achieve near-capacity performance at all SNRs.**

***Index Terms*—Broadcast channel, channel inversion, multiple-antenna multiple-user wireless, multiple-input multiple-output (MIMO), regularization, spatial equalization.**

I. Introduction

CURRENT information-theoretic interest in multiple-input multiple-output (MIMO) communications has shifted, in part, away from point-to-point links and into multiuser (or "broadcast") links. Recent work by Caire and Shamai [1] and others [2]–[5] has shown that many of the advantages of using multiple antennas in a single-user scenario also translate to large gains in multiuser scenarios. We investigate simple techniques to achieve this multiuser gain.

It is well known that the point-to-point capacity of an 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 [6]. It is also shown in [6] that K users, each with one antenna, can transmit to a single receiver with M antennas, and the sum capacity (total of transmission rates from 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, and other time-division duplex (TDD) systems where channel conditions can readily be learned by all parties. Some multiantenna multiuser 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 basestation, 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 multiantenna 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 [7], 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 [8] for the case where channel side information is known noncausally at the transmitter.

Several researchers have investigated practical techniques to achieve the sum capacity promised by dirty-paper coding. Nested lattices are used in [9] for the interference channel, as well as the general multiuser channel. Trellis precoding for the broadcast channel is presented in [10] and [11] as a practical technique for the multiuser channel. These techniques are generally in preliminary states of development. Dirty-paper techniques are natural candidates for achieving sum capacity in multiantenna multiuser links, because the transmitted signal for one user can be viewed as interference for another user, and this

Paper approved by R. Schober, the Editor for Detection, Equalization, and MIMO of the IEEE Communications Society. Manuscript received September 30, 2003; revised May 21, 2004. This work was supported in part by the National Science Foundation under Wireless Initiative Grant CCR-9979452 and Information Technology Grant CCR-0081476. This paper was presented in part at the Allerton Conference on Communications, Control, and Computing, Monticello, IL, 2003.

C. B. Peel is with the Communications Technology Laboratory, Swiss Federal Institute of Technology, CH-8092 Zurich, Switzerland (e-mail: chris.peel@ieee.org).

B. M. Hochwald is with the Mathematics of Communications Research Department, Bell Laboratories, Murray Hill, NJ 07974 USA (e-mail: hochwald@lucent.com).

A. L. Swindlehurst is with the Electrical and Computer Engineering Department, Brigham Young University, Provo, UT 84602 USA (e-mail: swindle@ee.byu.edu).

Digital Object Identifier 10.1109/TCOMM.2004.840638

interference is known to the transmitter (the transmitter knows everybody's channel). However, it has not been shown that dirty-paper coding is necessary for achieving the majority of the capacity. Unlike Costa's original premise that the transmitter knows the interference but cannot control it, in our scenario, the transmitter creates all of the signals, and thereby can also control the interference seen by all the users.

In this two-part paper, we show that a suitably modified form of channel inversion can achieve near-sum-capacity performance. Channel inversion is one of the simplest modulation techniques for the multiuser channel [12]. 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. In this first part, we show that the sum rate for channel inversion (sometimes also referred to as "zero-forcing (ZF) beamforming" [1]) in its plain form is poor. We develop a regularized form of inversion that improves performance, especially at low signal-to-noise ratios (SNRs). We find the regularization parameter that maximizes the signal-to-interference-plus-noise ratio (SINR) at each receiver.

While regularization improves performance significantly, especially at low SNRs, another step is still needed to obtain near-capacity performance. The second part of this paper [13] describes a vector-perturbation technique that is used in conjunction with regularization to obtain good performance at all SNRs.

II. Model

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

$$y_k = \sum_{m=1}^{M} h_{k,m} x_m + w_k \tag{1}$$

where $h_{k,m}$ 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 ith antenna, and w_k is standard complex-Gaussian receiver noise seen at the kth user. The corresponding vector equation is

$$\mathbf{y} = H\mathbf{x} + \mathbf{w} \tag{2}$$

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 $h_{k,m}$ as elements. We assume that $\mathrm{E}\mathbf{w}\mathbf{w}^* = \sigma^2 I$ and impose the power constraint $\mathrm{E}||\mathbf{x}||^2 = 1$.

It is often convenient to construct an unnormalized signal $\mathbf{s}$, such that

$$\mathbf{x} = \frac{\mathbf{s}}{\sqrt{\gamma}} \tag{3}$$

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

$$\mathbf{x} = \frac{\mathbf{s}}{\sqrt{\mathrm{E}\gamma}}. \tag{4}$$

In this case, $\mathrm{E}||\mathbf{x}||^2 = 1$. Equation (3) has the advantage that $\mathrm{E}\gamma$ does not need to exist (we see later that in simple channel inversion, $\mathrm{E}\gamma = \infty$), but has the disadvantage that the receivers generally need to know γ, a channel- and data-dependent quantity, to decode their data properly. In the normalization (4), the receiver needs to know only $\mathrm{E}\gamma$, which is neither channel- nor data-dependent. Although it is more practical to use (4) (when it exists), we choose, for convenience, in most of our simulations to use the instantaneous power normalization (3). A discussion of the expected performance difference of using (3) versus (4) may be found in Section V of Part II. Generally, we find that the performance difference to be very small. Our simulations, therefore, represent the performance of either normalization, and we assume that the receivers need to know only $\mathrm{E}\gamma$.

We concentrate on the scenario where all K users are serviced at the same data rate. 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. Many of our theoretical results are obtained for large M and K limits, because the limiting results are often tractable. Nevertheless, we often consider M as small as four in our examples.

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

$$C_{\mathrm{sum}} = \mathrm{E} \sup_{D \in \mathcal{A}} \log |I_M + \rho H^* D H| \tag{5}$$

where $\mathcal{A}$ is the set of $K \times K$ diagonal matrices with nonnegative elements, such that $\mathrm{tr}\ D = 1$, and we define $\rho = 1/\sigma^2$. The Hermitian transpose of H is denoted H^*. We assume the logarithm is base two, and therefore C_{sum} is measured in bits/channel use. Although the total transmitted power is one, the quantity ρ is directly related to, but is not necessarily the same as, the SNR at each receiver. By simply choosing $D = (1/K)I_K$, we can easily infer that C_{sum} grows linearly with $\min(M, K)$. The expectation in (5) assumes that coding is done over multiple intervals with independent H. The maximization in (5) has no simple closed-form solution, so we compute (5) numerically using a gradient-type method as needed, 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 ρ 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 (DOF) (which has a more dramatic linear effect). On the other hand, when $K > M$, we know from (5) that although transmitting to at least M out of the K users simultaneously uses all of our available DOFs, 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.

In (2), the users all have the same average (but not instantaneous) received signal power, so our model assumes that the users are similar distances from the access point and are not in deep shadow fades. We also 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.

III. Channel Inversion: Some Old and New Results

A. *An Old Result* $(K < M)$

Channel inversion, when done at the transmitter, is sometimes known as ZF precoding, and entails deciding that the symbols $\{u_1, \dots, u_K\}$ seen at receivers $1, \dots, K$ should be chosen independently, according to the independent data desired for users $1, \dots, K$. We assume that the entries of the vector $\mathbf{u} = [u_1, \dots, u_K]^T$ are chosen from the same constellation with $\mathrm{E}|u_k|^2 = 1$ (ensuring equal rate to the users), and the transmitter then sets

$$\mathbf{s} = H^*(HH^*)^{-1}\mathbf{u}. \tag{6}$$

Generally, the inverse in (6) can be done only when $\beta = M/K \geq 1$. In this case, the asymptotic (as M and K go to infinity in this fixed ratio) sum rate of channel inversion is [15]

$$\lim_{\{M,K\}\to\infty} \frac{C_{\mathrm{ci}}}{M} = \frac{1}{\beta}\log\left(1 + \rho(\beta - 1)\right). \tag{7}$$

Let $\beta_0 = \beta_0(\rho)$ be the β that maximizes (7). Then $\beta_0 > 1$ is the optimum antenna/user ratio, and at this ratio, we can get to within roughly 80% of C_{sum} (5) computed at the same ratio [15]. However, at other ratios, the difference between C_{ci} and C_{sum} can become much more pronounced. For example, we see that as $\beta \to 1$, we have $C_{\mathrm{ci}}/M \to 0$. The implication is that for $K = M$, the sum rate of raw channel inversion does not increase linearly with K (or M), while C_{sum} clearly does. We analyze this shortcoming more closely in the next section.

B. *A New Result* $(K = M)$

When $K = M$, channel inversion (6) becomes simply

$$\mathbf{s} = H^{-1}\mathbf{u}. \tag{8}$$

This equation can obviously be problematic when H is poorly conditioned, and this problem manifests itself in the normalization constant (3)

$$\gamma = ||\mathbf{s}||^2 = \mathbf{u}^*(HH^*)^{-1}\mathbf{u}. \tag{9}$$

Let the entries of $\mathbf{u}$ be zero-mean unit-variance independent complex-Gaussian random variables. Then γ has density [15]

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

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

The received data at the kth user is

$$y_k = \frac{u_k}{\sqrt{\gamma}} + w_k. \tag{11}$$

The receivers all know γ, and we assume that K is large enough so that any user's data does not significantly affect the value of γ. Then, conditioned on γ, the channel becomes a scaled Gaussian channel; the capacity of this channel is

$$C_{\mathrm{ci},k} = \mathrm{E}\log\left(1 + \frac{\rho}{\gamma}\right) = \int_0^\infty d\gamma \log\left(1 + \frac{\rho}{\gamma}\right) K\frac{\gamma^{K-1}}{(1+\gamma)^{K+1}}. \tag{12}$$

A change of variables yields

$$C_{\mathrm{ci},k} = \int_0^\infty d\gamma \log\left(1 + \frac{\rho\gamma}{K}\right) \frac{1}{\left(\frac{\gamma}{K}+1\right)^{K+1}}. \tag{13}$$

Using the large K approximation $1/((\gamma/K+1)^{K+1}) \sim e^{-\gamma}$ in (13) (we omit the technical details showing that this substitution is valid in the integral) gives

$$C_{\mathrm{ci},k} \approx \int_0^\infty d\gamma \log\left(1 + \frac{\rho\gamma}{K}\right) e^{-\gamma} = e^{\frac{K}{\rho}} E_1\left(\frac{K}{\rho}\right)\log e \tag{14}$$

where

$$E_1(x) = \int_x^\infty dt \frac{e^{-t}}{t} \tag{15}$$

is the exponential integral. Since there are K users, each with receive (11), the sum rate for channel inversion is approximated for large $K = M$ by

$$C_{\mathrm{ci}} \approx K e^{\frac{K}{\rho}} E_1\left(\frac{K}{\rho}\right)\log e \text{ bits per channel use.} \tag{16}$$

We finally use the approximation $E_1(x) \sim e^{-x}/x$ for large x [16, p. 229] to conclude that

$$\lim_{K\to\infty} C_{\mathrm{ci}} = \rho \log e \text{ bits per channel use.} \tag{17}$$

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.

An explanation for this poor capacity comes from looking at the eigenvalues of $(HH^*)^{-1}$ (or singular values of H^{-1}). As shown in [17, Th. 5.5], 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

$$p(\mu) = \left(\frac{K}{\mu^2}\right) e^{-\frac{K}{\mu}} \tag{18}$$

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.

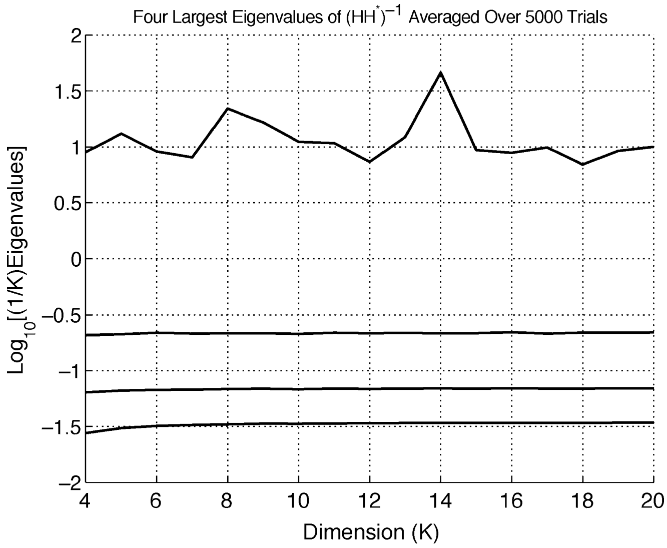


Fig. 1. Numerical comparison of the mean behavior of the four largest eigenvalues of $(HH^*)^{-1}$ as a function of K. The figure was generated using 5000 trials, and the eigenvalues are normalized by K. The largest eigenvalue has an erratic plot because its true mean is infinite (for all K), and it is clearly orders of magnitude larger than the remaining eigenvalues.

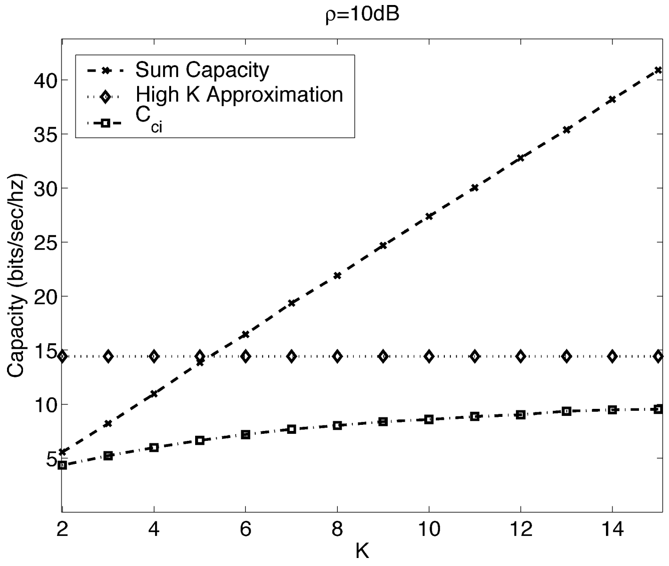


Fig. 2. Comparison of sum capacity (5) (dashed line) as a function of K for $\rho = 10$ dB, with the channel-inversion sum rate (K times the value in (12)) (dash-dotted line). Rather than growing linearly, C_{ci} approaches the large-K limit (17), which is shown as a dotted line.

It turns out that the remaining $K-1$ eigenvalues of $(HH^*)^{-1}$ are significantly better behaved. See Fig. 1 for a numerical comparison of the largest four eigenvalues of $(HH^*)^{-1}$ as a function of K. In fact, the smallest eigenvalue of $(HH^*)^{-1}$ concentrates (probabilistically) around $1/(4K)$ as $K \to \infty$ [17]. Therefore, any approach to improve channel inversion must seek to reduce the effects of the largest eigenvalue.

Fig. 2 shows the sum rate for channel inversion evaluated numerically, the large-K expression (17) and the sum capacity (5). We can see that as the number of transmit antennas and users grow simultaneously, the sum rate for channel inversion approaches $\rho \log e$, while the sum capacity grows linearly.

We assume that $K = M$ in the remainder of the paper.

IV. 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 $\mathbf{s}$ using (8), we use

$$\mathbf{s} = H^*(HH^* + \alpha I_K)^{-1}\mathbf{u}. \tag{19}$$

After going through the channel, the unnormalized signal $\mathbf{s}$ becomes

$$H\mathbf{s} = HH^*(HH^* + \alpha I)^{-1}\mathbf{u}. \tag{20}$$

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.

To evaluate the amount of desired signal and interference, we use the decomposition $HH^* = Q\Lambda Q^*$ for nonnegative diagonal eigenvalue matrix Λ and unitary eigenvector matrix Q to find

$$H\mathbf{s} = Q\frac{\Lambda}{\Lambda + \alpha I}Q^*\mathbf{u}. \tag{21}$$

(We use the convention that commuting matrices can be treated as scalars, and therefore, may appear in fractional form.) The (unnormalized) signal and interference received by user k is the kth entry of $H\mathbf{s}$. Using (21), we may find this entry

$$\begin{aligned}[H\mathbf{s}]_k = &\left[q_{k,1}\frac{\lambda_1}{\lambda_1+\alpha} \quad \cdots \quad q_{k,K}\frac{\lambda_K}{\lambda_K+\alpha}\right] \\ &\times \begin{bmatrix} q_{1,1}^* & \cdots & q_{K,1}^* \\ \vdots & \ddots & \vdots \\ q_{1,K}^* & \cdots & q_{K,K}^* \end{bmatrix} \begin{bmatrix} u_1 \\ \vdots \\ u_K \end{bmatrix}\end{aligned} \tag{22}$$

where $q_{k,\ell}$ is the (k,ℓ)th entry of the matrix Q. The (unnormalized) desired signal term in (22) is

$$\left(\sum_{\ell=1}^{K} \frac{\lambda_\ell}{\lambda_\ell + \alpha}|q_{k,\ell}|^2\right) u_k. \tag{23}$$

All of the remaining terms in (22) involving $u_\ell (\ell \neq k)$ are interference.

The kth user models its (normalized) received signal as

$$y_k = \left(\frac{1}{\sqrt{\gamma}}\right)\left(\sum_{\ell=1}^{K} \frac{\lambda_\ell}{\lambda_\ell + \alpha}|q_{k,\ell}|^2\right) u_k + w_k' \tag{24}$$

where the Gaussian w_k' combines the additive receiver noise w_k and the interference. The receiver makes its decisions about the transmitted signal by forming the likelihood function from (24).

The amount of interference is determined by $\alpha > 0$; when $\alpha = 0$, we return to (8). It is clear that, no matter how poorly conditioned H is, the inverse in (19) can be made to behave as well as desired by choosing α large enough. We examine the optimum value of α to choose. The amount of interference increases with α, so one possible metric for choosing α is to maximize the SINR in (24). We compute the SINR by computing the expected power of the desired signal and dividing it by the expected power of the interference plus noise.

The noise power at each receiver is given by σ^2. From (2), the signal (without noise) observed at the K receivers is $H\mathbf{x} = (1/\sqrt{\mathrm{E}\gamma})H\mathbf{s}$ (we assume that the average power normalization (4) is used). We need to examine the relative strengths of the desired signal and interference at each receiver; we first examine the behavior of $\gamma = \|\mathbf{s}\|^2$. We use the eigenvalue/eigenvector decomposition of HH^* to obtain

$$\begin{aligned}\gamma &= \mathbf{u}^*(HH^*+\alpha I)^{-1}HH^*(HH^*+\alpha I)^{-1}\mathbf{u}\\ &= \mathrm{tr}\left[(Q\Lambda Q^*+\alpha I)^{-1}Q\Lambda Q^*(Q\Lambda Q^*+\alpha I)^{-1}\mathbf{u}\mathbf{u}^*\right]\\ &= \mathrm{tr}\left[Q(\Lambda+\alpha I)^{-1}Q^*Q\Lambda Q^*Q(\Lambda+\alpha I)^{-1}Q^*\mathbf{u}\mathbf{u}^*\right]\\ &= \mathrm{tr}\left[\frac{\Lambda}{(\Lambda+\alpha I)^2}Q^*\mathbf{u}\mathbf{u}^*Q\right].\end{aligned}$$

We assume that the data $u_1,\ldots,u_K$ are independently chosen with zero mean and unit variance. Taking the conditional expectation of γ with respect to $\mathbf{u}$ and using $\mathrm{E}\mathbf{u}\mathbf{u}^* = I_K$, we get

$$\mathrm{E}\gamma = \mathrm{tr}\left[\frac{\Lambda}{(\Lambda+\alpha I)^2}\right] = \sum_{\ell=1}^{K}\frac{\lambda_\ell}{(\lambda_\ell+\alpha)^2}. \tag{25}$$

We find it convenient to take expectations only with respect to $\mathbf{u}$ and Q when evaluating the quantities needed to compute the SINR. The expectations with respect to Λ are generally difficult. We show later that, fortunately, the final result does not require us to take the expectation with respect to Λ.

From (21), the total expected power in $H\mathbf{s}$ is

$$\begin{aligned}\mathrm{E}\|H\mathbf{s}\|^2 &= \mathrm{E}\mathbf{u}^*Q\left(\frac{\Lambda}{\Lambda+\alpha I}\right)^2Q^*\mathbf{u}\\ &= \sum_{\ell=1}^{K}\frac{\lambda_\ell^2}{(\lambda_\ell+\alpha)^2}.\end{aligned} \tag{26}$$

We again avoid the expectation with respect to Λ. The desired signal for the kth user is given by (23). To find the expected power of the desired signal, we compute the expectation over Q in Appendix A, using the fact that Q and Λ are statistically independent [17]

$$\begin{aligned}\text{Desired} &= \mathrm{E}\left(\sum_{\ell=1}^{K}\frac{\lambda_\ell}{\lambda_\ell+\alpha}|q_{k,\ell}|^2\right)^2\\ &= \frac{1}{K(K+1)}\left[\left(\sum_{\ell=1}^{K}\frac{\lambda_\ell}{\lambda_\ell+\alpha}\right)^2+\sum_{\ell=1}^{K}\left(\frac{\lambda_\ell}{\lambda_\ell+\alpha}\right)^2\right].\end{aligned} \tag{27}$$

This is the unnormalized power of the desired signal at the receiver. Observe that this power is one when $\alpha = 0$ (plain channel inversion). The normalized power divides (27) by $\mathrm{E}\gamma$.

The total signal and interference power at any receiver is $1/K$th of the total (unnormalized) power appearing at all the receivers (26), which is $(1/K)\sum_{\ell=1}^{K}\lambda_\ell^2/(\lambda_\ell+\alpha)^2$. Hence, subtracting off the power of the desired signal (27) leaves the power of the interference $u_\ell(\ell \neq k)$ at receiver k as

$$\begin{aligned}\text{Undesired} = \frac{1}{K}\sum_{\ell=1}^{K}\left(\frac{\lambda_\ell}{\lambda_\ell+\alpha}\right)^2 - \frac{1}{K(K+1)}\\ \times\left[\left(\sum_{\ell=1}^{K}\frac{\lambda_\ell}{\lambda_\ell+\alpha}\right)^2+\sum_{\ell=1}^{K}\left(\frac{\lambda_\ell}{\lambda_\ell+\alpha}\right)^2\right].\end{aligned} \tag{28}$$

This is the unnormalized power of the interference at each receiver. Observe that this power is zero when $\alpha = 0$. The normalized power divides (28) by $\mathrm{E}\gamma$.

Putting (27) and (28) together, normalized by γ as given by (25), yields (29), shown at the bottom of the page. Because of the symmetry in the distribution of H, (29) is not a function of the user k. Rather than optimize (29) directly over α, we prefer to optimize a simpler large-K approximation to (29).

The large-K approximation follows from removing the second summation in the numerator of (29), which is dwarfed by the first summation, and replacing $K(K+1)$ by K^2. We then obtain

$$\mathrm{SINR} \approx \frac{\left(\sum_{\ell=1}^{K}\frac{\lambda_\ell}{\lambda_\ell+\alpha}\right)^2}{\sigma^2K^2\sum_{\ell=1}^{K}\frac{\lambda_\ell}{(\lambda_\ell+\alpha)^2}+K\sum_{\ell=1}^{K}\left(\frac{\lambda_\ell}{\lambda_\ell+\alpha}\right)^2-\left(\sum_{\ell=1}^{K}\frac{\lambda_\ell}{\lambda_\ell+\alpha}\right)^2}. \tag{30}$$

Remarkably, the large-K approximation (30) is maximized for $\alpha \geq 0$ at $\alpha_{\mathrm{opt}} = K\sigma^2 = K/\rho$, *independently of* $\lambda_1,\ldots,\lambda_K$. For a proof, consult Appendix B. Simulations indicate that (30) is close to the true SINR, for even small values of K. We see that α_{opt} is proportional to K and the noise variance. As we decrease the noise variance at each receiver, thereby increasing the SNR, $\alpha_{\mathrm{opt}} \to 0$.

The above analysis applies to any eigenvalue distribution of the channel matrix H. Our analysis only uses the fact that the eigenvector matrix Q has the so-called isotropic distribution, whose defining characteristic is that pre- or postmultiplying Q by any unitary matrix does not affect its distribution (see [18] and references therein). Physically, this means that the channel is not affected by arbitrary rotations, and that paths between the antennas and the users are statistically equivalent. It is this feature that allows us to examine the SINR of any user and claim that this analysis applies equally to the remaining users. Our analysis, therefore, applies to other channel distributions with this rotational-invariance property, and not just a Gaussian H.

We mention that the reviewers of our manuscript made a connection between our SINR analysis and a minimum mean-

$$\mathrm{SINR} = \frac{\left(\sum_{\ell=1}^{K}\frac{\lambda_\ell}{\lambda_\ell+\alpha}\right)^2+\sum_{\ell=1}^{K}\left(\frac{\lambda_\ell}{\lambda_\ell+\alpha}\right)^2}{\sigma^2K(K+1)\sum_{\ell=1}^{K}\frac{\lambda_\ell}{(\lambda_\ell+\alpha)^2}+K\sum_{\ell=1}^{K}\left(\frac{\lambda_\ell}{\lambda_\ell+\alpha}\right)^2-\left(\sum_{\ell=1}^{K}\frac{\lambda_\ell}{\lambda_\ell+\alpha}\right)^2} \tag{29}$$

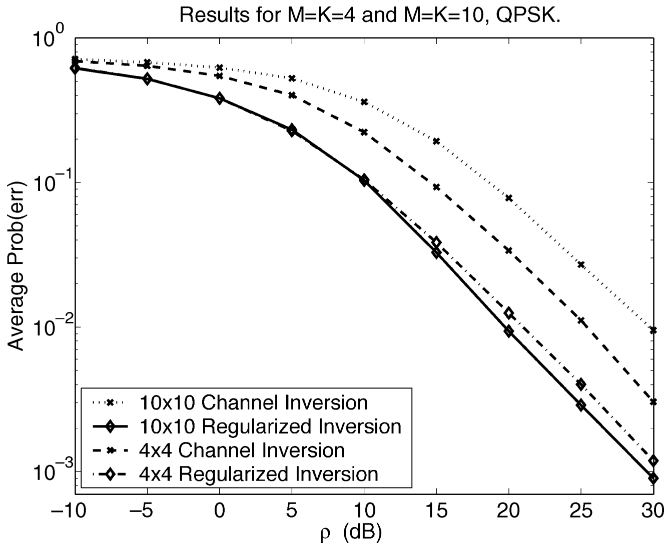


Fig. 3. Comparison of the SEP for plain (8) and regularized (19) channel inversion for $K = 4$ and $K = 10$. The raw error rate as a function of K worsens for plain channel inversion, but improves (slightly) for regularized inversion.

square error (MMSE) analysis available in [19] (and references therein). The MMSE cost function is

$$\arg\min_{P,\gamma} \varepsilon(P,\gamma) = \mathrm{E}||\mathbf{u} - \sqrt{\mathrm{E}\gamma}\mathbf{y}||^2 \tag{31}$$

expressing the square difference between the data vector $\mathbf{u}$ and a scaled version of what is received $\mathbf{y}$ (as a vector). It is assumed that the transmitted signal is $P\mathbf{u}$ and has a unit-energy constraint $\mathrm{E}||P\mathbf{u}||^2 = 1$. As shown in [19], the optimizing P of (31) is $P_{\mathrm{MMSE}} = H^*(HH^* + \alpha_{\mathrm{MMSE}}I_K)^{-1}$, where $\alpha_{\mathrm{MMSE}} = K/\rho$. Thus, despite the difference in our per-user cost function (29) and the vector MMSE cost function (31), the optimizing α's are very similar. However, they are not equal, except in the large-K limit.

V. Performance, Capacity, and Discussion

Three figures show the trends in performance. Fig. 3 shows the symbol-error probability (SEP) for plain and regularized channel inversion as a function of ρ for $K = 4$ and $K = 10$. The curves indicate that while the performance of plain channel inversion worsens with K, the performance of regularized inversion improves slightly with K.

A comparison of the sum capacity and sum rates for regularized and plain channel inversion as a function of K is shown in Fig. 4. The sum rate for regularized channel inversion is obtained using a numerical estimate of the SINR with $\alpha = K/\rho$

$$C_{\mathrm{reg}} \approx K\log(1 + \mathrm{SINR}). \tag{32}$$

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

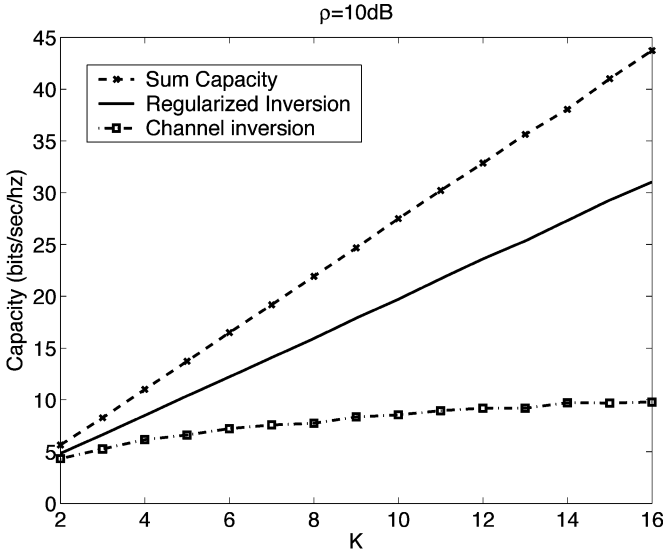


Fig. 4. Comparison of the sum capacity (5) (dashed line) as a function of K (where $M = K$) for $\rho = 10$ dB, with the regularized channel-inversion sum rate (32) (solid line) and the plain channel-inversion sum rate (dash-dotted line). Unlike plain channel inversion, regularized inversion has linear growth with K.

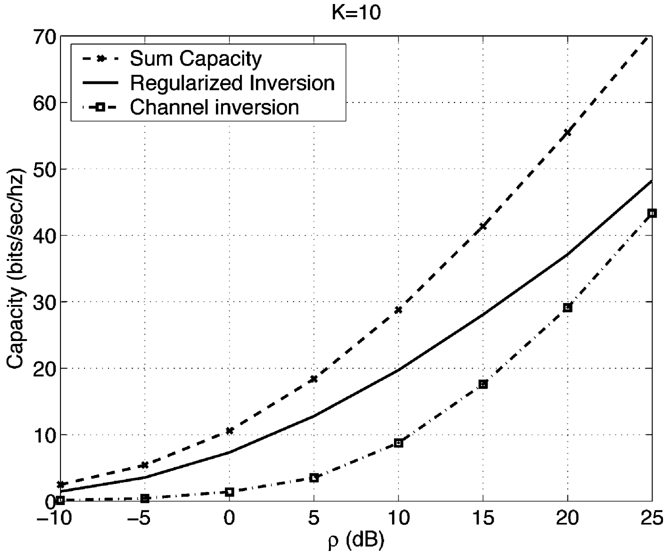


Fig. 5. Comparison of sum capacity (5) (dashed line) as a function of ρ for $K = M = 10$, with the regularized channel-inversion sum rate (32) (solid line) and the plain channel-inversion sum rate (dash-dotted line). At low power, regularized inversion approaches C_{sum}, while for high ρ, it approaches C_{ci}.

Fig. 5 shows that for a fixed K, as $\rho \to \infty(\sigma^2 \to 0)$, the sum rate of regularized inversion approaches plain inversion $C_{\mathrm{reg}} \to C_{\mathrm{ci}}$. Thus, we still do not have a modulation technique which is close to capacity for all ρ and K.

These three figures show that although regularization is a big improvement over plain inversion, a gap to capacity remains, especially at high SNR. This gap is dramatically reduced in the next paper [13], which shows how to combine regularization with a carefully chosen integer vector perturbation of the data to be transmitted to reduce the power of the transmitted signal dramatically.

APPENDIX

A. Expectation Over Q of (27)

We begin with calculations involving the elements of any row of Q. Let $q=q_{k,\ell}$ and $q'=q_{k,\ell'}$ for any $\ell'\neq\ell$. We use the following probability densities from [18]:

$$\begin{aligned}p(q)&=\frac{K-1}{\pi}\left(1-|q|^2\right)^{K-2},\quad 0\le|q|^2\le1\\ p(q,q')&=\frac{(K-1)(K-2)}{\pi^2}\left(1-|q|^2-|q'|^2\right),\\ &\qquad 0\le|q|^2+|q'|^2\le1.\end{aligned}$$

These distributions are with respect to the complex plane. Let $r=|q|^2$ and $r'=|q'|^2$. Then the transformation $q=\sqrt{r}(\cos\theta+i\sin\theta)$ yields

$$p(r)=(K-1)(1-r)^{K-2},\quad 0\le r\le1 \tag{33}$$

$$\begin{aligned}p(r,r')&=(K-1)(K-2)(1-r-r')^{K-3}\\ &\qquad 0\le r+r'\le1.\end{aligned} \tag{34}$$

Then, (33) implies

$$\mathrm{E}|q|^4=\mathrm{E}r^2=(K-1)\int_0^1 dr\, r^2(1-r)^{K-2}=\frac{2}{K(K+1)} \tag{35}$$

where we omit the integration by parts calculations. Equation (34) implies that

$$\begin{aligned}\mathrm{E}|q|^2|q'|^2&=\mathrm{E}rr'\\ &=(K-1)(K-2)\int_0^1 dr\, r\int_0^{1-r} dr'\, r'(1-r-r')^{K-3}\\ &=\frac{1}{K(K+1)}.\end{aligned} \tag{36}$$

Using (35) and (36), we may compute the expectation in (27)

$$\begin{aligned}&\mathrm{E}\left(\sum_{\ell=1}^{K}\frac{\lambda_\ell}{\lambda_\ell+\alpha}|q_{k,\ell}|^2\right)^2\\ &=\mathrm{E}\left[\sum_{\ell=1}^{K}\left(\frac{\lambda_\ell}{\lambda_\ell+\alpha}\right)^2|q_{k,\ell}|^4\right.\\ &\qquad\left.+\sum_{\substack{\ell,m=1\\ \ell\neq m}}^{K}\frac{\lambda_\ell\lambda_m}{(\lambda_\ell+\alpha)(\lambda_m+\alpha)}|q_{k,\ell}|^2|q_{k,m}|^2\right]\\ &=\frac{2}{K(K+1)}\sum_{\ell=1}^{K}\left(\frac{\lambda_\ell}{\lambda_\ell+\alpha}\right)^2+\frac{1}{K(K+1)}\\ &\quad\times\sum_{\substack{\ell,m=1\\ \ell\neq m}}^{K}\frac{\lambda_\ell\lambda_m}{(\lambda_\ell+\alpha)(\lambda_m+\alpha)}\\ &=\frac{1}{K(K+1)}\left[\left(\sum_{\ell=1}^{K}\frac{\lambda_\ell}{\lambda_\ell+\alpha}\right)^2+\sum_{\ell=1}^{K}\left(\frac{\lambda_\ell}{\lambda_\ell+\alpha}\right)^2\right].\end{aligned}$$

B. Proof That SINR is Maximized When $\alpha=K\sigma^2$

We find the stationary points by setting the derivative equal to zero

$$\frac{d}{d\alpha}\frac{\left(\sum\frac{\lambda}{\lambda+\alpha}\right)^2}{\sigma^2K^2\sum\frac{\lambda}{(\lambda+\alpha)^2}+K\sum\left(\frac{\lambda}{\lambda+\alpha}\right)^2-\left(\sum\frac{\lambda}{\lambda+\alpha}\right)^2}=0$$

where the sum is over $\lambda_1,\ldots,\lambda_K$. After some algebraic manipulations, the expression above becomes

$$\begin{aligned}0=&\sum\frac{\lambda}{\lambda+\alpha}\left[\sigma^2K\sum\frac{\lambda}{(\lambda+\alpha)^3}+\sum\frac{\lambda^2}{(\lambda+\alpha)^3}\right]\\ &-\sum\frac{\lambda}{(\lambda+\alpha)^2}\left[\sigma^2K\sum\frac{\lambda}{(\lambda+\alpha)^2}+\sum\left(\frac{\lambda}{\lambda+\alpha}\right)^2\right]\\ 0=&\sigma^2K\sum\frac{\lambda}{\lambda+\alpha}\sum\frac{\lambda}{(\lambda+\alpha)^3}+\sum\frac{\lambda}{\lambda+\alpha}\sum\frac{\lambda^2}{(\lambda+\alpha)^3}\\ &-\sigma^2K\left(\sum\frac{\lambda}{(\lambda+\alpha)^2}\right)^2-\sum\frac{\lambda}{(\lambda+\alpha)^2}\sum\left(\frac{\lambda}{\lambda+\alpha}\right)^2.\end{aligned}$$

We note that this equality is true for any α when $K=1$. For $K>1$, further manipulations convert this equation into

$$0=\sum_{k<\ell}\frac{\lambda_k\lambda_\ell(\lambda_k-\lambda_\ell)^2(K\sigma^2-\alpha)}{(\lambda_k+\alpha)^3(\lambda_\ell+\alpha)^3}.$$

Hence, provided that $\lambda_1,\ldots,\lambda_K$ are not all equal, the derivative is zero if and only if $\alpha=K\sigma^2$. That this stationary point is a maximum follows from checking the second derivative, which, in the interest of saving paper, we do not do here.

ACKNOWLEDGMENT

The authors thank T. Marzetta at Bell Laboratories for help with the computations in Appendix A. They also acknowledge the helpful comments and suggestions of the anonymous reviewers.

REFERENCES

[1] G. Caire and S. Shamai, "On the achievable throughput of a multi-antenna Gaussian broadcast channel," *IEEE Trans. Inf. Theory*, vol. 43, pp. 1691–1706, Jul. 2003.

[2] W. Yu and J. Cioffi, "Sum capacity of a Gaussian vector broadcast channel," in *Proc. IEEE Int. Symp. Inf. Theory*, Jul. 2002, p. 498.

[3] P. Viswanath and D. Tse, "Sum capacity of the vector Gaussian broadcast channel and uplink–downlink duality," *IEEE Trans. Inf. Theory*, vol. 49, pp. 1912–1921, Aug. 2003.

[4] S. Vishwanath, N. Jindal, and A. Goldsmith, "Duality, achievable rates, and sum capacity of Gaussian MIMO broadcast channels," *IEEE Trans. Inf. Theory*, vol. 49, pp. 2658–2668, Aug. 2003.

[5] H. Weingarten, Y. Steinberg, and S. Shamai, "The capacity region of the Gaussian MIMO broadcast channel," in *Proc. Conf. Inf. Sci. Syst. (CISS)*, Princeton, NJ, Mar. 2004, pp. 7–12.

[6] I. E. Telatar, "Capacity of multi-antenna Gaussian channels," *Eur. Trans. Telecommun.*, vol. 10, pp. 585–595, Nov./Dec. 1999.

[7] M. Costa, "Writing on dirty paper," *IEEE Trans. Inf. Theory*, vol. IT-29, pp. 439–441, May 1983.

[8] S. I. Gelfand and M. S. Pinsker, "Coding for channel with side information," *Problemi Peredachi Informatsii*, vol. 9, no. 1, pp. 19–31, 1980.

[9] R. Zamir, S. Shamai, and U. Erez, "Nested linear/lattice codes for structured multiterminal binning," *IEEE Trans. Inf. Theory*, vol. 48, pp. 1250–1276, Jun. 2002.

[10] W. Yu and J. Cioffi, "Trellis precoding for the broadcast channel," in *Proc. IEEE GLOBECOM*, Nov. 2001, pp. 1344–1348.

[11] J. Kusuma and K. Ramchandran, "Communicating by cosets and applications to broadcast," in *Proc. Conf. Inf. Sci. Syst. (CISS)*, Princeton, NJ, Mar. 2002, [CD-ROM].

[12] T. Haustein, C. von Helmolt, E. Jorswieck, V. Jungnickel, and V. Pohl, "Performance of MIMO systems with channel inversion," in *Proc. 55th IEEE Veh. Technol. Conf.*, vol. 1, Birmingham, AL, May 2002, pp. 35–39.

[13] C. B. Peel, B. M. Hochwald, and A. L. Swindlehurst, "A vector-perturbation technique for near-capacity multiantenna multiuser communication—Part II: Perturbation," *IEEE Trans. Commun.*, to be published.

[14] U. Fincke and M. Pohst, "Improved methods for calculating vectors of short lengths in a lattice, including a complexity analysis," *Math. Comput.*, vol. 44, pp. 463–471, Apr. 1985.

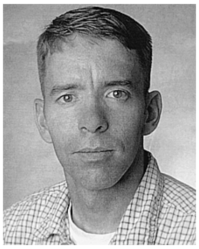
[15] B. M. Hochwald and S. Vishwanath, "Space–time multiple access: Linear growth in the sum rate," in *Proc. 40th Allerton Conf. Comput., Commun., Control*, Monticello, IL, Oct. 2002, pp. 387–396.

[16] M. Abramowitz and I. Stegun, *Handbook of Mathematical Functions With Formulas, Graphs, and Mathematical Tables*, 1st ed. Washington, DC: U.S. Dept. of Commerce, 1972.

[17] A. Edelman, "Eigenvalues and condition numbers of random matrices," Ph.D. dissertation, Dept. of Math., Mass. Inst. Technol., Cambridge, MA, 1989.

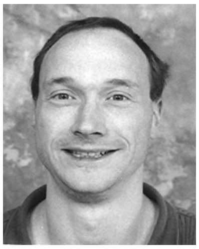
[18] T. L. Marzetta and B. M. Hochwald, "Capacity of a mobile multiple-antenna communication link in Rayleigh flat fading," *IEEE Trans. Inf. Theory*, vol. 45, pp. 139–157, Jan. 1999.

[19] M. Joham, K. Kusume, M. H. Gzara, and W. Utschick, "Transmit Wiener filter for the downlink of TDD DS-CDMA systems," in *Proc. IEEE 7th Symp. Spread-Spectrum Technol., Applicat.*, Prague, Czech Republic, Sep. 2002, pp. 9–13.



Christian B. Peel (S'93–M'98) received the B.S. (*magna cum laude*) and M.S. degrees in electrical engineering from Utah State University (USU), Logan, UT, in 1995 and 1997, respectively, and the Ph.D. degree in electrical engineering from Brigham Young University (BYU), Provo, UT, in 2004.

From 1992 to 1994, he was with the Space Dynamics Laboratory, USU, working on infrared sensor calibration. From 1993 to 1994, he attended the Siberian Aerospace Academy, Krasnoyarsk, Russia, on a U.S. Information Agency scholarship. He worked as Research Assistant (1995-1997) and Research Engineer (1997-1999) with the Electrical Engineering Department, USU, doing research on image and video compression. From 2000-2004, he was a Research Assistant with BYU, working on space–time modulation. He visited the Mathematics of Communications Department, Bell Laboratories, Murray Hill, NJ, in the fall of 2002, where he investigated coding techniques for the multiple-antenna broadcast channel. He is currently a Postdoctoral Researcher with the Communication Technology Laboratory, Swiss Federal Institute of Technology (ETH), Zurich, Switzerland.



Bertrand Hochwald (S'90–M'95) was born in New York, NY. He received the undergraduate degree from Swarthmore College, Swarthmore, PA, and the M.S. degree in electrical engineering from Duke University, Durham, NC. In 1989, he enrolled at Yale University, New Haven, CT, where he received the M.A. degree in statistics and the Ph.D. degree in electrical engineering.

From 1986 to 1989, he worked for the United States Department of Defense, Fort Meade, MD. In 1995–1996 he was a Research Associate and Visiting Assistant Professor at the Coordinated Science Laboratory, University of Illinois, Urbana-Champaign. He joined the Mathematics of Communications Research Department at Lucent Technologies Bell Laboratories, Murray Hill, NJ, in September 1996, where he is now a Distinguished Member of the Technical Staff. He holds several patents in the field of multiantenna wireless communication.

Dr. Hochwald is the recipient of several achievement awards while with the Department of Defense and the Prize Teaching Fellowship at Yale.

A. Lee Swindlehurst (S'83–M'84–SM'99–F'04) received the B.S. (*summa cum laude*) and M.S. degrees in electrical engineering from Brigham Young University (BYU), Provo, Utah, in 1985 and 1986, respectively, and the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA, in 1991.

From 1983 to 1984, he was with Eyring Research Institute, Provo, UT, as a Scientific Programmer. During 1984–1986, he was a Research Assistant in the Department of Electrical Engineering, BYU, working on various problems in signal processing and estimation theory. From 1985–1988, he was affiliated with the Information Systems Laboratory, Stanford University. From 1986 to 1990, he was with ESL, Inc., Sunnyvale, CA, where he was involved in the design of algorithms and architectures for several radar and sonar signal processing systems. He joined the faculty of the Department of Electrical and Computer Engineering, BYU, in 1990, where he holds the position of Full Professor. During 1996–1997, he held a joint appointment as a Visiting Scholar at both Uppsala University, Uppsala, Sweden, and at the Royal Institute of Technology, Stockholm, Sweden. His research interests include sensor array signal processing for radar and wireless communications, detection and estimation theory, and system identification.

Dr. Swindlehurst is currently serving as Secretary of the IEEE Signal Processing Society, and is a past Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING, a past member of the Statistical Signal and Array Processing Technical Committee in the IEEE Signal Processing Society, and past Vice-Chair of the Signal Processing for Communications Technical Committee within the same society. He has served as the Technical Program Chair for the 1998 IEEE Digital Signal Processing Workshop and for the 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. He was awarded an Office of Naval Research Graduate Fellowship for 1985–1988. He is also a recipient of the 2000 IEEE W. R. G. Baker Prize Paper Award, and is coauthor of a paper that received a Signal Processing Society Young Author Best Paper Award in 2001.