TOTAL CAPACITY OF MULTIACCESS VECTOR CHANNELS

by

Pramod Viswanath and Venkat Anantharam

Memorandum No. UCB/ERL M99/47

24 May 1999

TOTAL CAPACITY OF MULTIACCESS VECTOR CHANNELS

by

Pramod Viswanath and Venkat Anantharam

Memorandum No. UCB/ERL M99/47

24 May 1999

ELECTRONICS RESEARCH LABORATORY

College of Engineering
University of California, Berkeley
94720

Total Capacity of Multiaccess Vector Channels

Pramod Viswanath and Venkat Anantharam*
{pvi, ananth}@eecs.berkeley.edu
EECS Department, U C Berkeley CA 94720

24 May 1999

Abstract

The well known waterfilling power allocation policy maximizes the sum capacity of parallel Gaussian channels. We consider multiaccess vector channels with additive colored Gaussian noise and asymmetric user power constraints and completely characterize the sum capacity of this channel. We show that the sum capacity of the multiaccess vector channel is upper bounded by that of corresponding parallel Gaussian channels and that our solution (optimal powers and user signal directions) has a waterfilling structure. Two common examples in a wireless communication system that fall under this model are direct sequence code division multiaccess and multiaccess channels with multiple antennas at the receiver. The multiaccess vector channel models communication from users within a cell to the base station and interference from those users communicating with neighboring base stations is modeled by additive colored noise. Our characterization of the sum capacity allows us to conclude a *Schur-saddle* property: the sum capacity is *Schur convex* in the additive noise covariance and *Schur-concave* in the user powers.

1 Introduction

This paper considers multiaccess vector channels: multiaccess channels where users have *multiple degrees of freedom.* Two multiaccess channels that fall under the purview of this

*Research of the authors is supported by the National Science Foundation under under grant IRI 97-12131.

model and are commonly used in wireless communication systems are Direct-Sequence Code Division MultiAccess (DS-CDMA) and Space Division MultiAccess (SDMA) which is a multiaccess channel with multiple antennas at the receiver. The number of degrees of freedom in DS-CDMA model is the processing gain and in the antenna model it is the number of antennas at the receiver. The *signal direction* at the receiver of any user in the CDMA model is its *spreading sequence* (assuming frequency flat fading) and in the antenna model it is the vector of path gains from the user to the different antennas at the receiver. We assume colored additive noise in the multiaccess channel independent of the users, which models interference from users talking to neighboring base stations. This assumption means that the interference from outside the cell cannot be controlled but can be measured and estimated statistically. We restrict ourselves to the case when the users are symbol synchronous. A fundamental performance measure of multiaccess channels is *sum capacity* (equivalently spectral efficiency), defined as the maximum sum of rates of users per unit degree of freedom at which the users can transmit reliably. Our focus in this paper will be to identify the largest possible sum capacity as the signal directions of the users vary. In the DS-CDMA model the corresponding signal directions will be *optimal signature sequences* of the users, optimal in the sense of yielding the largest sum capacity of the multiaccess channel.

In previous work, the capacity region of a symbol synchronous DS-CDMA channel was characterized in [6]. The problem of characterizing the maximum sum capacity of DS-CDMA channels with white additive noise was first attempted in [4], which solved the symmetric user power constraint case. In [7] the general case of asymmetric user powers with additive white noise was solved and a simple recursive algorithm was provided to construct the corresponding optimal signature sequences. In the SDMA model with white additive noise, [5] characterized the capacity region and also obtained an expression for the maximum sum capacity when user powers are symmetric. The results in [7] covers the asymmetric user power constraints case for the SDMA model (with white additive noise) as well. In this paper we completely characterize the maximum sum capacity of the multiaccess vector channel with asymmetric user powers and colored additive noise.

We first show that the sum capacity of the multiaccess vector channel is upper bounded by the sum capacity of a sequence of parallel Gaussian channels. Here the number of parallel channels equals the number of degrees of freedom of the vector channel. The additive Gaussian noise variances in the parallel channels are the eigenvalues of the covariance matrix of the additive colored noise in the vector channel. Also the power available for the transmitter in the parallel Gaussian channels is equal to the sum of the power constraints of the users. Our characterization of sum capacity and the optimal user signal directions and powers in the multiaccess vector channel has the following waterfilling interpretation. We refer to the directions of eigenvectors of the covariance matrix of the additive colored noise as "channels".

1. If all the user power constraints are "not too far apart" *relative* to the problem size (number of degrees of freedom and number of users) and the additive noise variances then the user signal directions are chosen so that the power in the channels follows the waterfilling strategy. This is a necessary and sufficient condition for the sum capacity of the multiaccess vector channel to equal the sum capacity of the corresponding parallel Gaussian channels.

2. If one of the user power constraints is large *relative* to the problem size and the other user power constraints and the additive noise variances, that user alone should be in the channel with the *smallest* noise variance (this is achieved when that that user has signal direction equal to the eigenvector of the covariance matrix of the additive colored noise corresponding to the smallest eigenvalue). Then the problem reduces to that of one lesser user and one lesser degree of freedom.

3. If one of the noise variances is *relatively* large, then the user signal directions are chosen orthogonal to that direction. The problem then reduces to that of one lesser degree of freedom.

Our solution makes the notions of "not too far apart" and "relatively large", in the description above, precise. This characterization of the optimal user signal directions allows to derive the following qualitative properties of the sum capacity:

1. For fixed user power constraints, the sum capacity is convex in the additive noise covariance matrices. Furthermore, for fixed eigenvectors of the noise covariance matrix, the sum capacity *increases* when the eigenvalues of the covariance matrix are "spread apart" while maintaining the same average noise variance.

2. For fixed additive noise variances, the sum capacity is concave in the user power constraints. Furthermore, the sum capacity *decreases* when the user powers are "spread apart" while maintaining the same total user power constraint.

The precise notion of "spread apart" is provided by the *majorization partial order* on real vectors and we provide a brief summary of this partial order in Appendix A. We explain briefly the multiaccess vector channel model and the problem formulation in Section 2. Section 3 contains our main results on the characterization and properties of the maximum sum capacity. Section 4 contains the proofs of our main results and we conclude with a summary of the optimal user signal directions in Section 5.

2 Multiaccess Vector Channels and Sum Capacity

2.1 Model

There are K users in the channel and N denotes the number of degrees of freedom. In the DS-CDMA model, this means that the processing gain is N, and in the antenna model it means that the number of antennas at the receiver is N. Both K and N will be fixed throughout this paper. A baseband representation of the received signal in one symbol interval at the receiver is

$$Y = \sum_{i=1}^{K} s_i X_i + Z \tag{1}$$

where $s_1, \ldots, s_K$ are the received signal directions of the users, thought of as elements of $\mathbb{R}^N$. In the CDMA model, s_i is the received signature sequence of user i and in the antenna model it is the vector of path gains from user i to each of the antennas. We assume that the received signal directions are normalized such that the energy per unit degree of freedom is unity, i.e., $s_i^t s_i = 1$. The input symbols are represented by independent real random variables $X_1, \ldots, X_K$. The received user powers are given by $\mathrm{Var}(X_i) = p_i$ for each user i. Z is an additive colored Gaussian noise vector with covariance matrix Σ with positive eigenvalues $\sigma_1^2 \leq \sigma_2^2 \leq \cdots \leq \sigma_N^2$.

2.2 Problem Formulation

Fix the signal directions $s_1, \ldots, s_K$. The sum capacity of the vector MAC in (1) is

$$\max I(X_1, \ldots, X_K; Y)$$

where the maximum of mutual information between the inputs and the output vector Y is over all distributions $F_1, \ldots, F_K$ on $\mathbb{R}$ with variances upper bounded by $p_1, \ldots, p_K$ respectively. Proceeding as in [4], we see that this maximum is achieved when the distributions of all the random variables are Gaussian and thus we arrive at the following generalization of the result of [6] for colored additive noise. The sum capacity in nats per unit degree of freedom of the multiaccess channel in (1) is given by

$$C_{sum}(S, D, \Sigma) = \frac{1}{2N} \log \det \left(I + \Sigma^{-1} S D S^t \right) \tag{2}$$

where we have written $S = [s_1, \ldots, s_K]$ and $D = \mathrm{diag}\{p_1, \ldots, p_K\}$. Our main focus in this paper is to characterize the maximum sum capacity:

$$C_{opt}(D, \Sigma) \stackrel{\mathrm{def}}{=} \max_{S \in \mathcal{S}} C_{sum}(S, D, \Sigma) \tag{3}$$

where $\mathcal{S}$ is the set of all $N \times K$ real matrices with all columns having l_2 norm equal to 1. Since $\forall S \in \mathcal{S}$, we have $QS \in \mathcal{S}$ for every orthonormal matrix Q, it follows from the structure of C_{sum} in (2) that $C_{opt}(D, \Sigma)$ depends only on the eigenvalues of Σ.

2.3 Parallel Gaussian Channels

Consider the following parallel Gaussian channels (our notation is from Section 10.4 of [1]):

$$Y_j = X_j + Z_j, \;\; Z_j \sim \mathcal{N}\left(0, \sigma_j^2\right) \;\; j = 1 \ldots N$$

where the Gaussian noise is independent from channel to channel. The total power constraint on the input is $\mathbb{E}\left[\sum_{j=1}^N X_j^2\right] \leq P$. Denoting the sum capacity (the maximum sum of rates per unit channel at which all information can be transmitted in each of the channels reliably) of this channel by $C_p(P, \Sigma)$ we have

$$C_p(P, \Sigma) = \max_{\left\{(\eta_1, \ldots, \eta_N) \in \mathbb{R}_+^N : \sum_{j=1}^N \eta_j = P\right\}} \frac{1}{2N} \sum_{j=1}^N \log\left(1 + \frac{\eta_j}{\sigma_j^2}\right) . \tag{4}$$

For notational ease, we have represented the noise variances by a covariance matrix Σ that has eigenvalues $\sigma_1^2, \ldots, \sigma_N^2$. It is very well known that the optimal allocation of powers follows the waterfilling policy $\mathbb{E}\left[X_j^2\right] = \eta_j^* = \left(\beta - \sigma_j^2\right)^+$ for some $\beta > 0$ such that $\sum_{j=1}^N \eta_j^* = P$. A further explicit expression for the waterfilling policy is as follows: Define the set $\mathcal{K}_{wf}$ to be

$$\left\{ k : \sigma_k^2 > \frac{P + \sum_{j=1}^N \sigma_j^2 1_{\{\sigma_k^2 > \sigma_j^2\}}}{N - \sum_{j=1}^N 1_{\{\sigma_j^2 \geq \sigma_k^2\}}} \right\} .$$

Observe that if $k \in \mathcal{K}_{wf}$ then every l such that $\sigma_l^2 \geq \sigma_k^2$ also belongs to $\mathcal{K}_{wf}$. Since we have ordered the variances $\sigma_1^2 \leq \sigma_2^2 \leq \cdots \leq \sigma_N^2$, the set $\mathcal{K}_{wf}$ is of the form $\{k, \ldots N\}$ for some $1 < k \leq N+1$ (if $k = N+1$, then by convention we take $\mathcal{K}_{wf}$ to be empty). The waterfilling policy is simply

$$\eta_l^* = 0, \; k \leq l \leq N, \text{ and } \eta_j^* = \frac{P + \sum_{i=1}^{k-1} \sigma_i^2}{k-1} - \sigma_j^2, \;\; 1 \leq j < k . \tag{5}$$

Then (4) becomes

$$C_p(P, \Sigma) = \frac{1}{2N} \sum_{j=1}^{k-1} \log\left(\frac{P + \sum_{i=1}^{k-1} \sigma_i^2}{(k-1)\,\sigma_j^2}\right) .$$

On the other hand, for every $S \in \mathcal{S}$, the sum capacity of the multiaccess vector channel (1) is

$$\begin{aligned} C_{sum}(S, D, \Sigma) &= \frac{1}{2N} \log \det \left(I + \Sigma^{-1} S D S^t \right) \\ &= \frac{1}{2N} \log \det \left(I + \operatorname{diag} \left\{ \sigma_1^{-2}, \ldots, \sigma_N^{-2} \right\} Q S D S^t Q^t \right) \quad \text{for some orthogonal } Q \\ &\leq \frac{1}{2N} \sum_{i=1}^{N} \log \left(1 + \frac{d_i}{\sigma_i^2} \right) \end{aligned}$$

where we have denoted the diagonal entries of $QSDS^tQ^t$ by $d_1, \ldots, d_n$ and used the Hadamard inequality in the derivation of the last step. Since $\sum_{j=1}^{N} d_i = \operatorname{tr} SDS^t = \operatorname{tr} D$, comparing with (4) we arrive at the following simple upper bound to $C_{opt}(D, \Sigma)$:

$$C_{opt}(D, \Sigma) \leq C_p(\operatorname{tr} D, \Sigma) \ .$$

Though this upper bound is completely expected, our characterization of the sum capacity $C_{opt}(D, \Sigma)$ shows that if the user powers are not too "spread apart" (in a relative sense that depends on the size of the problem and Σ) the upper bound can actually be attained. We identify necessary and sufficient conditions on the user power constraints and the eigenvalues of the noise covariance matrix so that this upper bound is met.

3 Sum Capacity Characterization

Our main result is the solution of the optimization problem in (3) and thus the characterization of C_{opt}. Our solution completely characterizes the structure of the *optimal* signal directions (the S that achieve the maximum in (3)) and we also provide a combinatorial algorithm that explicitly constructs the optimal signature sequences. We also analyze properties of C_{opt} as a function of D and Σ and give intuitive explanations for these properties.

3.1 Solution of the optimization problem (3)

To begin with, observe that the dimension of the linear span of the signal directions is at most K, and thus if $N > K$ we should always restrict the signal directions to that subspace (of dimension at most K) which contains the eigenvectors of Σ corresponding to the smallest K eigenvalues. Hence, without loss of generality, we assume that $K \geq N$. We begin with the following definition.

Definition 3.1 *For any* $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$, *let*

$$x_{[1]} \geq \cdots \geq x_{[n]}$$

denote the components of x *in decreasing order, called the* order statistics *of* x.

Towards solving the optimization problem (3) we first make a (natural) change of variables from $S \in \mathcal{S}$ to the vector of eigenvalues (denoted by μ) of $SDS^t + \text{diag}\{\sigma_1^2, \ldots, \sigma_N^2\}$. With this change of variable, we claim the following:

Lemma 3.1

$$C_{opt}(D, \Sigma) = \max_{\mu \in \mathcal{L}} \frac{1}{2N} \sum_{i=1}^{N} \log\left(\frac{\mu_i}{\sigma_i^2}\right) \tag{6}$$

where $\mathcal{L}$ *is a polyhedron in the positive orthant of* $\mathbb{R}^N$ *defined by*

$$\mathcal{L} = \left\{ (\mu_1, \ldots, \mu_N) : \begin{array}{c} \mu_i \geq \sigma_i^2, \ \forall i = 1 \ldots N \\ \sum_{j=1}^{i} \mu_j \geq \sum_{j=1}^{i} p_{[j]} + \sigma_j^2, \ \forall i = 1 \ldots N-1 \\ \sum_{j=1}^{N} \mu_j = \sum_{j=1}^{K} p_j + \sum_{j=1}^{N} \sigma_j^2 \end{array} \right\}. \tag{7}$$

We relegate the proof of this result to the next section. We now proceed to give a combinatorial algorithm that solves the optimization problem (6) in at most N steps, and propose the following greedy algorithm $\mathcal{A}$ on $\mathcal{L}$. Recall that we have ordered the eigenvalues of Σ as $\sigma_1^2 \leq \sigma_2^2 \leq \cdots \leq \sigma_N^2$.

Input $K, N, (p_1, \ldots, p_K)$ and $(\sigma_1^2, \ldots, \sigma_N^2)$.

Output $\mu^* \in \mathcal{L}$.

Update

1. Initialization: $i = 1, j = N$ and $\mu_k^* = 0, \ \forall k = 0 \ldots N$.
2. Termination: If $i \geq j$ stop and output the vector $(\mu_1^*, \ldots, \mu_N^*)$. Else, go to Step 3.
3. Let

$$\eta = \max\left\{ \sigma_j^2, \frac{\sum_{k=1}^{K} p_{[k]} + \sum_{m=1}^{N} \sigma_m^2 - \sum_{m \notin \{i,\ldots,j\}} \mu_m^*}{j-i+1}, \frac{1}{l} \sum_{k=0}^{l-1} \left(p_{[i+k]} + \sigma_{i+k}^2 \right), i \leq l \leq j \right\} \tag{8}$$

   (a) If $\eta = \sigma_j^2$ then set $\mu_j^* := \sigma_j^2$ and $j := j - 1$. Go to Step 2.

   (b) If $\eta = \frac{\sum_{k=1}^{K} p_{[k]} + \sum_{m=1}^{N} \sigma_m^2 - \sum_{m \notin \{i,\ldots,j\}} \mu_m^*}{j-i+1}$ then set $\mu_m^* := \eta, \forall m = i \ldots, j$ and $i := j$. Go to Step 2.

(c) If $\eta = \frac{1}{l}\sum_{k=0}^{l-1}\left(p_{[i+k]} + \sigma^2_{i+k}\right)$ for some $l \in \{i, \ldots, j\}$ then set $\mu^*_m := \eta, \forall m = i, \ldots, l$ and $i := l + 1$. Go to Step 2.

Our formal claim on the behavior of this algorithm is below. We delay the proof of this result to the next section and now analyze some properties of the algorithm $\mathcal{A}$ that shed some insight into the structure of μ^*, the vector of eigenvalues of $S^*DS^{*t} + \text{diag}\{\sigma^2_1, \ldots, \sigma^2_N\}$. We conclude this section with some properties of the maximum sum capacity $C_{opt}(D, \Sigma)$.

Theorem 3.1 *The output μ^* of the combinatorial algorithm $\mathcal{A}$ achieves the maximum in (6).*

3.2 Properties of Algorithm $\mathcal{A}$

1. The function $(\mu_1, \ldots, \mu_N) \mapsto \sum_{i=1}^N \log \frac{\mu_i}{\sigma^2_i}$ is concave and thus the optimal solution μ^* is the vector with components "least spread out" in $\mathcal{L}$.

2. The algorithm stops after at most N steps. Since at least one component of μ^* is updated in each step, this observation is completely clear.

3. The updates of the components of μ^* by $\mathcal{A}$ are in non-increasing order. Hence $\mathcal{A}$ is a *greedy* algorithm in the sense that the algorithm first sets the largest component of μ to the smallest value it can attain and then reduces the problem to one lesser dimension.

4. The special case of $\Sigma = \sigma^2 I$ was addressed in [7] and algorithm $\mathcal{A}$ reduces to the following simple form (Section 3, [7]): Define the set $\mathcal{K} \subset \{1, \ldots, N\}$ to be

$$\left\{k : p_k > \frac{\sum_{j=1}^K p_j 1_{\{p_k > p_j\}}}{N - \sum_{j=1}^K 1_{\{p_j \geq p_k\}}}\right\}.$$

It follows that if $k \in \mathcal{K}$ then for every user l with power constraint $p_l \geq p_k$ also belongs to $\mathcal{K}$. The optimal solution μ^* is simply

$$\mu^*_l = p_l + \sigma^2, l \in \mathcal{K} \text{ and } \mu^*_j = \frac{\sum_{i=1}^{k-1}(p_i + \sigma^2)}{N - |\mathcal{K}|}, j \notin \mathcal{K}.$$

The physical intuition is that for every $k \in \mathcal{K}$, the user k is *oversized*, i.e., its power is large *relative* to the power constraints of the other users and the degrees of freedom. Every oversized user is given an independent channel (In the DS-CDMA context, this is done by allocating oversized users signature sequences that are orthogonal to all the other signature sequences).

5. In the special case when $D = pI$, then again, algorithm $\mathcal{A}$ has a simple structure. Observe that in this case, Case 3(c) of the algorithm will never be reached and this makes the algorithm have the following simple form. Define the set $\mathcal{K}$ to be

$$\left\{ k : \sigma_k^2 > \frac{Kp + \sum_{j=1}^{N} \sigma_j^2 1_{\{\sigma_k^2 > \sigma_j^2\}}}{N - \sum_{j=1}^{N} 1_{\{\sigma_j^2 \geq \sigma_k^2\}}} \right\} .$$

Observe that if $k \in \mathcal{K}$ then every l such that $\sigma_l^2 \geq \sigma_k^2$ also belongs to $\mathcal{K}$. Thus $\mathcal{K}$ is of the form $\{k, \ldots, N\}$ for some $1 < k \leq N+1$ (by convention $k = N+1$ denotes $\mathcal{K}$ to be empty). The algorithm $\mathcal{A}$ simply outputs

$$\mu_l = \sigma_l^2, l \leq k \leq N \text{ and } \mu_j = \frac{Kp + \sum_{i=1}^{k-1} \sigma_i^2}{k-1}, \ 1 \leq j < k .$$

The physical intuition is that for every $k \in \mathcal{K}$, the "channel" (the direction specified by the eigenvector of Σ corresponding to the eigenvalue σ_k^2) k is *oversized* and has noise variance σ_k^2 large *relative* to the other noise variances and the number of users and degrees of freedom in the MAC. We recognize that the simple form of $\mathcal{A}$ coincides with the waterfilling policy in (5). Thus the case of $D = pI$ is a sufficient condition for the sum capacity C_{opt} to be equal to the corresponding parallel Gaussian channels sum capacity C_p. A formal statement that identifies a necessary and sufficient condition will be made in Section 5.

3.3 Properties of $C_{opt}(D, \Sigma)$

Consider a multiaccess channel with additive white noise and variance σ^2. Suppose we make one of the noise variances, say σ_N^2 much larger than the rest while keeping the average of the variances equal to σ^2. The users can avoid using signals in the direction of the eigenvector of Σ corresponding to the large eigenvalue σ_N^2 and benefit from a reduced average noise variance (since the overall average noise variance is still σ^2). Thus we expect that the maximal sum capacity of the latter channel will be more than the maximal sum capacity of the additive white noise channel. We make this intuitive idea precise in the following proposition.

Proposition 3.1 *Fix D, the diagonal matrix of user powers. Then $C_{opt}(D, \Sigma)$ is convex in Σ. Also, $C_{opt}(D, \Sigma) > C_{opt}\left(D, \tilde{\Sigma}\right)$ for every Σ and $\tilde{\Sigma}$ be such that $(\sigma_1^2, \ldots, \sigma_N^2)$ majorizes $(\tilde{\sigma}_1^2, \ldots, \tilde{\sigma}_N^2)$.*

Majorization is a partial order that makes precise the notion that the components of one vector are "more spread out" than are those of another vector with the same sum of components. Appendix A has a short introduction to this partial order. Thus Proposition 3.1

says that for fixed user power constraints, C_{opt} increases if the noise variance becomes "more colored" while keeping the total noise variance ($\mathrm{tr}\Sigma$) constant. On the other hand keeping the additive noise variances fixed, if the user power constraints are asymmetric keeping the total user power fixed it is intuitive that there is lesser flexibility in choosing μ. We make this precise below:

Proposition 3.2 *For fixed Σ, $C_{opt}(D,\Sigma)$ is concave in D and furthermore, for every $D \neq \tilde{D}$ such that $(p_1,\ldots,p_K)$ majorizes $(\tilde{p}_1,\ldots,\tilde{p}_K)$ we have $C_{opt}\left(\tilde{D},\Sigma\right) > C_{opt}(D,\Sigma)$.*

We conclude that $C_{opt}(\Sigma,D)$ is a concave (and Schur-concave) function in D and convex (and Schur-convex) in Σ (see Appendix A for the notation). Thus C_{opt} is a *saddle function* in D and Σ (in fact C_{opt} is also a "Schur-saddle function" in the sense of the results of Propositions 3.1 and 3.2). This saddle function is reminiscent of the famous Shannon saddle function property of mutual information:

$$I\left(\bar{X}_g; S\bar{X}_g + Z\right) \geq I\left(\bar{X}_g; S\bar{X}_g + W\right) \geq I\left(\bar{X}; S\bar{X} + W\right) .$$

where $\bar{X}$ is a K dimensional random vector with covariance matrix D and $\bar{X}_g$ is a K dimensional Gaussian random vector with the same covariance matrix D. Also, Z is a N dimensional noise vector with covariance matrix Σ and W is a N dimensional Gaussian noise vector with the same covariance matrix Σ. Our result says that C_{opt}, the maximum value of $I\left(\bar{X}; S\bar{X} + W\right)$ (maximum over $S \in \mathcal{S}$ and independent distributions on $\bar{X}$ subject to a variance constraint), is a saddle function in D and Σ. The formal proofs of Propositions 3.1 and 3.2 are in the next section.

4 Proofs of Main Result

4.1 Proof of Lemma 3.1

We begin with the following claim. For any $n \times n$ positive semidefinite matrices A and B with vector of eigenvalues $\left(\lambda_1^{(A)},\ldots,\lambda_n^{(A)}\right)$ and $\left(\lambda_1^{(B)},\ldots,\lambda_n^{(B)}\right)$ respectively,

$$\log\det(I + AB) \leq \max_{\pi\in S_n} \sum_{i=1}^{n} \log\left(1 + \lambda_{\pi(i)}^{(A)}\lambda_i^{(B)}\right) \tag{9}$$

where S_n is the permutation group of size n. This claim follows direction from Corollary 1 of [2]. Fix $S \in \mathcal{S}$. The set $\{QS : Q \in O(N)\}$ is also in $\mathcal{S}$ (here $O(N)$ is the standard notation

for the orthogonal group) and for every element QS in this set, the matrix $QSD(QS)^t$ has the same vector of eigenvalues denoted by $\left(\lambda_1^{(S)}, \ldots, \lambda_N^{(S)}\right)$. Furthermore, using (9) there exists a $Q \in O(N)$ such that

$$\log\det\left(I + \Sigma^{-1}QSD(QS)^t\right) = \sum_{i=1}^{N} \log\left(1 + \frac{\lambda_i^{(S)}}{\sigma_i^2}\right) .$$

Thus (3) can be rewritten as

$$C_{opt} = \max_{S \in \mathcal{S}} \frac{1}{2N} \sum_{i=1}^{N} \log\left(1 + \frac{\lambda_i^{(S)}}{\sigma_i^2}\right) . \tag{10}$$

In Section 3 of [7] we show that the set spanned by $\lambda^{(S)}$ as S varies over $\mathcal{S}$ is given by

$$\mathcal{L}_1' = \left\{(\lambda_1, \ldots, \lambda_N) \in \mathbb{R}_+^N : (\lambda_1, \ldots, \lambda_N, 0, \ldots, 0) \text{ majorizes } (p_1, \ldots, p_K)\right\} . \tag{11}$$

For every $\lambda \in \mathcal{L}_1'$, Section 4 of [7] outlines a procedure to construct $S \in \mathcal{S}$ (in N steps) such that SDS^t has vector of eigenvalues λ. We briefly recall below the key steps of this procedure which will be useful later in the paper as well. Fix $\lambda \in \mathcal{L}_1'$. Appealing to Lemma A.3, there exists a symmetric matrix H with eigenvalues $\lambda_1, \ldots, \lambda_N, 0, \ldots, 0$ and diagonal elements $p_1, \ldots, p_K$. Let $v_1, \ldots, v_N \in \mathbb{R}^K$ be the normalized eigenvectors of H corresponding to the eigenvalues $\lambda_1, \ldots, \lambda_N$. Let $V^t = [v_1 v_2 \ldots v_N]$. If we let Λ to be the diagonal matrix with entries $\lambda_1, \ldots, \lambda_N$, then $H = V^t \Lambda V$. Now define $S = \Lambda^{\frac{1}{2}} V D^{-\frac{1}{2}}$. Then, since the square of the l_2 norms of the columns of S are the diagonal elements of S^tS, we verify that $S^tS = D^{-\frac{1}{2}} H D^{-\frac{1}{2}}$ has unit diagonal entries concluding that $S \in \mathcal{S}$. Section 4 of [7] also outlines a simple recursive algorithm that constructs a symmetric matrix given its diagonal entries and eigenvalues. Writing $\mu_i = \lambda_i + \sigma_i^2$ we can rewrite (10) as

$$C_{opt} = \max_{\mu \in \mathcal{L}_2} \sum_{i=1}^{N} \log \frac{\mu_i}{\sigma_i^2} \tag{12}$$

where

$$\mathcal{L}_1 = \left\{(\mu_1, \ldots, \mu_N) : \begin{array}{l} \mu_i \geq \sigma_i^2, \ \forall i = 1 \ldots N \\ (\mu_1 - \sigma_1^2, \ldots, \mu_N - \sigma_N^2, 0, \ldots, 0) \text{ majorizes } (p_1, \ldots, p_K) \end{array}\right\} .$$

Define the set

$$\mathcal{L} = \mathcal{L}_1 \bigcap \left\{\mu : \mu_1 - \sigma_1^2 \geq \mu_2 - \sigma_2^2 \geq \cdots \geq \mu_N - \sigma_N^2 \geq 0\right\} .$$

and observe that this definition is identical to that in (7). Consider the following claim:

$$\forall \mu \in \mathcal{L}_1, \text{ there exists } \tilde{\mu} \in \mathcal{L} \text{ such that } \mu \text{ majorizes } \tilde{\mu} . \tag{13}$$

The map $C : (\mu_1, \ldots, \mu_N) \mapsto \sum_{i=1}^{N} \log \frac{\mu_i}{\sigma_i^2}$ is Schur-concave (see Definition A.3) in $\mathcal{L}_1$ and thus $C(\mu) \leq C(\tilde{\mu})$ for every $\tilde{\mu}$ majorized by μ. This combined with (13) completes the proof of the lemma. We now show (13). Suppose $\mu \in \mathcal{L}_1$ and $\mu \notin \mathcal{L}$. Then there exits at least one pair of indices (i, j) such that

$$i > j, \text{ and } \mu_i - \sigma_i^2 > \mu_j - \sigma_j^2 .$$

Define the vector $\tilde{\mu}$ that differs from μ only in the components indexed by i and j as

$$\tilde{\mu}_i = \mu_j - \sigma_j^2 + \sigma_i^2 \text{ and } \tilde{\mu}_j = \mu_i - \sigma_i^2 + \sigma_j^2.$$

It is seen that

$$\begin{aligned} \tilde{\mu} &\in \mathcal{L}_1 \text{ and that } \mu_i - \sigma_i^2 > \mu_j - \sigma_j^2 \text{ and} \\ \tilde{\mu}_i &= \alpha\mu_i + (1-\alpha)\mu_j,\ \tilde{\mu}_j = (1-\alpha)\mu_i + \alpha\mu_j \text{ where } \alpha = \frac{\sigma_i^2 - \sigma_j^2}{\mu_i - \mu_j} \in [0, 1) . \end{aligned}$$

Thus $\tilde{\mu}$ is majorized by μ (See Example A.2). By interchanging every pair (i, j) with this property we can sequentially construct $\tilde{\mu}$ that is majorized by μ and belongs to $\mathcal{L}$. This verifies the claim in (13). □

4.2 Proof of Theorem 3.1

We denote the optimization problem in (6) by $\mathcal{P} = (K, N, (p_1, \ldots, p_K), (\sigma_1^2, \ldots, \sigma_N^2))$ and the region over which the Schur-concave function

$$C : (\mu_1, \ldots, \mu_N) \mapsto \frac{1}{2N} \sum_{i=1}^{N} \log \frac{\mu_i}{\sigma_i^2}$$

is maximized by $\mathcal{L}(\mathcal{P})$. We begin with some preliminary observations about algorithm $\mathcal{A}$.

1. If $\frac{\sum_{i=1}^{K} p_i + \sum_{i=1}^{N} \sigma_i^2}{N} \geq \max\left\{\sigma_N^2, \frac{\sum_{i=1}^{l}(p_{[i]} + \sigma_i^2)}{l},\ l = 1 \ldots N-1\right\}$ then algorithm $\mathcal{A}$ output μ^* has all equal components. Hence we have that μ^* is majorized by μ for any $\mu \in \mathcal{L}$ (see Example A.1). This will complete the claim that μ^* is indeed the optimizing argument. We henceforth assume that this case does not occur.

2. The algorithm $\mathcal{A}$ updates the components of μ^* in non increasing order. In other words, the order of updates of algorithm $\mathcal{A}$ is $\mu^*_{[1]}, \ldots, \mu^*_{[N]}$.

3. For any $\mu \in \mathcal{L}$ we have $\mu^*_{[1]} \leq \mu_{[1]}$. This observation is trivial since η in Step 3 of $\mathcal{A}$ is always less than or equal to $\mu_{[1]}$ for every $\mu \in \mathcal{L}$.

4. If $\mu_j^* = \sigma_j^2$ for some j then it follows that $\mu_l^* = \sigma_l^2$ for every $j \leq l \leq N$.

Our proof that the output of algorithm $\mathcal{A}$ is optimal is by induction. First consider the case $N = 2$ and arbitrary $K \geq 2$. Since for every $\mu \in \mathcal{L}$ we have $\mu_{[1]}^* \leq \mu_{[1]}$ and $\mu_1^* + \mu_2^* = \mu_1 + \mu_2$, we conclude that μ^* is majorized by μ and thus $C(\mu^*) \geq C(\mu)$. This completes the proof. We now make the induction hypothesis that the output of $\mathcal{A}$ is optimal for all $N \leq n$ and all $K \geq N$. We show that the output of $\mathcal{A}$ is optimal for $N = n+1$ and any $K \geq n+1$. Let us denote this problem by $\mathcal{P} = (K, N, (p_1, \ldots, p_K), (\sigma_1^2, \ldots, \sigma_N^2))$. Suppose $\mu \in \mathcal{L}$ is the optimal argument to the optimization problem in (6) and the output μ^* of $\mathcal{A}$ is such that $\left(\mu_{[1]}^*, \ldots, \mu_{[N]}^*\right) \neq \left(\mu_{[1]}, \ldots, \mu_{[N]}\right)$. We now proceed to get a contradiction to the hypothesis that μ is the optimal solution to (6).

1. Suppose $\mu_{[1]} = \mu_{[1]}^*$.

   (a) If $\mu_{[1]}^* = \sigma_N^2 (= \mu_N^*)$, then $\left(\mu_{[2]}^*, \ldots, \mu_{[N]}^*\right)$ is the output of $\mathcal{A}$ to the reduced problem $\mathcal{P}' = \left(K, N-1, (p_1, \ldots, p_K), \left(\sigma_1^2, \ldots, \sigma_{N-1}^2\right)\right)$. By hypothesis $\mu_{[1]} = \mu_{[1]}^* = \sigma_N^2$ and thus $(\mu_1, \ldots, \mu_{N-1}) \in \mathcal{L}\left(\mathcal{P}'\right)$. By the induction hypothesis $C\left(\mu_{[2]}^*, \ldots, \mu_{[N]}^*\right) \geq C(\mu_1, \ldots, \mu_{N-1})$. Since $\mu_{[j]} \neq \mu_{[j]}^*$ for some $j \in \{2, \ldots, N\}$ we have by the strict Schur-concavity of C that $C(\mu^*) > C(\mu)$. Thus we arrive at a contradiction to the hypothesis that μ is the optimal argument in (6) completing the proof.

   (b) If $\mu = \frac{\sum_{i=1}^{l}\left(p_{[i]} + \sigma_i^2\right)}{l}$ for some $l \in \{1, \ldots, N-1\}$ then from $\mathcal{A}$ we have $\mu^*[1] = \mu_1^* = \cdots = \mu_l^*$. Using the fact that $\mu \in \mathcal{L}$ we arrive at $\mu_{[1]}^* = \mu_{[1]} = \mu_1 = \cdots = \mu_l$. Thus $\left(\mu_{l+1}^*, \ldots, \mu_N^*\right)$ and $(\mu_{l+1}, \ldots, \mu_N)$ belong to $\mathcal{L}\left(\mathcal{P}'\right)$ where $\mathcal{P}' = \left(K-l, N-l, (p_{l+1}, \ldots, p_K), \left(\sigma_{l+1}^2, \ldots, \sigma_N^2\right)\right)$. By the induction hypothesis $\left(\mu_{l+1}^*, \ldots, \mu_N^*\right)$ is the optimal argument of C in $\mathcal{L}\left(\mathcal{P}'\right)$ and hence $C\left(\mu_{l+1}^*, \ldots, \mu_N^*\right) > C(\mu_{l+1}, \ldots \mu_N)$ contradicts the hypothesis that μ is the optimal argument of C in $\mathcal{L}(\mathcal{P})$, completing the proof.

   (c) As observed earlier, we do not need to consider the case when $\mu_{[1]}^* = \frac{\sum_{i=1}^{K} p_i + \sum_{i=1}^{N} \sigma_i^2}{N}$ since in this case μ^* is the optimal argument.

2. Henceforth we take $\mu_{[1]} > \mu_{[1]}^*$.

   (a) Let $\mu_{[1]}^* = \sigma_N^2 = \mu_N^*$. Let $1 \leq j \leq N$ be the largest index such that $\mu_{[1]} = \mu_j$.

      i. Suppose $j = N$. Since $N\mu_{[1]} > \sum_{i=1}^{N} p_i + \sum_{i=1}^{N} \sigma_i^2$ there is some $1 \leq l < N$ such that $\mu_l < \mu_N = \mu_{[1]} = \sigma_N^2$. Thus we can define a vector $\tilde{\mu}$ differing from

μ only in components indexed by N and l as follows:

$$\tilde{\mu}_N = \mu_N - \epsilon,\ \tilde{\mu}_l = \mu_l + \epsilon, \quad \text{where } \epsilon = \min\left\{\frac{\mu_N - \sigma_N^2}{2}, \frac{\mu_N - \mu_j}{2}\right\} > 0\,.$$

It is clear that $\tilde{\mu} \in \mathcal{L}(\mathcal{P})$ and that $\tilde{\mu}$ is majorized by μ. Thus $C(\tilde{\mu}) < C(\tilde{\mu})$ and we arrive at a contradiction to our hypothesis that μ was optimal on $\mathcal{L}(\mathcal{P})$.

ii. Suppose $j \neq N$. Suppose further that $\mu_1 = \mu_2 = \ldots = \mu_j$. By definition we have $\mu_{j+1} < \mu_j$ and by hypothesis that $j\mu_j > j\sigma_N^2 > \sum_{i=1}^{j}\left(p_{[i]} + \sigma_i^2\right)$. Hence we can define a vector $\tilde{\mu}$ differing from μ only in components j and $j+1$ by

$$\tilde{\mu}_j = \mu_j - \epsilon,\ \tilde{\mu}_{j+1} = \mu_{j+1} + \epsilon, \quad \text{where } \epsilon = \min\left\{\frac{\mu_j - \mu_{j+1}}{2}, \frac{\mu_j - \sigma_j^2}{2}, \frac{\mu_j - \sum_{i=1}^{j}\left(p_{[i]} + \sigma_i^2\right)}{2}\right\}\,.$$

It is clear that $\tilde{\mu} \in \mathcal{L}(\mathcal{P})$ and that $\tilde{\mu}$ is majorized by μ and we arrive at a contradiction as before. Now suppose that all of $\mu_1, \ldots \mu_j$ are not equal. Then there must exist some $1 \leq l < j$ such that $\mu_l < \mu_j$. Then we can define a vector $\tilde{\mu}$ differing from μ only in the components j and l by

$$\tilde{\mu}_j = \mu_j - \epsilon,\ \tilde{\mu}_l = \mu_l + \epsilon, \quad \text{where } \epsilon = \min\left\{\frac{\mu_j - \mu_l}{2}, \frac{\mu_j - \sigma_j^2}{2}\right\}\,.$$

As before $\tilde{\mu} \in \mathcal{L}(\mathcal{P})$ and $\tilde{\mu}$ is majorized by μ arriving at the contradiction.

(b) Let $\mu_{[1]}^* = \frac{\sum_{i=1}^{l}\left(p_{[i]} + \sigma_i^2\right)}{l}$ for some $1 \leq l < N$. Then $\mu_{[1]}^* \geq \sigma_N^2$. Let $1 \leq j \leq N$ be the largest index such that $\mu_{[1]} = \mu_j$.

i. Let $j = N$. By definition and hypothesis, $\mu_N = \mu_{[1]} > \mu_{[2]}^* \geq \sigma_N^2$. Furthermore, $N\mu_{[1]} = N\mu_N > \sum_{i=1}^{K} p_i + \sum_{i=1}^{N} \sigma_i^2$. Hence there exists $1 \leq l < N$ such that $\mu_l < \mu_N$. We can then define a vector $\tilde{\mu}$ differing from μ only in the components indexed by N and l by

$$\tilde{\mu}_N = \mu_N - \epsilon,\ \tilde{\mu}_l = \mu_l + \epsilon, \quad \text{where } \epsilon = \min\left\{\frac{\mu_N - \mu_l}{2}, \frac{\mu_N - \sigma_N^2}{2}\right\}$$

and the contradiction follows as before.

ii. Let $j \neq N$. Suppose $\mu_1 = \mu_2 = \cdots = \mu_j$. Then we have $j\mu_j = j\mu_{[1]} > j\mu_{[1]}^* \sum_{i=1}^{j}\left(p_{[i]} + \sigma_i^2\right)$. Furthermore $\mu_j > \mu_{[1]}^* \geq \sigma_j^2$ and $\mu_{j+1} < \mu_j$. Hence we can define a vector $\tilde{\mu}$ differing from μ only in the components indexed by j and $j+1$ by

$$\tilde{\mu}_j = \mu_j - \epsilon,\ \tilde{\mu}_{j+1} = \mu_{j+1} + \epsilon, \quad \text{where } \epsilon = \min\left\{\frac{\mu_j - \mu_{j+1}}{2}, \frac{\mu_j - \sigma_j^2}{2}, \frac{\mu_j - \sum_{i=1}^{j}\left(p_{[i]} + \sigma_i^2\right)}{2}\right\}$$

arriving at the contradiction as before.

This exhausts all the cases and completes the proof of Theorem 3.1. □

4.3 Proof of Propositions 3.1 and 3.2

Fix Σ and consider $\tilde{D}$ and D such that $(\tilde{p}_1, \ldots, \tilde{p}_K)$ majorizes $(p_1, \ldots, p_K)$. In the notation of the proof of Lemma 3.1 using the transitivity of the partial order of majorization we have

$$\mathcal{L}_1'\left(K, N, \tilde{D}\right) \subseteq \mathcal{L}_1'\left(K, N, D\right) .$$

Thus $\mathcal{L}\left(K, N, \tilde{D}, \Sigma\right) \subseteq \mathcal{L}\left(K, N, D, \Sigma\right)$ and C_{opt} is Schur-concave in D for fixed Σ. To see concavity, fix D_1 and D_2. From (12) and (13) we can write for $j = 1, 2$

$$C_{opt}\left(D_j, \Sigma\right) = \frac{1}{2N} \max_{\lambda \in \mathcal{L}'(K,N,D_j)} \sum_{i=1}^{N} \log\left(1 + \frac{\lambda_i}{\sigma_i^2}\right) \tag{14}$$

where

$$\mathcal{L}'\left(K, N, D\right) \stackrel{\text{def}}{=} \mathcal{L}_1'\left(K, N, D\right) \bigcap \left\{\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N \geq 0\right\} .$$

Observe now that if $\lambda^{(j)} \in \mathcal{L}'\left(K, N, D_j\right)$ for $j = 1, 2$, then for every $\alpha \in (0, 1)$

$$\alpha\lambda^{(1)} + (1 - \alpha)\,\lambda^{(2)} \in \mathcal{L}'\left(K, N, \alpha D_1 + (1 - \alpha)\, D_2\right) . \tag{15}$$

Using the concavity of the logarithm, we have for every $\alpha \in (0, 1)$, from (14),

$$\begin{aligned}
\alpha C_{opt}\left(D_1, \Sigma\right) + (1 - \alpha)\, C_{opt}\left(D_2, \Sigma\right) &\leq \max_{\left\{\lambda^{(j)} \in \mathcal{L}'(K,N,D_j),\, j=1,2\right\}} \frac{1}{2N} \sum_{i=1}^{N} \log\left(1 + \frac{\alpha\lambda_i^{(1)} + (1 - \alpha)\,\lambda_i^{(2)}}{\sigma_i^2}\right) \\
&\leq \max_{\left\{\lambda \in \mathcal{L}'(K,N,\alpha D_1 + (1-\alpha) D_2)\right\}} \sum_{i=1}^{N} \log\left(1 + \frac{\lambda_i}{\sigma_i^2}\right) \\
&= C_{opt}\left(\alpha D_1 + (1 - \alpha)\, D_2, \Sigma\right)
\end{aligned}$$

where we used (15) in the second step. This shows Proposition 3.2. Now fix D, Σ and $\tilde{\Sigma}$. Here Σ and $\tilde{\Sigma}$ are such that the vector of their eigenvalues (arranged in nondecreasing order)

$$\mathrm{eig}\left(\Sigma\right) \stackrel{\text{def}}{=} \left(\sigma_1^2, \ldots, \sigma_N^2\right) \text{ majorizes } \mathrm{eig}\left(\tilde{\Sigma}\right) \stackrel{\text{def}}{=} \left(\tilde{\sigma}_1^2, \ldots, \tilde{\sigma}_N^2\right) .$$

We will show that

$$C_{opt}\left(D, \Sigma\right) = C_{opt}\left(D, \mathrm{eig}\left(\Sigma\right)\right) \geq C_{opt}\left(D, \mathrm{eig}\left(\tilde{\Sigma}\right)\right) = C_{opt}\left(D, \tilde{\Sigma}\right) .$$

Appealing to Lemma A.2, it suffices to show that for every T-transform T

$$C_{opt}\left(D, \mathrm{eig}\left(\Sigma\right)\right) \geq C_{opt}\left(D, T\left(\mathrm{eig}\left(\Sigma\right)\right)\right) . \tag{16}$$

Without loss of generality, $T(\mathrm{eig}(\Sigma))$ can be taken to be different from $\mathrm{eig}(\Sigma)$ only in the entries indexed by i and j for some $1 \leq i < j \leq N$. We can then write for some $\alpha \in (0,1)$

$$T(\mathrm{eig}(\Sigma))_i = \tilde{\sigma}_i^2 \stackrel{\mathrm{def}}{=} \alpha\sigma_i^2 + (1-\alpha)\sigma_j^2, \text{ and } T(\mathrm{eig}(\Sigma))_j = \tilde{\sigma}_j^2 \stackrel{\mathrm{def}}{=} \alpha\sigma_j^2 + (1-\alpha)\sigma_i^2 .$$

Now, from (14),

$$\begin{aligned} C_{opt}(D, \mathrm{eig}(\Sigma)) &= \frac{1}{2N} \max_{\lambda \in \mathcal{L}'(K,N,D)} \left\{ \log\left(1+\frac{\lambda_i}{\sigma_i^2}\right) + \log\left(1+\frac{\lambda_j}{\sigma_j^2}\right) + \sum_{l \neq i,j} \log\left(1+\frac{\lambda_l}{\sigma_l^2}\right) \right\} \\ &= C_{opt}(D, T(\mathrm{eig}(\Sigma))) + \frac{1}{2N} \max_{\lambda \in \mathcal{L}'(K,N,D)} \left\{ \log\left(1+\frac{\lambda_i}{\sigma_i^2}\right) + \log\left(1+\frac{\lambda_j}{\sigma_j^2}\right) - \right. \\ &\qquad \left. \log\left(1+\frac{\lambda_i}{\tilde{\sigma}_i^2}\right) - \log\left(1+\frac{\lambda_j}{\tilde{\sigma}_j^2}\right) \right\} . \end{aligned} \tag{17}$$

Now, on $\mathcal{L}'(K,N,D)$ we have $\lambda_i \geq \lambda_j$ and recalling that $\sigma_i^2 \leq \sigma_j^2$ we arrive at

$$\log\left(1+\frac{\lambda_i}{\sigma_i^2}\right) + \log\left(1+\frac{\lambda_j}{\sigma_j^2}\right) \geq \log\left(1+\frac{\lambda_i}{\sigma_j^2}\right) + \log\left(1+\frac{\lambda_j}{\sigma_i^2}\right) . \tag{18}$$

Using the convexity of the map $x \mapsto \log\left(1+\frac{a}{x}\right),;\ a > 0$, and applying (18) in (17) we have shown that $C_{opt}(D, \mathrm{eig}(\Sigma)) \geq C_{opt}(D, T(\mathrm{eig}(\Sigma)))$. To see the convexity of C_{opt} in the covariance matrix Σ, fix any two noise covariance matrices Σ and $\tilde{\Sigma}$. From (14) we have, using the convexity of the map $x \mapsto \log\left(1+\frac{a}{x}\right);\ a > 0$,

$$\begin{aligned} \alpha C_{opt}(D,\Sigma) + (1-\alpha) C_{opt}\left(D, \tilde{\Sigma}\right) &\geq \frac{1}{2N} \max_{\lambda \in \mathcal{L}'(K,N,D)} \sum_{i=1}^{N} \log\left(1+\frac{\lambda_i}{\alpha\sigma_i^2 + (1-\alpha)\tilde{\sigma}_i^2}\right) \\ &= C_{opt}\left(D, \alpha\mathrm{eig}(\Sigma) + (1-\alpha)\mathrm{eig}\left(\tilde{\Sigma}\right)\right) \\ &\geq C_{opt}\left(D, \mathrm{eig}\left(\alpha\Sigma + (1-\alpha)\tilde{\Sigma}\right)\right) \\ &= C_{opt}\left(D, \alpha\Sigma + (1-\alpha)\tilde{\Sigma}\right) \end{aligned}$$

where we used Lemma A.1 in conjunction with the earlier proof of the Schur-convexity of C_{opt} for fixed D (expressed in (16)) in arriving at the last but one step. This shows the convexity of C_{opt} in Σ and completes the proof of Proposition 3.1. □

5 Discussion

The general scheme to construct the optimal signal directions is contained in our proofs of the main results: Lemma 3.1 and Theorem 3.1. We summarize this construction below.

Denote the eigenvector of Σ corresponding to the eigenvalue σ_j^2 by q_j for $1 \leq j \leq N$ and write $Q = [q_1, \ldots, q_N]$. We first use algorithm $\mathcal{A}$ to generate μ^* and construct the vector $\lambda^* = (\mu_1^* + \sigma_1^2, \ldots, \mu_N^* + \sigma_N^2)$. We then use the recursive algorithm in Section 4 of [7] to construct $S \in \mathcal{S}$ such that the matrix SDS^t is the diagonal matrix diag $\{\lambda_1^*, \ldots, \lambda_N^*\}$. Then the optimal signal directions S^* are given by QS. However, the structure of $\mathcal{A}$ yields more insight into the nature of the optimal signal directions S^* and this allows the following more succinct characterization and construction of the optimal signal directions.

1. We begin with the first iteration of $\mathcal{A}$. If Step 3(a) is reached, then we set $\lambda_N^* = 0$. This means that the optimal signal directions are orthogonal to q_N. Thus this allows us to recursively reduce the problem to one with only $N-1$ degrees of freedom.

2. If Step 3(b) is reached, then we set $\lambda_j^* = \frac{\mathrm{trD}+\mathrm{tr}\Sigma}{N} - \sigma_j^2$, $1 \leq j \leq N$. We use the recursive algorithm of Section 4, [7] to construct $S \in \mathcal{S}$ such that $SDS^t = \mathrm{diag}\,\{\lambda_1^*, \ldots, \lambda_N^*\}$. Then the optimal signal directions are $S^* = QS$. This step terminates the algorithm and completes the construction.

3. If Step 3(c) is reached for $1 \leq l < N$, then we have $\lambda_{[j]}^* = \lambda_j^* = \frac{\sum_{i=1}^{l}(p_{[i]}+\sigma_i^2)}{l} - \sigma_j^2$, $1 \leq j \leq l$. For expository ease, we assume that the users are ordered according to their power constraints, i.e. $p_1 \geq p_2 \geq \cdots \geq p_K$. By hypothesis that Step 3(c) is reached for l, and by construction, $(\lambda_1^*, \ldots, \lambda_l^*)$ majorizes the vector $(p_1, \ldots, p_l)$. Thus we can use the procedure in Section 4, [7] (summarized in the proof of Lemma 3.1) to construct a $l \times l$ matrix S_l such that $S_l \mathrm{diag}\,\{p_1, \ldots, p_l\}\, S_l^t = \mathrm{diag}\,\{\lambda_1^*, \ldots \lambda_l^*\}$. We construct the optimal signal directions for the first l users (these have the largest power constraints) as $[s_1^*, \ldots, s_l^*] = [q_1, \ldots, q_l]\, S_l$. The following is a key observation: recall that the output μ^* of $\mathcal{A}$ is in $\mathcal{L}$ and hence $(\lambda_1^*, \ldots, \lambda_N^*, 0, \ldots, 0)$ must majorize $(p_1, \ldots, p_K)$. By construction of $\lambda_{[i]}^* = \lambda_i^*$, $1 \leq i \leq l$ above, we must have

$$\left(\lambda_{l+1}^*, \ldots, \lambda_N^*, 0 \ldots, 0\right) \text{ majorizes } (p_{l+1}, \ldots, p_K)\ . \tag{19}$$

Recalling the construction of $S^* \in \mathcal{S}$ from λ^* from the proof of Lemma 3.1, we see from (19) that we can construct the $N-l \times K-l$ matrix $S_{\bar{l}}$ such that $S_{\bar{l}} \mathrm{diag}\,\{p_{l+1}, \ldots, p_K\}\, S_{\bar{l}}^t = \mathrm{diag}\left\{\lambda_{l+1}^*, \ldots, \lambda_N^*\right\}$. We then let the optimal signal directions for the remaining $K-l$ users to be $\left[s_{l+1}^*, \ldots, s_K^*\right] = [q_{l+1}, \ldots, q_N]\, S_{\bar{l}}$. We emphasize the point that each of the optimal signal directions of the first l users are orthogonal to each one of the optimal signal directions of the remaining $K-l$ users. Furthermore, the first l user signal directions span the l dimensional subspace span $\{q_1, \ldots, q_l\}$ while the signal directions of the remaining $K-l$ users span the orthogonal complement of this subspace. Thus if Step 3(c) is reached in the first iteration of $\mathcal{A}$, this observation allows us to identify the user signal directions for the first l users and recursively reduce the problem to one of l fewer users and l fewer degrees of freedom.

4. In Section 3.2 and in the proof of Theorem 3.1 we have identified certain key properties of $\mathcal{A}$. We summarize below the physical insights gained from these observations.

   (a) Consider the first iteration of $\mathcal{A}$. Step 3(a) is reached if the largest noise variance σ_N^2 is "much larger" than the other noise variances (in a sense made precise by $\mathcal{A}$). Our optimal signal directions are chosen in this case to be orthogonal to q_N and thus they avoid "directions" of high noise variance. We emphasize that this step is never reached if all the noise variances are equal.

   (b) Suppose Step 3(c) is reached for some $1 \leq l < N$. This means that the average of the largest l user power constraints is "much larger" than other averages of user powers in a sense that depends on the noise variances as well (made precise by $\mathcal{A}$) and the optimal signal directions are assigned to these l users so that they span a subspace (of dimension l) given by span $\{q_1, \ldots, q_l\}$. Thus these signal directions lie in the subspace with least noise and furthermore the other user signal directions are orthogonal to this subspace. We emphasize that this step is never reached if all the user powers are equal.

In Section 3.2 we observed that if $D = pI$ then the algorithm $\mathcal{A}$ output is identical to the waterfilling policy in (5). Thus $D = pI$ is a sufficient condition for the sum capacity C_{opt} to be equal to the upper bound C_p. The following proposition identifies the *precise* condition when this equality holds.

Proposition 5.1 $C_{opt}(D, \Sigma) = C_p(\mathrm{tr}D, \Sigma)$ *if and only if*

$$\max\left\{\frac{\sum_{i=1}^{K} p_i + \sum_{j=1}^{N} \sigma_j^2}{N}, \sigma_N^2\right\} \geq \max_{l=1\ldots N-1}\left\{\frac{1}{l}\sum_{i=1}^{l}\left(p_{[i]} + \sigma_i^2\right)\right\} . \tag{20}$$

Proof Our first claim is that a necessary and sufficient for equality of C_{opt} and C_p is that Step 3(c) of algorithm $\mathcal{A}$ is never reached. This follows the fact that when Step 3(c) is never reached, $\mathcal{A}$ simply reduces to the waterfilling allocation of (5) in Section 2.3 (this reduction of $\mathcal{A}$ was shown explicitly for the case $D = pI$ in Section 3.2). Necessity of (20) for Step 3(c) to be not reached in the first iteration of $\mathcal{A}$ is obvious from inspection. We complete the proof by showing the sufficiency of (20) for Step 3(c) to be never reached. In the first iteration of $\mathcal{A}$, (20) ensures that Step 3(c) is not reached. Furthermore, observe that if Step 3(b) is reached, then the algorithm terminates (this statement is true on any iteration of the algorithm). Now suppose that for some $1 \leq k < N$ iterations, $\mathcal{A}$ has not terminated and Step 3(c) has never been reached. This means that for each of the first k iterations, Step 3(a) is reached. It follows that $\mu_{[i+1]}^* = \mu_{N-i} = \sigma_{N-i}^2$, $i = 0, \ldots k-1$. We

now show that in iteration $k+1$, Step 3(c) is not reached. In iteration $k+1$, Step 3(c) is not reached if

$$\max_{\{1\le l\le N-k\}} \left\{ \frac{1}{l} \sum_{j=1}^{l} \left(p_{[j]} + \sigma_j^2 \right) \right\} \le \max \left\{ \frac{\sum_{i=1}^{K} p_i + \sum_{j=1}^{N-k} \sigma_j^2}{N-k}, \sigma_{N-k}^2 \right\} . \tag{21}$$

By the induction hypothesis for iteration k that $\mathcal{A}$ did not terminate, we have $\sum_{i=1}^{K} p_i + \sum_{j=1}^{N-k+1} \sigma_j^2 \le (N-k+1)\,\sigma_{N-k+1}^2$. Thus

$$\begin{aligned} \frac{\sum_{i=1}^{K} p_i + \sum_{j=1}^{N-k} \sigma_j^2}{N-k} &\le \frac{(N-k+1)\,\sigma_{N-k+1}^2}{N-k} - \frac{\sigma_{N-k+1}^2}{N-k}, \\ &= \sigma_{N-k+1}^2 \le \sigma_N^2 . \end{aligned}$$

Using this inequality and (20), we have shown (21) completing the proof. □

A Majorization

Majorization makes precise the vague notion that the components of a vector x are "less spread out" or "more nearly equal" than are the components of a vector y by the statement x is majorized by y.

Definition A.1 *For $x, y \in \mathbb{R}^n$, say that x is majorized by y (or y majorizes x) if*

$$\begin{aligned} &\textstyle\sum_{i=1}^{k} x_{[i]} \le \sum_{i=1}^{k} y_{[i]}, \quad k = 1 \ldots n-1 \\ &\textstyle\sum_{i=1}^{n} x_{[i]} = \sum_{i=1}^{n} y_{[i]} \end{aligned}$$

A comprehensive reference on majorization and its applications is [3]. A simple (trivial, but important) example of majorization between two vectors is the following:

Example A.1 *For every $a \in \mathbb{R}^n$ such that $\sum_{i=1}^{n} a_i = 1$,*

$$(a_1, \ldots, a_n) \quad \textit{majorizes} \quad \left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n} \right)$$

Another important example that we will use often is the following.

Example A.2 *Fix $a \in \mathbb{R}^n$ not of the form when all its components are equal. Let i and j be a pair of indices such that $a_i > a_j$. Define $\tilde{a} \in \mathbb{R}^n$ that differs from a in only the components indexed by i and j by*

$$\tilde{a}_i = a_i - \epsilon,\ \tilde{a}_j = a_j + \epsilon,\ \textit{for some } \epsilon \in \left[0, \frac{a_i - a_j}{2}\right) .$$

Then $\tilde{a}$ is majorized by a.

Let A and B be two symmetric matrices of dimension $n \times n$. Let λ^A and λ^B denote the vectors of eigenvalues of A and B respectively. The following result (Theorem 9.G.1 in [3]) shows that the eigenvalues of $A + B$ (the components of the vector λ^{A+B}) are less spread out than the sum of the order statistics of the eigenvalues of A and B:

Lemma A.1 *For any two symmetric matrices A and B,*

$$\left(\lambda_1^{A+B}, \ldots, \lambda_n^{A+B}\right)^t \textit{ is majorized by } \left(\lambda_{[1]}^A + \lambda_{[2]}^B, \ldots, \lambda_{[n]}^A + \lambda_{[n]}^B\right)^t$$

A permutation matrix Q of dimension $n \times n$ is a matrix with each entry equal to either 0 or 1 such that each row and column has exactly one entry equal to 1. A *T-transform* is a doubly stochastic matrix of the form

$$T = \alpha I + (1 - \alpha) Q$$

for some $\alpha \in [0, 1]$ and some permutation matrix Q with $n - 2$ diagonal entries equal to 1. To see the operation of a T-transform, let $y = (y_1, \ldots, y_n) \in \mathbb{R}^n$. Let $Q_{kl} = Q_{lk} = 1$ for some indices $k < l$. Then $Qy = (y_1, \ldots, y_{k-1}, y_l, y_{k+1}, \ldots, y_{l-1}, y_k, y_{l+1}, \ldots, y_K)$ and hence

$$Ty = (y_1, \ldots, y_{k-1}, \alpha y_k + (1 - \alpha) y_l, y_{k+1}, \ldots, y_{l-1}, \alpha y_l + (1 - \alpha) y_k, y_{l+1}, \ldots, y_n)$$

The following is a fundamental result from the theory of majorization (Lemma 2.B.1 in [3]).

Lemma A.2 *If x is majorized by y then there exists a sequence of T-transforms $T_1, \ldots, T_n$ such that $x = T_l \cdots T_2 T_1 y$ and $l < K$.*

It is well known that the sum of diagonal elements of a matrix is equal to the sum of its eigenvalues. When the matrix is symmetric the *precise* relationship between the diagonal elements and the eigenvalues is that of majorization:

Lemma A.3 (Theorem 9.B.1 and 9.B.2 in [3]) *Let H be a symmetric matrix with diagonal elements $h_1, \ldots, h_n$ and eigenvalues $\lambda_1, \ldots, \lambda_n$ we have*

$$(\lambda_1, \ldots, \lambda_n) \quad \textit{majorizes} \quad (h_1, \ldots, h_n)$$

That $h = (h_1, \ldots, h_n)$ and $\lambda = (\lambda_1, \ldots, \lambda_n)$ cannot be compared by an ordering stronger than majorization is the consequence of the following converse: If $h_1 \geq \cdots \geq h_n$ and $\lambda_1 \geq \cdots \lambda_n$ are $2n$ numbers such that λ majorizes h, then there exists a real symmetric matrix H with diagonal elements $h_1, \ldots, h_n$ and eigenvalues $\lambda_1, \ldots, \lambda_n$.

We will also need the following definition:

Definition A.2 *A real valued function $\phi : \mathbb{R}^n \to \mathbb{R}$ is said to be Schur-concave if for all $x, y \in \mathcal{R}^n$ such that y majorizes x we have $\phi(x) \geq \phi(y)$. Say that ϕ is strictly Schur-concave if y majorizes x and $y \neq x$ implies that $\phi(x) > \phi(y)$.*

An important class of Schur-concave functions is the following (Theorem 3.C.1 in [3]):

Example A.3 *If $g : \mathbb{R} \to \mathbb{R}$ is concave then the symmetric concave function $\phi(x) = \sum_{i=1}^{n} g(x_i)$ is Schur-concave.*

A function g is Schur-convex if $-g$ is Schur-concave.

References

[1] T. M. Cover, and J. A. Thomas, *Elements of Information Theory*, New York: Wiley, 1991.

[2] S. W. Drury, "A bound for the determinant of certain Hadamard products and for the determinant of the sum of two normal matrices", *Linear Algebra and Applications*, 199:329-338, 1994.

[3] A. W. Marshall, and I. Olkin, *Inequalities: Theory of Majorization and its applications*, Academic Press, 1979.

[4] M. Rupf and J. L. Massey, "Optimum sequence multisets for Synchronous code-division multiple-access channels", *IEEE Transactions on Information Theory*, Vol 40(4), July 1994, pp. 1261-1266.

[5] B. Suard, G. Xu, H. Liu and T. Kailath, "Uplink channel capacity of space-division-multiple-access schemes", *IEEE Transactions on Information Theory*, Vol 44(4), July 1998, pp 1468-1476.

[6] S. Verdú, "Capacity region of Gaussian CDMA channels: the symbol-synchronous case", in *Proc. 24th Allerton Conf.*, Oct. 1986, pp. 1025-1034

[7] P. Viswanath and V Anantharam, "Optimal sequences and sum capacity of synchronous CDMA systems", *IEEE Transactions on Information Theory*, vol. 45(6), Sept. 1999, pp. 1984-1991.