De-hyping Transmit Diversity in Modern MIMO Cellular Systems

Nihar Jindal
University of Minnesota
Minneapolis, MN55455, USA
Email: nihar.jindal@umn.edu

Angel Lozano
Universitat Pompeu Fabra
08005 Barcelona, Spain
Email: angel.lozano@upf.edu

Abstract—**A contemporary perspective on the tradeoff between transmit antenna diversity and spatial multiplexing is provided. It is argued that, in the context of modern cellular systems and for the operating points of interest, transmission techniques that utilize all available spatial degrees of freedom for multiplexing outperform techniques that explicitly sacrifice spatial multiplexing for diversity. Reaching this conclusion, however, requires that the channel and some key system features be adequately modeled; failure to do so may bring about starkly different conclusions. As a specific example, this contrast is illustrated using the 3GPP Long-Term Evolution system design.**

I. Introduction

Multipath fading is one of the most fundamental features of wireless channels. Fortunately, fades are very localized in space and frequency: a sufficient change in the transmitter or receiver location or in the frequency leads to a roughly independent realization of the fading process. Motivated by this *selectivity*, the concept of *diversity* is borne: rather than making the success of a transmission entirely dependent on a single fading realization, hedge the transmission's success across multiple ones.

The notion of diversity, however, is indelibly associated with channel uncertainty. If the transmitter knows the instantaneous channel state, then it can match its transmission to the channel in such a way that the error probability depends only on the noise. Diversity techniques, which aim precisely at mitigating the effects of channel uncertainty, are then beside the point. Although perhaps evident, this point is often neglected. In models commonly used to evaluate diversity techniques, for instance, the channel fades very slowly yet there is no transmitter adaptation. This is at odds with how contemporary cellular systems operate, where link adaptation is a chief feature.

A longer version of this work has been submitted to *IEEE Trans. on Wireless Communications*.

II. Channel Model and Performance Metrics

Let $n_{\rm T}$ and $n_{\rm R}$ denote, respectively, the number of transmit and receive antennas. Assuming that OFDM is used to decompose the channel into N parallel non-interfering tones, the received signal on the ith tone is

$$\mathbf{y}_i = \mathbf{H}_i \mathbf{x}_i + \mathbf{n}_i \tag{1}$$

where $\mathbf{H}_i$ is the $n_{\rm R} \times n_{\rm T}$ channel matrix on that tone, $\mathbf{y}_i$ is the $n_{\rm R} \times 1$ received signal, $\mathbf{n}_i$ is thermal noise, IID circularly symmetric complex Gaussian with unit variance, and $\mathbf{x}_i$ is the $n_{\rm T} \times 1$ transmitted signal subject to a power constraint SNR, i.e., $E[||\mathbf{x}_i||^2] \leq \mathsf{SNR}$. The receiver has perfect knowledge of the N channel matrices; transmitter CSI (channel state information) is further discussed in Section IV.

For a particular realization of $\mathbf{H}_1, \ldots, \mathbf{H}_N$, the average mutual information (bits/s/Hz) thereon is

$$\mathcal{I}(\mathsf{SNR}) = \frac{1}{N} \sum_{i=1}^{N} I(\mathbf{x}_i; \mathbf{y}_i). \tag{2}$$

where $I(\mathbf{x}_i; \mathbf{y}_i)$ depends on the transmission strategy (cf. Section V).

The outage probability for rate R (bits/s/Hz) is then

$$P_{\text{out}}(\text{SNR}, R) = \Pr\{\mathcal{I}(\text{SNR}) < R\}. \tag{3}$$

With suitably strong codes, the outage probability is an accurate approximation to the block error probability [1] and we shall thus use both notions interchangeably henceforth.

As justified in Section IV, modern systems operate at a target error probability. Hence, the primary performance metric is the maximum rate, at each SNR, such that this target ϵ is not exceeded, i.e.,

$$R_\epsilon(\text{SNR}) = \max_\zeta\{\zeta : P_{\text{out}}(\text{SNR}, \zeta) \leq \epsilon\}. \tag{4}$$

III. The Outage-Rate Tradeoff and the DMT

Eq. (3) specifies the tradeoff between outage and rate at any SNR, but closed forms do not exist in general for (3). This led to the introduction of metrics whose tradeoff can be more succinctly expressed. In particular, the diversity order was introduced as a proxy for P_{out}. The traditional notion of diversity order equals the asymptotic slope of the outage-SNR curve (in log-log scale) for a fixed R. Although meaningful in early wireless systems, where R was indeed fixed, this is not particularly indicative of contemporary systems where R is increased with SNR. A more general formulation was introduced in [2], where R depends on SNR according to some function $R = f(\text{SNR})$. The diversity order

$$d = -\lim_{\text{SNR}\to\infty} \frac{\log P_{\text{out}}(\text{SNR}, f(\text{SNR}))}{\log \text{SNR}} \tag{5}$$

still captures the asymptotic slope of the outage-SNR curve (in log-log scale), albeit now for increasing R. A proxy for rate, termed the multiplexing gain, was further introduced as

$$r = \lim_{\text{SNR}\to\infty} \frac{f(\text{SNR})}{\log \text{SNR}}, \tag{6}$$

which is the asymptotic slope, in bits/s/Hz/(3 dB), of the rate-SNR curve.

The DMT (diversity-multiplexing tradeoff) specifies the (r, d) pairs that are achievable for $\text{SNR} \to \infty$, and thus characterizes the tradeoff between r and d [2]. For a quasi-static channel where each coded block is subject to a single fading realization, the DMT specifies that $\min(n_{\text{T}}, n_{\text{R}}) + 1$ distinct DMT points are feasible, each corresponding to a multiplexing gain $0 \leq r \leq \min(n_{\text{T}}, n_{\text{R}})$ and a diversity order

$$d(r) = (n_{\text{T}} - r)(n_{\text{R}} - r). \tag{7}$$

The full DMT frontier of achievable (r, d) pairs is obtained by connecting these points with straight lines.[1] Each transmit-receive architecture is associated with a DMT frontier, associated with the outage-rate relationship for that architecture.

The DMT governs the speed at which P_{out} decreases with SNR: if the rate grows as $r \log \text{SNR}$, then P_{out} decreases (ignoring sub-polynomial terms) as $\text{SNR}^{-d(r)}$. The DMT is thus a coarse description, through the proxies d and r, of the fundamental tradeoff between outage and rate. This coarseness arises from the definitions of r and d, which (*i*) are asymptotic, thereby restricting the validity of the insights to the high-power regime, and (*ii*) involve only the slopes of the outage-SNR and rate-SNR curves, ignoring constant offsets. Indeed, d does not suffice to determine P_{out} at a given SNR but simply quantifies the speed at which it falls with SNR. Similarly, r does not suffice to determine R, but it only quantifies how it grows with SNR.

Notice that $d(0)$ corresponds to the traditional notion of diversity order, i.e., with a fixed rate. A multiplexing gain $r = 0$ signifies a rate that does not increase (polynomially) with the SNR while $d = 0$ indicates an outage probability that does not decrease (polynomially) with the SNR.

IV. Modeling Modern Cellular Systems

Cellular systems have experienced major changes as they evolved. Besides MIMO, features of modern systems—that in many cases were completely absent in earlier designs—include:

- Wideband channelizations and OFDM.
- Powerful channel codes [3].
- Link adaptation, and specifically rate control via variable modulation and coding [4].
- ARQ (automatic repeat request) and H-ARQ (hybrid-ARQ) [5].

[1] If the coded block spans several fading realizations, then this additional time/frequency selectivity leads to larger diversity orders but does not increase the maximum value of r [2].

- Packet switching, with time- and frequency-domain scheduling for low-velocity users.

These features, in turn, have had a major impact on the operational conditions:

- There is a target block error probability, on the order of 1%, at H-ARQ termination. Link adaptation loops are tasked with selecting the rate in order to maintain performance tightly around this operating point.
- The channels of low-velocity users can be tracked and fed back to the transmitter thereby allowing for link adaptation to the supportable rate, scheduling on favorable time/frequency locations, and possibly precoding.
- The channels of high-velocity users vary too quickly in time to allow for feedback of CSI. Thus, the signals of such users are dispersed over the entire available bandwidth thereby taking advantage of extensive frequency selectivity. In addition, time selectivity is naturally available because of the high velocity.

Certain traditional wisdoms and models do not apply in these conditions, and this directly affects the role of transmit antenna diversity.

A. Low Velocity

At low velocities, timely feedback regarding the current state of the channel is feasible. This fundamentally changes the nature of the communication problem: all uncertainty is removed except for the noise. With powerful coding handling noise, outages are essentially eliminated. Transmit diversity techniques, whose goal is precisely to reduce outages, become inappropriate. Rate maximization becomes the overriding design principle.

In multiuser settings, furthermore, CSI feedback is collected from many users and scheduling offers additional degrees of freedom. In this case, transmit diversity techniques can actually be detrimental by reducing multiuser scheduling gains [6].

B. High Velocity

At high velocities, the fading is too rapid to be tracked. The link adaptation loops can therefore only match the rate to the average channel conditions. The scheduler, likewise, can only respond to average conditions and thus it is not possible to transmit only to users with favorable instantaneous channels; we thus need not distinguish between single-user and multiuser settings.

It is in this high-velocity scenario where transmit diversity is enticing. Frequency-flat analyses are likely to indicate that dramatic reductions in outage probability can be had by increasing d. On these grounds, transmission strategies that operate efficiently at the full-diversity DMT point have been developed. The value of these strategies for modern cellular systems, however, is questionable because:

1) The outage need not be reduced below the target error probability.
2) Diversity is plentiful already since *i)* by the same token that the fading is too rapid to be tracked, it offers time selectivity, and *ii)* in this regime, modern systems distribute the signals over large swaths of bandwidth thereby reaping abundant frequency selectivity.

Within the DMT framework, a fixed outage probability corresponds to $d = 0$, i.e., to the full multiplexing gain achievable by the architecture at hand. Thus, the R_ϵ-maximizing architectures for $\text{SNR} \to \infty$ are those that can attain the maximum multiplexing gain $r = \min(n_\text{T}, n_\text{R})$. Due to the nature of the DMT, however, this holds asymptotically in the SNR. The extent to which it holds for SNR values of interest in a selective channel can only be determined through a more detailed (non-asymptotic) study, such as the case study presented in the next section.

V. Case Study: A Modern MIMO-OFDM System

Let us consider the exemplary system described in Table I, which is loosely based on the 3GPP LTE design [7]. (With only slight modifications, this system could be made to conform with 3GPP2 UMB or with IEEE 802.16 WiMAX.) Every feature relevant to the discussion at hand is modeled:

- A resource block spans 12 OFDM tones over 1 ms. Since 1 ms corresponds to 14 OFDM symbols, this amounts to 168 symbols. In the high-

TABLE I
MIMO-OFDM SYSTEM PARAMETERS

Tone spacing	15 kHz
OFDM Symbol duration	71.5 μs
Bandwidth	10 MHz (600 usable tones)
Resource block	12 tones over 1 ms
H-ARQ	Incremental redundancy
H-ARQ transmission spacing	6 ms
Max. number H-ARQ rounds	6
Power delay profile	12-ray TU
Doppler spectrum	Clarke-Jakes
Max. Doppler frequency	185 Hz
Antenna correlation	None

velocity regime, the 12 tones are interspersed uniformly over 10 MHz of bandwidth.[2] There are 600 usable tones on that bandwidth and thus every 50th tone is allocated to the user at hand; the rest are available for other users.

- Every coded block spans up to 6 H-ARQ transmission rounds, each corresponding to a resource block, with successive rounds spaced by 6 ms for a maximum temporal span of 31 ms. The H-ARQ process terminates as soon as decoding is possible. An error is declared if decoding is not possible after 6 rounds.
- The channel exhibits continuous Rayleigh fading with a Clarke-Jakes spectrum and a 180-Hz maximum Doppler frequency. (This corresponds, for example, to 100 Km/h at 2 GHz.) The power delay profile is given by the 12-ray TU (typical urban) channel detailed in Table II. The r.m.s. delay spread equals 1 μs.
- The antennas are uncorrelated to underscore the roles of both diversity and multiplexing.

The impulse response describing each of the $n_{\rm T} n_{\rm R}$ entries of the channel matrix at each tone is

$$h(t,\tau) = \sum_{j=1}^{12} \sqrt{\alpha_j} c_j(t) \delta(t-\tau_j) \tag{8}$$

where the delays $\{\tau_j\}_{j=1}^{12}$ and the powers $\{\alpha_j\}_{j=1}^{12}$ are specified in Table II and $\{c_j(t)\}_{j=1}^{12}$ are independent complex Gaussian processes with a Clarke-Jakes spectrum.

Rate and outage were defined in Section II without H-ARQ. With H-ARQ, the length of each coded block is now variable. With IR (incremental redundancy) specifically, mutual information is accumulated over successive H-ARQ rounds [8]. If we let $\mathcal{M}_k(\mathsf{SNR})$ denote the mutual information after k rounds, then the number of rounds needed to decode a block is the smallest integer K such that

$$\mathcal{M}_K(\mathsf{SNR}) > 6\, R_\epsilon(\mathsf{SNR}) \tag{9}$$

where $K \leq 6$. A one-bit success/failure notification of is fed back after a decoding attempt following each H-ARQ round. An outage is declared if

$$\mathcal{M}_6(\mathsf{SNR}) \leq 6\, R_\epsilon(\mathsf{SNR}) \tag{10}$$

and the effective rate is

$$\mathcal{R}_\epsilon(\mathsf{SNR}) = \frac{6\, R_\epsilon(\mathsf{SNR})}{E[K]} \tag{11}$$

The initial rate is selected such that the outage at H-ARQ termination is precisely $\epsilon = 1\%$. This corresponds to choosing an initial rate of $6\, R_\epsilon$ where R_ϵ corresponds to (4) with the mutual information averaged over the 168 symbols within each H-ARQ round and then summed across the 6 rounds.

In order to contrast the benefits of transmit diversity and spatial multiplexing, we shall evaluate two representative transmission techniques:

- A transmit diversity strategy that converts the MIMO channel into an effective scalar channel with signal-to-noise ratio $\frac{\mathsf{SNR}}{n_{\rm T}} \, {\rm Tr}\{\mathbf{H}_i(k)\mathbf{H}_i^\dagger(k)\}$ where $\mathbf{H}_i(k)$ denotes the channel for the ith symbol on the kth H-ARQ round. By applying a strong outer code, the mutual information after k rounds is, at most, (14) at the bottom

TABLE II
TU POWER DELAY PROFILE

Delay (μs)	Power (dB)
0	-4
0.1	-3
0.3	0
0.5	-2.6
0.8	-3
1.1	-5
1.3	-7
1.7	-5
2.3	-6.5
3.1	-8.6
3.2	-11
5	-10

[2]For low velocity users, in contrast, the 12 tones are contiguous so that their fading can be efficiently described and fed back for link adaptation and scheduling purposes.

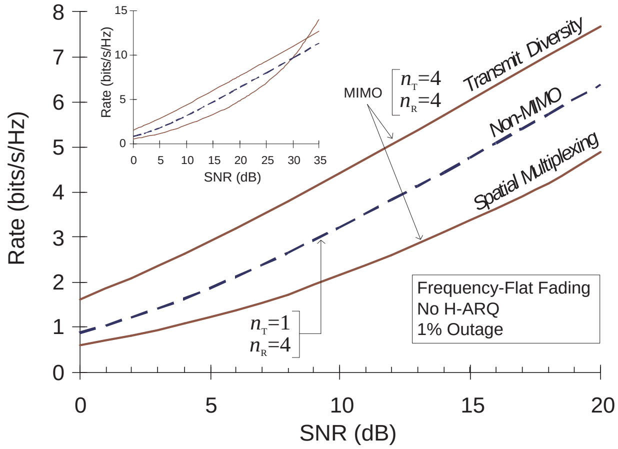

Fig. 1. Main plot: MMSE-SIC spatial multiplexing v. transmit diversity with $n_{\rm T} = n_{\rm R} = 4$ in a frequency-flat Rayleigh-faded channel with no H-ARQ. Also shown is the non-MIMO reference ($n_{\rm T} = 1$, $n_{\rm R} = 4$). Inset: Same curves over a wider SNR range.

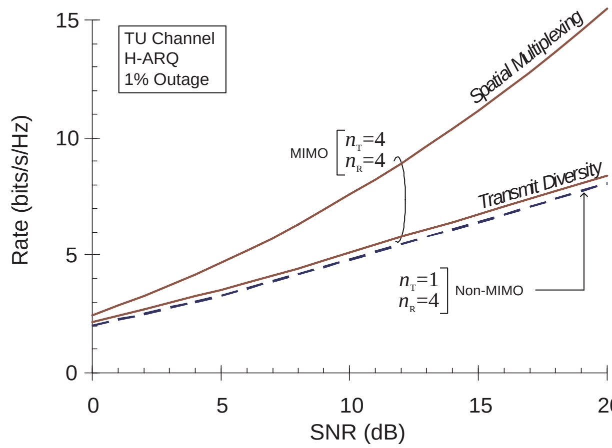

Fig. 2. MMSE-SIC spatial multiplexing v. transmit diversity with $n_{\rm T} = n_{\rm R} = 4$ in the channel described in Tables I–II. Also shown is the non-MIMO reference ($n_{\rm T} = 1$, $n_{\rm R} = 4$).

of the following page [2]. Transmit diversity strategies provide full diversity order, but their multiplexing gain cannot exceed $r = 1$, i.e., one information symbol for every vector $\mathbf{x}_i$ in (1). Note that, when $n_{\rm T} = 2$, (14) is achieved by Alamouti transmission [9].

- A basic MMSE-SIC spatial multiplexing strategy where a separate coded signal is transmitted from each antenna, all of them at the same rate [10]. The receiver attempts to decode the signal transmitted from the first antenna. An MMSE filter is applied to whiten the interference from the other signals and the first signal experiences a signal-to-noise ratio

$$\mathbf{h}_{i,1}^{\dagger}(k)\left(\mathbf{H}_{i,1}(k)\mathbf{H}_{i,1}^{\dagger}(k)+\frac{n_{\rm T}}{\mathsf{SNR}}\mathbf{I}\right)^{-1}\mathbf{h}_{i,1}(k)$$

during the kth H-ARQ round. If successful, the effect of the first signal is subtracted from the received samples and decoding of the second signal is attempted, and so forth. An outage is declared if any of the $n_{\rm T}$ coded signals cannot support the transmitted rate. The aggregate mutual information over the $n_{\rm T}$ antennas after k H-ARQ rounds is given by (15) at the bottom of the page, where $\mathbf{h}_{i,m}(\ell)$ is the mth column of $\mathbf{H}_i(\ell)$ while $\mathbf{H}_{i,m}(\ell) = [\mathbf{h}_{i,m+1}(\ell)\mathbf{h}_{i,m+2}(\ell)\cdots\mathbf{h}_{i,n_{\rm T}}(\ell)]$. While deficient in terms of diversity order, this strategy yields full multiplexing gain, $r = \min(n_{\rm T}, n_{\rm R})$, when $d = 0$. This MMSE-SIC structure is representative of the single-user MIMO mode in LTE [7].

Let $n_{\rm T} = n_{\rm R} = 4$ and consider first a simplistic model where the fading is frequency-flat and there is no H-ARQ. Every coded block is subject to a single fading realization. Under such model, the spectral efficiencies achievable with 1% outage, $\mathcal{R}_{0.01}(\mathsf{SNR})$, are compared in Fig. 1. Transmit diversity is uniformly superior to spatial multiplexing

$$\mathcal{M}_k(\mathsf{SNR}) = \sum_{\ell=1}^{k}\frac{1}{168}\sum_{i=1}^{168}\log\left(1+\frac{\mathsf{SNR}}{n_{\rm T}}\,\mathrm{Tr}\left\{\mathbf{H}_i(\ell)\mathbf{H}_i^{\dagger}(\ell)\right\}\right) \quad (14)$$

$$\mathcal{M}_k(\mathsf{SNR}) = n_{\rm T}\min_{m=1,\cdots,n_{\rm T}}\left\{\sum_{\ell=1}^{k}\frac{1}{168}\sum_{i=1}^{168}\log\left(1+\mathbf{h}_{i,m}^{\dagger}(\ell)\left(\mathbf{H}_{i,m}(\ell)\mathbf{H}_{i,m}^{\dagger}(\ell)+\frac{n_{\rm T}}{\mathsf{SNR}}\mathbf{I}\right)^{-1}\mathbf{h}_{i,m}(\ell)\right)\right\} \quad (15)$$

in the SNR range of interest. The curves eventually cross, as the DMT predicts and the inset in Fig. 1 confirms, (the asymptotic slope of spatial multiplexing is $r = 4$ bits/s/Hz/(3 dB) while $r = 1$ for transmit diversity and for non-MIMO), but this crossover does not occur until beyond 30 dB.

Still with $n_{\rm T} = n_{\rm R} = 4$, consider now the richer model in Tables I–II. The mutual information for each coded block is averaged over tones and symbols and accumulated over H-ARQ rounds. The corresponding comparison is presented in Fig. 2. In this case, transmit diversity offers a negligible advantage whereas spatial multiplexing provides ample gains with respect to non-MIMO.

The stark contrast between the behaviors observed under the different models is explained by the abundant time/frequency selectivity neglected by the simple model and actually present in the system. This renders transmit antenna diversity superfluous, not only asymptotically but at every SNR.

VI. Conclusion

While narrowband channelizations and non-adaptive links were the norm in cellular systems, antenna diversity was highly effective at mitigating fading. In modern systems, however, this is no longer the case. Link adaptivity and scheduling have rendered transmit diversity undesirable for low-velocity users whereas abundant time/frequency selectivity has rendered it superfluous for high-velocity users. Moreover, MIMO has opened the door for a much more effective use of antennas: spatial multiplexing.

Of all DMT points, therefore, the zero-diversity one stands out in importance. Techniques, even suboptimum ones, that can provide full multiplexing are most appealing to modern cellular systems whereas techniques that achieve full diversity order but fall short on multiplexing gain are least appealing. Our findings further the conclusion in [11], where a similar point is made solely on the basis of the multiplexing gain for frequency-flat channels.

The trend for the foreseeable future is a sustained increase in system bandwidth, which is bound to only shore up the above conclusion. At the same time, exceptions do exist, e.g., control channels.

Our study has only required evaluating well-known techniques under realistic models and at the appropriate operating points. Indeed, a more general conclusion that can be drawn is that, over time, the evolution of wireless systems has rendered some of the traditional models and wisdoms obsolete. In particular:

- Time/frequency selectivity should always be properly modeled.
- Performance assessments are to be made at the correct operating point, particularly in terms of error probability.
- The assumptions regarding transmit CSI must be consistent with the regime being considered. At low velocities, adaptive rate control based on instantaneous CSI should be incorporated; at high velocities, only adaptation to average channel conditions should be allowed.

References

[1] E. Biglieri, J. Proakis, and S. Shamai, "Fading channels: Information-theoretic and communication aspects," *IEEE Trans. Inform. Theory*, vol. 44, no. 6, pp. 2619–2692, Oct. 1998.

[2] L. Zheng and D. Tse, "Diversity and multiplexing: A fundamental tradeoff in multiple antenna channels," *IEEE Trans. Inform. Theory*, vol. 49, pp. 1073–1096, May 2003.

[3] G. D. Forney and D. J. Costello, "Channel coding: The road to channel capacity," *Proceedings of IEEE*, vol. 95, no. 6, pp. 1150–1177, June 2007.

[4] A. J. Goldsmith, "The capacity of downlink fading channels with variable rate and power," *IEEE Trans. Veh. Technol.*, vol. 46, pp. 569–580, Aug. 1997.

[5] D. Costello, J. Hagenauer, H. Imai, and S. B. Wicker, "Applications of error-control coding," *IEEE Trans. Inform. Theory*, vol. 44, no. 6, pp. 2531–2560, Oct. 1998.

[6] H. Viswanathan and S. Venkatesan, "The impact of antenna diversity in packet data systems with scheduling," *IEEE Trans. on Communications*, vol. 52, pp. 546–549, Apr. 2004.

[7] S. Sesia, I. Toufik, and M. Baker (Editors), *The UMTS Long Term Evolution: From Theory to Practice*, Wiley, 2009.

[8] G. Caire and D. Tuninetti, "The throughput of hybrid-ARQ protocols for the Gaussian collision channel," *IEEE Trans. Inform. Theory*, vol. 47, no. 5, pp. 1971–1988, Jul. 2001.

[9] S. M. Alamouti, "A simple transmit diversity technique for wireless communications," *IEEE J. Select. Areas Commun.*, vol. 16, pp. 1451–1458, Oct. 1998.

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

[11] B. Varadarajan and J. R. Barry, "The outage capacity of linear space-time codes," *IEEE Trans. Wireless Communications*, vol. 6, no. 6, pp. 2642–2648, Nov. 2005.