Dropping Users in a Multi-Antenna Broadcast Channel

Chau Yuen
Institute for Infocomm Research
cyuen@i2r.a-star.edu.sg

Bertrand M. Hochwald
Beceem Communications
hochwald@beceem.com

Abstract - We analyze the sensitivity of the capacity of a multi-antenna multi-user system to the number of users being served simultaneously. We show analytically that the capacity loss of serving fewer users at low SNR is much smaller than at high SNR. The advantages of serving fewer users are many: multi-user algorithms often have better performance, lower complexity, and reduced channel-knowledge requirements. We provide guidelines on how many users to serve to get near-capacity performance with low complexity. For example, we show that the performance gain in some algorithms more than compensates for the loss in capacity in an eight-antenna eight-user system when only four users are served, and we are now approximately 2dB from capacity at very low SNR.

I. Introduction

Given a base-station or access-point with M transmit antennas serving a pool of $L = M$ autonomous single-antenna users who cannot cooperate with each other, we wish to approach the sum-capacity, especially at low SNR. This is a multi-antenna multi-user system, which is also sometimes referred to as a Gaussian vector broadcast channel. The sum-capacity of such a system is investigated in [1-7], assuming that channel state information (CSI) is available at the transmitter and receivers.

Achieving capacity in this multi-antenna multi-user channel ostensibly could in theory require us to serve all $L = M$ users simultaneously with the M transmit antennas. We show that we may often serve fewer than M users with little capacity penalty especially at low SNR. We quantify the penalty in this paper.

We focus on low SNR's or low SINR's (signal to interference-plus-noise) because many multi-user systems operate in an interference-dominated environment. Furthermore, some multi-user techniques, such as vector-perturbation described in [8], are able to approach the sum-capacity at high SNR but suffer some performance loss at low SNR. It is then important to improve the performance of these algorithms at low SNR. One simple way to reduce complexity and improve per-user performance is simply to serve fewer users. But serving fewer users is a viable strategy only if we do not suffer a large total throughput penalty in the process. By computing the sensitivity of the capacity to the number of users being served, we quantify the loss and show that it is small at low SNR's.

We express the sensitivity of the capacity to the number of users as follows: Given that we are serving a certain number of users at a certain SNR, if we now serve fewer users with the same number of antennas, how much do we need to increase the SNR to obtain the same total throughput? The reduction in number of users is the equivalent of a complexity reduction (that comes from less CSI requirements and simpler algorithm), while the increase in the SNR is the equivalent of a power penalty. For example, for eight transmit antennas, at ρ = 1dB (which is defined as the ratio of total transmit power to per-user receive noise power), we suffer a penalty of only 1.14dB in SNR (increase in ρ) if we serve only four users, each at one-fourth the sum-capacity, versus all eight users, each at one-eighth this same sum-capacity. The corresponding penalty at ρ = 10dB is more than 2.55dB, and the corresponding penalty at ρ = 20dB is more than 7.63dB. We sometimes call this power penalty a "loss".

Thus, to obtain capacity at low SNR's we may as well serve fewer users at once, with algorithms that have low complexity and good performance. Continuing the example of the previous paragraph, we break the original pool of eight users into two random groups of four and serve these groups individually and suffer only a 1.14dB power penalty at ρ = 1dB. In return, we obtain the benefit of requiring the CSI of only four users at the transmitter at one time. Furthermore, we may obtain performance gains from any algorithm used to serve the four users: for example, the vector-perturbation technique gains approximately 1.5dB gain over the same technique serving eight users (in the eight-antenna system of this example). We formulate the problem in detail in the next section.

II. Sensitivity Analysis

A) Capacity versus number of users being served

Consider a multi-antenna multi-user communication system with M transmit antennas and M users, each with one antenna. The sum-capacity is [2]:

$$C_M = \mathrm{E} \max_{\mathbf{D}_M,\mathrm{tr}(\mathbf{D}_M)=1} \log\det\left(\mathbf{I}_M + \rho\mathbf{H}^*\mathbf{D}_M\mathbf{H}\right) \qquad (1)$$

where $\mathbf{I}_M$ is an $M\times M$ identity matrix, $\mathbf{H}$ is the $M\times M$ channel matrix between every user and the transmitter, and $\mathbf{D}_M$ is an $M\times M$ positive diagonal matrix whose trace is unity. The elements of $\mathbf{H}$ are Rayleigh fading (complex-Gaussian) coefficients with mean zero and unit variance. Throughout

this paper, log is base-two, and natural log is denoted by ln. We ignore the numerical and algorithmic issues in optimizing $\mathbf{D}_M$ in (1).

To simplify the analysis, suppose in (1) that we set

$$\mathbf{D}_M = \frac{1}{M}\mathbf{I}_M \tag{2}$$

and obtain the following lower bound:

$$C_M \geq I_{\text{rand},M} \tag{3}$$

where

$$I_{\text{rand},M} = \mathrm{E}\log\det\left(\mathbf{I}_M + \frac{\rho}{M}\mathbf{H}^*\mathbf{H}\right). \tag{4}$$

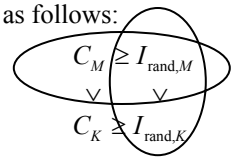
The suffix "rand" denotes that the users are chosen randomly from the pool and has operational meaning only when fewer than M are chosen.

The diagonal elements of $\mathbf{D}_M$ in (2) are related to the power assigned to different data streams of different users [2]. Rather surprisingly, we show that the lower bound in (4) (which assigns equal power to the data streams) is tight at both high and low ρ for large M. When $M = 1$, this lower bound is trivially tight. For large ρ it is shown in [5] that setting $\mathbf{D}_M = (1/M)\mathbf{I}_M$ is a good approximation. When ρ is small, we show in [9] that:

$$(1+\zeta)\, I_{\text{rand},M} \geq C_M \geq I_{\text{rand},M} \tag{5}$$

for large M and any $\zeta > 0$.

We next investigate the sensitivity of C_M to reducing the number of users K to a value less than M but keeping the number of antennas fixed at M. We call the resulting capacity C_K. The sensitivity of the capacity is determined by decreasing the number of users by a small amount and examining how much we must increase the SNR so as to keep the total capacity constant. Instead of working directly on the sum-capacity (which does not have a closed-form formula), we use the lower bound I and the general relationship is as follows:



(6)

where K is the design variable representing the number of users we serve with M antennas. We would like to find the difference between C_M and C_K as a function of K and ρ. We examine this gap indirectly as shown in equation (6) where we instead examine the gaps in the two circled inequalities. As shown in equation (5), we know that $I_{\text{rand},M}$ is a good approximation of C_M. Thus, we only need to investigate the gap between $I_{\text{rand},M}$ and $I_{\text{rand},K}$. If the difference between $I_{\text{rand},M}$ and $I_{\text{rand},K}$ is small, then we can ensure also that the difference between C_M and C_K is also small.

The approximation $I_{\text{rand},K}$ has the big advantage of a closed-form formula:

$$I_{\text{rand},K} = \mathrm{E}\log\det\left(\mathbf{I}_M + \frac{\rho}{K}\mathbf{H}^*\mathbf{H}\right) \approx K\mathrm{F}(\beta,\rho) \tag{7}$$

and $\mathrm{F}(\beta,\rho)$ is defined in [10] as:

$$\begin{aligned}\mathrm{F}(\beta,\rho) &= \frac{1}{\pi}\int_{\left(\sqrt{\beta}-1\right)^2}^{\left(\sqrt{\beta}+1\right)^2} \log(1+\rho\lambda)\sqrt{\frac{\beta}{\lambda} - \frac{1}{4}\left(1+\frac{(\beta-1)}{\lambda}\right)^2}\, d\lambda \\ &= \log\left(1+\rho\left(\sqrt{\beta}+1\right)^2\right) + (\beta+1)\log\left(\frac{1+\sqrt{1-a}}{2}\right) \\ &\quad -(\log e)\sqrt{\beta}\,\frac{1-\sqrt{1-a}}{1+\sqrt{1-a}} + (\beta-1)\log\left(\frac{1+\gamma}{\gamma+\sqrt{1-a}}\right)\end{aligned} \tag{8}$$

where $a = 4\rho\sqrt{\beta}\Big/\left(1+\rho\left(\sqrt{\beta}+1\right)^2\right)$, $\gamma = \left(\sqrt{\beta}-1\right)\Big/\left(\sqrt{\beta}+1\right)$, and where $\mathbf{H}$ is of dimension $K \times M$, and $\beta = M/K$.

The approximation of $I_{\text{rand},K} \approx K\mathrm{F}(\beta, \rho)$ above is valid when we consider a fixed β but both K and M are large [10]. The approximation makes the analysis tractable and is very accurate for even small values of M and K.

When we reduce the number of users from K to K', the ratio β increases to β', where $\beta' = M/K'$. In order to achieve the same capacity before this reduction, ρ must be increased to some ρ'. We define two quantities: ε the *complexity reduction coefficient,* and δ the *power penalty coefficient*. The quantity ε is defined as:

$$\varepsilon = \frac{\mathrm{d}\beta}{\beta} \quad \text{where} \quad \mathrm{d}\beta = \beta' - \beta \tag{9}$$

A large positive ε implies a large reduction in complexity. We are generally interested in infinitesimal changes, and an infinitesimal change in β is related to an infinitesimal change in K through

$$\varepsilon = \frac{\mathrm{d}\beta}{\beta} = -\frac{\mathrm{d}K}{K}. \tag{10}$$

Observe that ε divides the change in β by β. Hence, if $\varepsilon = 1$ the number of users is halved (i.e. $\beta' = 2\beta$). This is a notational convenience since it turns out that our final results are insensitive to the absolute number of antennas and users but are strong functions of the ratio β.

The power penalty coefficient δ is defined as:

$$\delta = \frac{\mathrm{d}\rho}{\rho} \quad \text{where} \quad \mathrm{d}\rho = \rho' - \rho \tag{11}$$

A large positive δ implies a large increase in SNR. The power penalty is related to the dB-change in power through

$$\begin{aligned}&\rho' = \rho(1+\delta) \\ &\Rightarrow \mathrm{d}\rho_{(\mathbf{dB})} = 10\log_{10}(1+\delta)\end{aligned} \tag{12}$$

We define the *sensitivity* as the ratio

$$\frac{\delta}{\varepsilon} = \frac{\mathrm{d}\rho/\rho}{\mathrm{d}\beta/\beta} = \frac{\mathrm{d}\rho}{\mathrm{d}\beta}\left(\frac{\beta}{\rho}\right) \tag{13}$$

where the changes in ρ and β are infinitesimal and such that the mutual information $I_{\text{rand},K}$ is kept constant. A small value for this ratio suggests that the number of users can be changed with little penalty in power. To obtain the sensitivity we solve

$$I_{\text{rand},K}(\rho) = I_{\text{rand},K'}(\rho') = \text{constant} \tag{14}$$

for infinitesimal changes in β and ρ.

Theorem 1: The sensitivity is:

$$\frac{\delta}{\varepsilon} = \frac{\mathrm{F}(\beta,\rho) - c_2(\beta,\rho)}{c_1(\beta,\rho)} \tag{15}$$

where c_1 and c_2 are

$$c_1(\beta,\rho) = a(\log e)\left[\frac{\left(\sqrt{\beta}+1\right)^2}{4\sqrt{\beta}} + d\right] \tag{16}$$

$$\begin{aligned} c_2(\beta,\rho) = {} & \beta \log\left(\frac{(1+\gamma)\left(1+\sqrt{1-a}\right)}{2\left(\gamma+\sqrt{1-a}\right)}\right) \\ & -(\log e)\frac{\left(\sqrt{\beta}-1\right)\left(1-\sqrt{1-a}\right)}{2\left(\gamma+\sqrt{1-a}\right)} \\ & -\frac{a(\log e)}{4}\left[\begin{array}{l} -\left(\sqrt{\beta}+1\right)+2d\left(\rho\beta-\rho-1\right) \\ +\dfrac{2\sqrt{\beta}}{\left(1+\sqrt{1-a}\right)^2} \end{array}\right] \end{aligned} \tag{17}$$

and

$$\begin{aligned} d = {} & \frac{\beta-1}{2\sqrt{1-a}\left(1+\rho\left(\sqrt{\beta}+1\right)^2\right)\left(\gamma+\sqrt{1-a}\right)} \\ & -\frac{\left(\sqrt{\beta}+1\right)^2+(\beta+1)\sqrt{1-a}}{2\sqrt{1-a}\left(1+\rho\left(\sqrt{\beta}+1\right)^2\right)\left(1+\sqrt{1-a}\right)^2} \end{aligned} \tag{18}$$

and a and γ are defined in (8). A proof of *Theorem 1* can be found in [9].

We notice that sensitivity δ/ε in (15) is a function of only ρ and β and is therefore "universal" in the sense that it does not depend on the specific values of the number of transmit antennas M and the number of users K but only their ratio.

The sensitivity is the ratio of incremental power to user reduction while achieving constant mutual information. A low value of δ/ε implies that the capacity is insensitive to the number of users being served; there is only a small penalty if we serve fewer users. On the other hand, a large value of δ/ε implies that the capacity is highly sensitive to the number of users being served.

Since the expression of sensitivity in (15) is rather complex, we look at some special cases and asymptotic results. For example, when $\beta = 1$ we obtain $\gamma = 0$ and we may simplify (15) to

$$\frac{\delta}{\varepsilon} = (\ln b)\left(1+\frac{b}{\rho}\right) - b\left(1+\frac{\rho}{b}\right) \tag{19}$$

where $b = \left(1+\sqrt{1+4\rho}\right)/2$.

We plot the sensitivity for $\beta = 1$ as given in (19) in Figure 1. One can see that the sensitivity is never negative because the mutual information is a non-decreasing function of the number of users being served. Furthermore, from Figure 1, the sensitivity for $\beta = 1$ can be separated into two regions, from $\rho = -40$dB to 0dB, the sensitivity δ/ε is small (i.e. $\delta/\varepsilon << 1$), but after ρ = 10dB δ/ε grows quickly. In Figure 1, we also plot the sensitivity for $\beta = 2, 4, 8$ based on (15). We observe from the figure that the sensitivity increases as β increases because a larger value of β means that we are already serving fewer users.

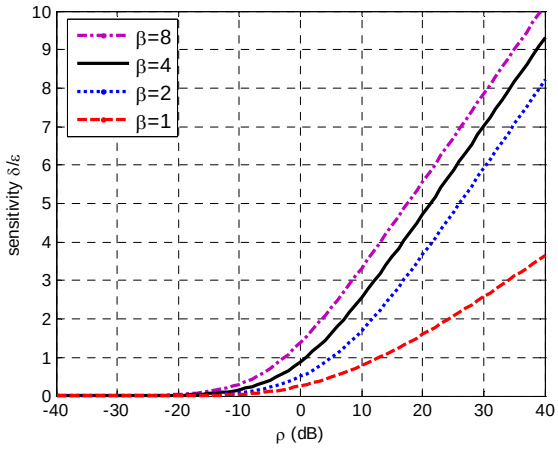


Figure 1: Sensitivity δ/ε for $\beta = 1, 2, 4, 8$. A low sensitivity means that we may reduce the number of users served with only a small power penalty.

We present the following asymptotics for the sensitivity:

$\rho \to 0$

$$\frac{\delta}{\varepsilon} = \frac{\beta\rho}{2} \tag{20}$$

$\rho \to \infty$

$$\frac{\delta}{\varepsilon} = \begin{cases} \ln(\rho)/2 - 1 & \text{if } \beta = 1 \\ \ln(\rho) + \ln(\beta-1) - 1 & \text{if } \beta > 1 \end{cases} \tag{21}$$

From (20) one can see that at low ρ the sensitivity is linear in ρ with slope $\beta/2$ and goes to zero as ρ goes to zero. This is seen in Figure 1. The effect of β is to increase the penalty multiplicatively as β increases, but since ρ is already small the effect is generally also small.

When ρ is large, (21) shows that there are two cases: $\beta = 1$ and $\beta > 1$. The effect of $\beta > 1$ is to shift the curve to

the left by $\ln(\beta - 1)$, which can be seen in Figure 1 in the curves for $\beta = 2$, 4, and 8 at high SNR.

B) Application of sensitivity

Suppose $K = M$ ($\beta = 1$). We ask: At what ρ we can reduce the number of users to be served by ½, ¼, or ⅛ (equivalently, make $\beta = 2$, 4, or 8) while suffering only ≈1 dB power loss?

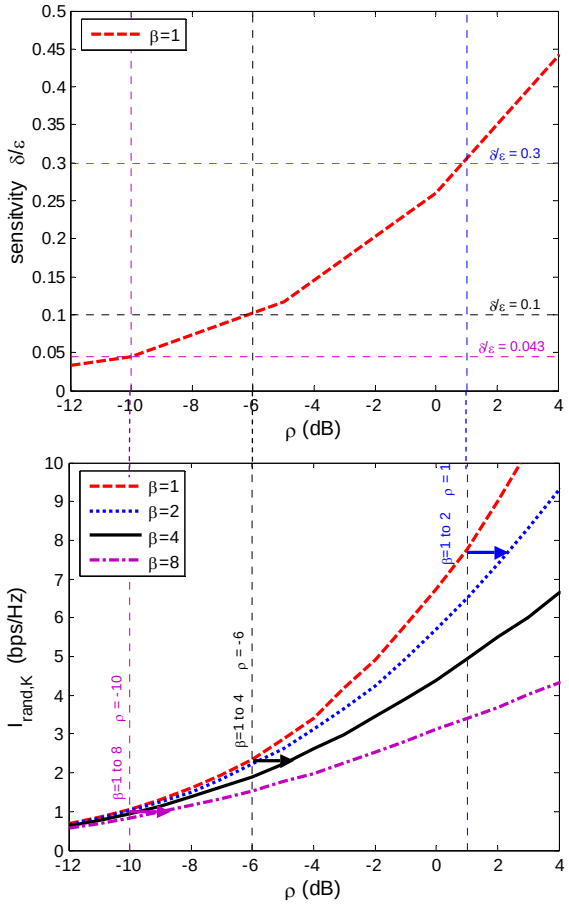


Figure 2: (upper figure) Sensitivity δ/ε for $\beta = 1$ showing the points where the power penalty is approximately 1.14 dB.

Figure 3: (lower figure) $I_{\text{rand,K}}$ for $M = 8$ with $\beta = 8, 4, 2, 1$ ($K = 1, 2, 4, 8$ users). These curves confirm that the power penalties at the operating points are slightly more than 1 dB.

The sensitivity δ/ε can be used to answer this question. The complexity reductions ε in (9) that correspond to changing β from 1 to 2, or from 1 to 4, or from 1 to 8 are $\varepsilon = 1$, 3, and 7. We choose the power penalty coefficient as $\delta = 0.3$ since we accept an estimated penalty of $10\log_{10}(1+0.3) \approx 1.14$dB. The corresponding sensitivities δ/ε for the three cases are then 0.3, 0.1, and 0.043.

We obtain the operating point ρ that yields these sensitivities by solving δ/ε in (19) with $\beta = 1$. Figure 2 is a plot of δ/ε versus ρ that shows our operating points. For example, the sensitivity is 0.3 when $\rho = 1$dB; this implies that when $\beta = 1$ there will be a loss of 1.14dB at $\rho = 1$dB when β is increased to 2. A similar loss is obtained at $\rho = -6$dB when β is increased to 4; at $\rho = -10$dB when β is increased to 8. These predicted values of ρ are tabulated in the second column of Table 1. Accepting a loss of approximately 1.14dB and with $M = 8$, we can therefore serve only $K = 4$ users ($\beta = 2$) at $\rho = 1$dB, $K = 2$ users ($\beta = 4$) at –6dB, $K = 1$ user ($\beta = 8$) at –10dB.

We verify these predicted operating points with the capacity curves, which are displayed in Figure 3. The actual power penalty when β is increased from 1 to 2, 4, or 8, corresponding to $\delta/\varepsilon = 0.3$, 0.1 and 0.043, is tabulated in the last column of Table 1. For example, Figure 3 shows that when $\rho = -6$dB, in order to achieve the same throughput as eight users ($\beta = 1$) with only two users ($\beta = 4$), about 1.30dB of extra power is needed. Similarly for the remaining two cases. Hence the exact penalties in ρ for three cases are close to the estimated values, but the estimated values are much more readily computed using the closed-form equation (19).

Table 1: Predicted ρ when penalty of approximately 1.14 dB is obtained as β is increased. Also shown is actual penalty at the predicted ρ when $M = 8$

Change in β	Predicted ρ from (19)	Actual penalty in ρ from Figure 3
1→ 2	1.00dB	1.40dB
1→ 4	-6.00dB	1.30dB
1→ 8	-10.00dB	1.30dB

III. Algorithm Performance Gain

The previous sections show that we can reduce the number of users at low SNR with only a small power penalty. We now show that reducing the number of users in some algorithms improve their performance sufficiently to overcome this penalty. We consider a system with $M = 8$ transmit antennas and $L = 8$ users at a total throughput of 8 bps/Hz. We use the vector-perturbation technique described in [8]. Throughout the simulations, it is assumed that rate-half turbo codes from the UMTS standard with feedforward polynomial $1+D+D^3$, feedback polynomial $1+D^2+D^3$, block length 10,000 bits, 20 inner iterations, and 8 outer loop iterations is used [11].

The vector-perturbation technique from [8] is summarized as follows:

$$\mathbf{y} = \frac{1}{\sqrt{\gamma}}\mathbf{HG}\left(\mathbf{u} + \tau \boldsymbol{l}\right) + \mathbf{n} \tag{22}$$

$$\mathbf{G} = \mathbf{H}^*\left(\mathbf{H}\mathbf{H}^* + \frac{\alpha}{\rho}\mathbf{I}\right)^{-1} \tag{23}$$

$$\tau = 2.5\left(|c|_{\max} + \Delta/2\right) \tag{24}$$

where **y** is received signal vectors for all K users, **H** is channel matrix, **u** and **n** are the signal and noise vectors for K users, **G** is the regularized-inverse precoding matrix, γ is a scalar such that total transmission power is normalized to one, α is the regularized-inverse parameter, $|c|_{\max}$ is the absolute value of the constellation symbol with largest magnitude, and Δ is the spacing between constellation points. The *perturbation vector* $\boldsymbol{l}$, which consists of only integer components, is obtained from the following optimization:

$$\boldsymbol{l} = \arg\min_{\boldsymbol{l}'} \left\{ \left(\mathbf{u} + \tau \boldsymbol{l}'\right)^* \left(\mathbf{G}^*\mathbf{G}\right) \left(\mathbf{u} + \tau \boldsymbol{l}'\right) \right\} \qquad (25)$$

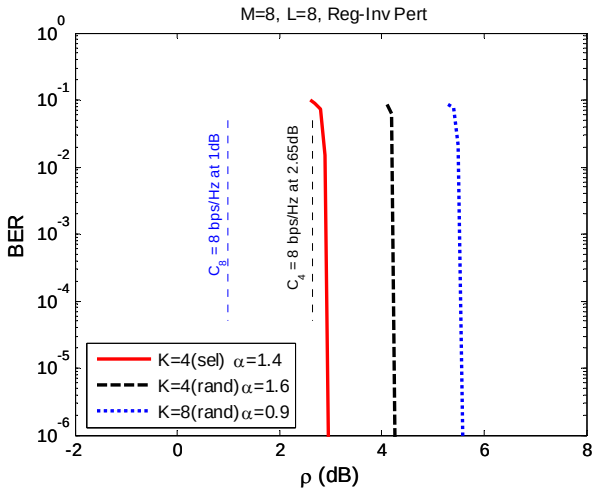


Figure 4: BER performance for the case of $M = 8$ when serving $K = 4$ (dashed line) or $K = 8$ (dotted line) users at a throughput of 8 bps/Hz. Also shown is the advantage of allowing the four users to be selected (solid line)

The BER performance of $M = 8$, $L = 8$, $K = 4$ or 8 users with regularized-inverse perturbation technique at a total throughput of 8 bps/Hz is shown in Figure 4. We observe from Figure 4 that by serving $K = 4$ users chosen randomly from the pool of 8 (the black dashed line, where each user uses a 16QAM constellation, corresponding to a coded rate of 2bps/Hz per user) can be much better than serving all $K = 8$ users (the blue dotted line, where each user uses a QPSK constellation, corresponding to a coded rate of 1bps/Hz per user). In this particular example, the overall performance gain is about 1.5dB despite the fact that serving 4 users has a power penalty of approximately 1.65dB (C_8 vs C_4). This illustrates the advantages of serving fewer users at low SNR, as the algorithm performance improvement outweighs the power penalty. Furthermore, to serve four users, we only require the CSI of any four users at a time, instead of all eight users, and the algorithm complexity decreases greatly.

One can also serve $K = 4$ selected users (the red solid line, where each user uses a 16QAM constellation). The selection process chooses the set of four users out of eight that maximize the mutual information between the transmitter and the receiver. As shown in Figure 4, selection has the best performance and is only 2dB away from C_8. The selection process, however, necessarily requires the CSI of all eight users.

IV. Conclusion

We proposed a sensitivity-based method for reducing the number of users served in a multi-antenna multi-user system at low SNR. We showed that the power penalty for serving a smaller number of users can be small and can, in fact, be overcome by both performance improvements and complexity reductions in algorithms that serve these users.

V. References

[1] G. Caire and S. Shamai, "On the achievable throughput of a multiantenna Gaussian broadcast channel", *IEEE Trans. on Information Theory*, vol. 49, pp. 1691-1706, July 2003.

[2] P. Viswanath and D. N. C. Tse, "Sum capacity of the vector Gaussian broadcast channel and uplink-downlink duality", *IEEE Trans. on Information Theory*, vol. 49, pp. 1912-1921, Aug. 2003.

[3] S. Vishwanath, N. Jindal, and A. Goldsmith, "Duality, achievable rates, and sum-rate capacity of Gaussian MIMO broadcast channels", *IEEE Trans. on Information Theory*, vol. 49, pp. 2658-2668, Oct. 2003.

[4] W. Yu and J. M. Cioffi, "Sum capacity of Gaussian vector broadcast channels", *IEEE Trans. on Information Theory*, vol. 50, pp. 1875-1892, Sept. 2004.

[5] N. Jindal, "High SNR analysis of MIMO broadcast channels", *IEEE ISIT 2005*.

[6] N. Jindal, S. Vishwanath, and A. Goldsmith, "On the duality of Gaussian multiple-access and broadcast channels", *IEEE Trans. on Information Theory*, vol. 50, pp. 768-783, May 2004.

[7] H. Weingarten, Y. Steinberg, and S. Shamai, "The capacity region of the Gaussian MIMO broadcast channel", *accepted for publication in IEEE Trans. on Information Theory*.

[8] B. M. Hochwald, C. B. Peel, and A. L. Sindlehurst, "A vector-perturbation technique for near-capacity multi-antenna multi-user communication – part II: perturbation", *IEEE Trans. on Communications*, vol. 53, pp. 537-544, March 2005.

[9] C. Yuen and B. M. Hochwald, "Achieving capacity of multi-antenna multi-user communication system at low SNR", *in preparation*.

[10] B. M. Hochwald, T. L. Marzetta, and B. Hassibi, "Space-time autocoding", *IEEE Trans. on Information Theory*, vol. 47, pp. 2761-2781, Nov 2001.

[11] B. M. Hochwald and S. T. Brink, "Achieving near-capacity on a multiple-antenna channel", *IEEE Trans. on Communications*, vol. 51, pp. 389-399, March 2003.