Multiuser Transmit Beamforming for Maximum Sum Capacity in Tactical Wireless Multicast Networks

Final Technical Report

by

Prof. Nikos Sidiropoulos

August, 2006

United States Army

EUROPEAN RESEARCH OFFICE OF THE U.S. ARMY

London, England

CONTRACT NUMBER N62558-03-C-0012

Telecommunication Systems Institute (TSI)

Technical University of Crete, Greece

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	August 2, 2006	Final Technical Report, 7/24/2003-7/23/2006

4. TITLE AND SUBTITLE
Multiuser Transmit Beamforming for Maximum Sum Capacity in Tactical Wireless Multicast Networks: Final Technical Report

5. FUNDING NUMBERS
Contract #: N62558-03-C-0012

6. AUTHOR(S)
Prof. Nikos Sidiropoulos, PI

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Telecommunication Systems Institute
Technical University of Crete
Kounoupidiana, Chania-Crete 73100
GREECE

8. PERFORMING ORGANIZATION REPORT NUMBER
12

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army European Research Office
Dr. James Harvey, Director
USARDSG-UK,
Edison House, 223 Old Marylebone Rd.,
London, NW1 5TH, U.K.

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT ***(Maximum 200 Words)*** Tactical wireless networks often comprise clusters of nodes, which are fed information from a head node. Transmit antenna arrays mounted on the head node (e.g., unmanned aerial vehicle) offer an attractive means of boosting capacity and assuring quality of service through transmit beamforming. The central goal of our research was to investigate efficient multiuser transmit beamforming strategies, and develop high-throughput low-complexity algorithms that will meet the needs of future tactical wireless networks. Sum capacity, quality of service, and fair service objectives were considered, under unicast and multicast scenarios. A key innovation of our work is the concept of physical layer multicasting, which affords significant capacity gains. A number of effective and efficient algorithms were developed, drawing upon and contributing to semidefinite relaxation (SDR) tools. Closely-related added-value topics of our research program included i) computationally efficient quasi-optimal multiple input multiple output detection (using lattice search, data association, and SDR tools); ii) accurate and scalable node localization from pairwise distance estimates; and iii) tracking of time-varying carrier signals (using and developing associated particle filtering tools). Our work on these topics has been reported in seven (IEEE, SIAM) journal papers and seven IEEE conference papers. Variants of some of our published algorithms are currently considered for adoption by industry.

14. SUBJECT TERMS Transmit beamforming, minimization of radiation power, quality of service, max-min fair, sum capacity, broadcasting, multicasting, convex optimization, semidefinite programming, NP-hard problems, semidefinite relaxation, lattice search, node localization, multidimensional scaling, tracking, intercept, particle filters.

15. NUMBER OF PAGES: 171

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT

CONTENTS

I. ABSTRACT

Tactical wireless networks often comprise clusters of nodes, which are fed information from a head node. Transmit antenna arrays mounted on the head node (e.g., unmanned aerial vehicle) offer an attractive means of boosting capacity and assuring quality of service through transmit beamforming. The central goal of our research was to investigate efficient multiuser transmit beamforming strategies, and develop high-throughput low-complexity algorithms that will meet the needs of future tactical wireless networks. Sum capacity, quality of service, and fair service objectives were considered, under unicast and multicast scenarios. A key innovation of our work is the concept of physical layer multicasting, which affords significant capacity gains. A number of effective and efficient algorithms were developed, drawing upon and contributing to semidefinite relaxation (SDR) tools. Closely-related added-value topics of our research program included i) computationally efficient quasi-optimal multiple input multiple output detection (using lattice search, data association, and SDR tools); ii) accurate and scalable node localization from pairwise distance estimates; and iii) tracking of time-varying carrier signals (using and developing associated particle filtering tools). Our work on these topics has been reported in seven (IEEE, SIAM) journal papers and seven IEEE conference papers. Variants of some of our published algorithms are currently considered for adoption by industry.

Keywords: Transmit beamforming, minimization of radiation power, quality of service, max-min fair, sum capacity, broadcasting, multicasting, convex optimization, semidefinite programming, NP-hard problems, semidefinite relaxation, lattice search, integer least squares, node localization, multidimensional scaling, tracking, intercept, particle filtering

II. Motivation and Problem Statement

Tactical wireless networks must seamlessly support diverse services, including command and control, "bulk" information dissemination (e.g., terrain maps), and large-scale surveillance and sensing (e.g., radar, alien signal interception, biochemical sensor networks). These come with equally diverse service needs: guaranteed quality of service for command and control, very high transmission rates for bulk information dissemination, reliable detection under stringent energy constraints for sensor networks. While truly seamless unified solutions are still way down the road, there is a number of enabling communication technologies and concepts that have emerged at the center stage of network science, particularly for tactical networks. These include

- The deployment of transmit antenna arrays, for assuring quality of service and/or higher data rates through spatial multiplexing;
- Wireless multicasting, as a means of improving spectral utilization and assuring quick and efficient delivery of mission-critical information;
- Effective strategies for vector decoding, as a means of improving spectral efficiency and robustness to jamming;
- Node localization, for sensing, routing, fading channel estimation, and situational awareness; and
- Carrier sensing and tracking, for signal intelligence, dynamic spectrum monitoring and access.

Our work in this project addresses many important aspects of the aforementioned enabling concepts and technologies. While many of our contributions specifically target applications in tactical networks, some have a clear dual use, e.g., in 802.16e fixed wireless systems and 4G cellular networks.

III. METHODOLOGY

Modern convex optimization / convex approximation underlies most of our work in this project. Specifically, semidefinite programming / semidefinite relaxation forms the basis of our design approach. More conventional optimization tools and concepts (e.g., water-filling, branch-and-bound) also come into play in certain algorithms, and particle filtering is the framework for our work on tracking of time-varying carrier signals.

IV. RESULTS

Our main results and findings are reviewed next, classified in four categories: Multiuser transmit beamforming (including sum capacity, quality of service, and fair service objectives); multiple input multiple output decoding; node localization; and tracking of time-varying carrier signals for synchronization, Doppler estimation, and signal intelligence applications. Conclusions are drawn and recommendations are made in the following section.

A. Multiuser Transmit Beamforming

A.1 Sum capacity objective

Multiuser transmit beamforming forms the core of our work under this project. The idea is to employ a transmit antenna array to create multiple beams directed towards the individual users, in order to increase the attainable throughput, as measured by sum capacity. In particular, we are interested in the practically important case of more users than transmit antennas, which requires user selection. Optimal solutions to this problem can be prohibitively complex for online implementation at the access point and entail so-called Dirty Paper (DP) precoding for known interference. Suboptimal solutions capitalize on multiuser (selection) diversity to achieve a significant

fraction of sum capacity at lower complexity cost. We analyzed the throughput performance in Rayleigh fading of a suboptimal greedy DP-based scheme proposed by Tu and Blum. We also proposed another user-selection method of the same computational complexity based on simple zero-forcing beamforming. Our results indicate that the proposed method attains a significant fraction of sum capacity, similar to Tu and Blums scheme, however at a much lower overall (design plus implementation) complexity; it thus, offers an attractive alternative to DP-based schemes.

A.2 Multicasting under Quality of Service (QoS) and Max-min Fair (MMF) objectives

Next, we considered the problem of transmit beamforming in the context of common information broadcasting or multicasting applications, wherein channel state information (CSI) is available at the transmitter. Unlike the usual blind isotropic broadcasting scenario, the availability of CSI allows transmit optimization. A minimum transmission power criterion was adopted, subject to prescribed minimum received signal-to-noise ratios (SNRs) at each of the intended receivers. A related maxmin SNR fair problem formulation was also considered subject to a transmitted power constraint. It was proven that both problems are NP-hard; however, suitable reformulation allows the successful application of semidefinite relaxation (SDR) techniques. SDR yields an approximate solution plus a bound on the optimum value of the associated cost/reward. SDR was motivated from a Lagrangian duality perspective, and its performance was assessed via pertinent simulations for the case of Rayleigh fading wireless channels. We found that SDR typically yields solutions that are within 3 to 4 dB of the optimum, which is often good enough in practice. In several scenarios, SDR generates exact solutions that meet the associated bound on the optimum value. This was illustrated using far-field beamforming for a uniform linear

transmit antenna array. Interestingly, these numerical experiments effectively led us to discover new and exact convex reformulations of the basic problem, via spectral factorization, applicable when the channel vectors are Vandermonde.

We also analyzed the approximation performance of the aforementioned broadcast beamforming algorithms theoretically. In particular, we showed that SDR provides an $O(m^2)$ approximation in the real case, and an $O(m)$ approximation in the complex case, where m is the total number of receivers. Moreover, we showed that these bounds are tight up to a constant factor. When the phase spread of the entries of the steering vectors is bounded away from $\pi/2$, we further established a certain constant factor approximation (depending on the phase spread but independent of the number of receivers, m and the number of transmit antennas, n) for both SDR and a convex quadratic programming restriction of the original NP-hard problem. Finally, we considered a related problem of finding a maximum norm vector subject to m convex homogeneous quadratic constraints. We showed that SDR provides an $O(1/ln(m))$ approximation, which is analogous to a result of Nemirovski, Roos and Terlaky for the real case.

Having settled the case of a single multicast group, we then generalized to multiple co-channel multicast groups. Two different design objectives were considered: minimizing total transmission power while guaranteeing a prescribed minimum signal-to-interference-plus-noise-ratio (SINR) at each receiver; and a fair approach maximizing the overall minimum SINR under a total power budget. The core problem is a multicast generalization of the multiuser downlink beamforming problem; the difference is that each transmitted stream is directed to multiple receivers, each with its own channel. Such generalization is relevant and timely, e.g., in the context of 802.16e wireless networks. The joint problem also contains single group multicast beamforming as a special case. The latter (and therefore also the former) is NP-hard. This motivates the

pursuit of computationally efficient quasi-optimal solutions. It was shown that Lagrangian relaxation coupled with suitable randomization / co-channel multicast power control loops yield computationally efficient high-quality approximate solutions. For a significant fraction of problem instances, the solutions generated this way are exactly optimal. Extensive numerical results using both simulated and measured wireless channels (courtesy of the University of Alberta, Canada) were presented to corroborate our main findings.

Whereas multi-group multicast transmit beamforming under SINR constraints is NP-hard in general, we have shown that, in the special case of Vandermonde steering vectors it is in fact a semidefinite problem, which can be exactly and efficiently solved.

We also considered various robust formulations for the problem of single-group multicasting, when the steering vectors are only approximately known. We obtained an elegant theoretical relationship between the optimal solutions of the original non-robust and associated robust formulations of the problem: the two are related via a simple (albeit solution-dependent) scaling. This relationship naturally suggests robust multicast beamforming approximation algorithms, through semidefinite relaxation of the original non-robust version of the problem.

B. Multiple Input Multiple Output Decoding

Multiple input multiple output (MIMO) communication links are now common in both commercial and tactical wireless networks, for spectral efficiency, fading, and jam-resilience considerations. The associated optimum vector decoding problem is known to be NP-hard. We developed two new computationally efficient MIMO decoding algorithms that afford very competitive symbol error rate (SER) performance. The first algorithm is a judicious combination of probabilistic data association (PDA) and sphere decoding (SD). The second is based on the

principle of semidefinite relaxation (SDR).

The key idea behind the hybrid PDA-SD detector is to reduce the dimension of the problem solved via SD by first running a single stage of the PDA to fix symbols that can be decoded with high reliability. This two-step algorithm attains a considerably better performance-complexity tradeoff than SD and PDA for low to moderate signal-to-noise ratio (SNR) or higher problem dimensions.

The second approach, based on SDR, has been specifically developed for MIMO systems employing high-order QAM constellations. The new approach affords improved detection performance compared to existing solutions of comparable worst-case complexity order, which is nearly cubic in the dimension of the transmitted symbol vector and independent of the constellation order for uniform QAM, or affine in the constellation order for non-uniform QAM.

C. Acquiring Channel State Information: Node Localization

Given a set of pairwise distance estimates between nodes, it is often of interest to generate a map of node locations. This is an old nonlinear estimation problem that has recently drawn interest in the signal processing community, due to the emergence of wireless sensor networks. Sensor maps are useful for estimating the spatial distribution of measured phenomena (including shadowing and fading), and for routing purposes. We proposed a two-stage algorithm that combines algebraic initialization and gradient descent. In particular, we borrowed an algebraic solution known as *Fastmap* from the database literature and adapted it to the sensor network context, using a specific choice of anchor/pivot nodes. The resulting estimates are fed to a gradient descent iteration. The overall algorithm offers very competitive performance at significantly lower complexity than existing solutions with similar estimation performance. For a certain mul-

tiplicative measurement noise model that is often adopted in the literature, we also derived the pertinent Cramér-Rao bound (CRB). Simulations indicate that the performance of our algorithm is close to the CRB when the network is (close to) fully connected, in the sense that every node can estimate its distance from all (most) other nodes. Our adaptation of Fastmap also turns out to make a big difference when used to initialize other iterative distributed estimation algorithms that have been developed specifically for sparse networks.

D. Synchronization, Doppler, and Intercept Issues: Particle Filtering Tools

In collaboration with Dr. Ananthram Swami, of ARL/Adelphi, we also investigated problems in time-varying frequency estimation. These appear in numerous pertinent applications: synchronization, Doppler frequency tracking, and signal intelligence, to name a few. We adopted a particle filtering (PF) framework, and contributed closed-form solutions for the optimal importance function, plus associated sampling procedures.

We first considered the problem of tracking the frequency and complex amplitude of a frequency-hopped complex sinusoid, using a novel stochastic state-space formulation that is naturally suited for the application of PF tools. The problem is of considerable interest for interference mitigation in frequency-hopped wireless networks, and for signal intelligence in military communications. The proposed particle filtering approach has a number of desirable features. It affords high-resolution estimates of carrier frequency and hop timing, manageable complexity (linear in the number of processed samples), and flexibility in tracking signals with irregular hopping patterns due to intentional timing jitter. The proposed state-space model is not only parsimonious, but fortuitous as well: it turns out that the associated optimal importance function (that minimizes the variance of the particle weights) can be computed in closed form, and thus

samples from it can be drawn using rejection techniques. Both prior and optimal importance sampling versions were developed and illustrated in pertinent simulations.

Next, we turned our attention to the problem of tracking the frequency and complex amplitude of a slowly time-varying (TV) harmonic signal. Similar to previous PF approaches to TV spectral analysis, we assumed that the frequency and complex amplitude evolve according to a Gaussian AR(1) model; but we concentrated on the important special case of a single TV harmonic. For this case, we showed that the optimal importance function can be computed in closed form. We also developed a suitable procedure to sample from the optimal importance function. The end result is a custom PF solution that is more efficient than generic ones, and can be used in a broad range of important applications that postulate a single TV harmonic component, e.g., TV Doppler estimation in communications and radar.

E. Publications

Summarizing the status of *journal papers* (5 appeared/accepted + 2 submitted for publication = 7 overall):

1. E. Karipidis, N.D. Sidiropoulos, Z.-Q. Luo, "Quality of Service and Max-min-fair Transmit Beamforming to Multiple Co-channel Multicast Groups, submitted to *IEEE Trans. on Signal Processing*, July 2006.

2. G. Latsoudas, N.D. Sidiropoulos, "A Fast and Effective Multidimensional Scaling Approach for Node Localization in Wireless Sensor Networks", submitted to *IEEE Trans. on Signal Processing*, July 2006.

3. N.D. Sidiropoulos, Z.-Q. Luo, "A Semidefinite Relaxation Approach to MIMO Detection for High-order QAM Constellations", *IEEE Signal Processing Letters*, to appear.

4. Z.-Q. Luo, N.D. Sidiropoulos, P. Tseng, S. Zhang, "Approximation Bounds for Quadratic Optimization with Homogeneous Quadratic Constraints", *SIAM Journal on Optimization*, to appear.

5. N.D. Sidiropoulos, T.N. Davidson, Z-Q (Tom) Luo, "Transmit Beamforming for Physical Layer Multicasting", *IEEE Trans. on Signal Processing*, 54(6, Part 1):2239-2251, June 2006.

6. G. Latsoudas, N.D. Sidiropoulos, "A Hybrid Probabilistic Data Association - Sphere Decoding Detector for Multiple-Input Multiple-Output Systems", *IEEE Signal Processing Letters*, 12(4):309-312, Apr. 2005.

7. G. Dimic, N.D. Sidiropoulos, "On Downlink Beamforming with Greedy User Selection: Performance Analysis and a Simple New Algorithm", *IEEE Trans. on Signal Processing*, 53(10):3857-3868, Oct. 2005.

Regarding *conference papers* (7 appeared/accepted; 2 in collaboration with Ananthram Swami, ARL/Adelphi, MD):

1. E. Tsakonas, N.D. Sidiropoulos, A. Swami, "Time-Frequency Analysis Using Particle Filtering: Closed-form Optimal Importance Function and Sampling Proceedure for a Single Time-varying Harmonic", in *Proc. Nonlinear Statistical Signal Processing Workshop: Classical, Unscented, and Particle Filtering Methods*, Sep. 13-15, 2006, Corpus Christi College, Cambridge, U.K., to appear.

2. N.D. Sidiropoulos, A. Swami, A. Valyrakis, "Tracking a Frequency-Hopped Signal Using Particle Filtering", *Proc. IEEE ICASSP 2006*, May 14-19, 2006, Toulouse, France.

3. E. Karipidis, N.D. Sidiropoulos, Z.-Q. (Tom) Luo, "Convex Transmit Beamforming For Downlink Multicasting to Multiple Co-channel Groups", *Proc. IEEE ICASSP 2006*, May 14-19, 2006, Toulouse, France.

4. E. Karipidis, N.D. Sidiropoulos, Z.-Q. (Tom) Luo, "Transmit Bemaforming to Multiple Co-channel Multicast Groups", in *Proc. IEEE CAMSAP 2005*, Dec. 12-14, 2005, Puerto Vallarta, Mexico.

5. G. Latsoudas, N.D. Sidiropoulos, "A Two-stage FASTMAP-MDS Approach for Node Localization in Sensor Networks", in *Proc. IEEE CAMSAP 2005*, Dec. 12-14, 2005, Puerto Vallarta, Mexico.

6. N.D. Sidiropoulos, T.N. Davidson, "Broadcasting with Channel State Information", in *Proc. IEEE SAM 2004*, July 18-21, Sitges, Barcelona, Spain.

7. G. Dimic, N.D. Sidiropoulos, "Low-Complexity Downlink Beamforming for Maximum Sum Capacity", in *Proc. IEEE ICASSP 2004*, May 17-21, Montreal, Quebec, Canada.

All journal and conference papers produced to date are included in the Annex.

Our research of course continues; in addition to the above, the following journal papers stemming from our *ICASSP06* conference paper are currently in progress

1. E. Karipidis, N.D. Sidiropoulos, Z.-Q. (Tom) Luo, "Far-field Multicast Beamforming of Uniform Linear Antenna Arrays is a Convex Problem", in preparation for submission to *IEEE Trans. on Signal Processing*.

2. E. Karipidis, N.D. Sidiropoulos, Z.-Q. (Tom) Luo, "Robust Transmit Beamforming for Multicasting", in preparation for submission to *IEEE Trans. on Signal Processing*.

V. Conclusions and Recommendations

A. *Multiuser Transmit Beamforming*

A.1 Sum capacity objective

We have considered two algorithms that capitalize on multiuser diversity to achieve a significant fraction of the multi-antenna downlink sum capacity when the number of users, M, is greater than the number of antennas, N. We have analyzed the throughput performance of the greedy zero-forcing dirty paper (gZF-DP) algorithm in independent Rayleigh fading, and characterized the pdf's of certain key parameters of interest. Determining the proper number of samples required for accurate Monte Carlo estimates is a difficult issue without a baseline. While the end result of gZF-DP performance analysis requires sequential numerical integration and is admittedly cumbersome, it does provide such a baseline and thus corroborates the results of Monte Carlo estimation. Also, numerical integration is simpler than Monte Carlo simulation for a small number of transmit antennas. Furthermore, our analysis allowed us to establish that at high SNR the throughput versus SNR slope of the gZF-DP algorithm is proportional to N.

We have also proposed another low-complexity algorithm, dubbed ZFS, which does not require DP coding at the transmitter. We have shown that the selection procedures of gZF-DP and ZFS algorithms have the same complexity order, $O(N^3M)$, which is significantly smaller than the complexity of the optimal algorithms when $M >> N$. We have evaluated the throughput performance of the ZFS algorithm via simulations. The results show that for a realistic number of transmit antennas, ZFS achieves a significant fraction of the throughput of gZF-DP and sum capacity, at a low coding and on-line computation cost. The simulation results also indicate that, at high SNR, ZFS achieves the same slope of throughput per dB of SNR as the

capacity-achieving strategy based on the use of DP coding for known interference cancellation and convex optimization.

Due to its simplicity, low complexity, and close to optimal performance, the proposed ZFS method offers an attractive alternative to earlier DP-based methods when $M >> N$. ZFS is hard to beat from a performance-complexity trade-off point of view. This is attributed to *multiple user selection* diversity, which generalizes the concept of multiuser diversity, due to Tse, by selecting to serve a *group* of users, versus a single user. We believe that ZFS has strong potential of being implemented in actual systems (there is recent follow-up work by Morgan, Huang, of Bell Labs / Lucent Technologies, as well as European industry R&D groups).

A.2 Multicasting under Quality of Service (QoS) and Max-min Fair (MMF) objectives

We have taken a new look at the broadcasting/multicasting problem when channel state information is available at the transmitter. We have proposed two pertinent problem formulations: minimizing transmitted power under multiple minimum received power constraints, and maximizing the minimum received power subject to a bound on the transmitted power.We have shown that both formulations are NP-hard optimization problems; however, their solution can often be well approximated using semidefinite relaxation tools. We have explored the relationship between the two formulations and also insights afforded by Lagrangian duality theory. In view of i) our extensive numerical experiments with simulated and measured data, verifying that semidefinite relaxation consistently yields good performance, ii) proof that the basic problem is NP-hard, and thus approximation is unavoidable, and iii) corroborating motivation provided by duality theory, we conclude that the approximate solutions provided herein offer useful designs across a broad range of applications.

The downlink beamforming problem was considered for the general case of multiple co-channel multicast groups, under two design criteria: QoS, in which we seek to minimize the total transmitted power while guaranteing a prescribed minimum SINR at all receivers; and a fair objective, in which we seek to maximize the minimum received SINR under a total power constraint. Both formulations contain single group multicast beamforming as a special case, and are therefore NP-hard. Computationally efficient quasi-optimal solutions were proposed by means of SDR and a combined randomization - multi-group multicast power control loop. Extensive numerical results have been presented, using both simulated (i.i.d. Rayleigh) and measured stationary outdoor wireless channel data, showing that the proposed algorithms yield high quality approximate solutions at a moderate complexity cost. Interestingly, our numerical findings indicate that the solutions generated by our algorithms are often exactly optimal, especially in the case of measured channels. In certain cases this optimality can be proven beforehand, and alternative convex reformulations of lower complexity have been constructed; in other cases, a theoretical worst-case bound on approximation accuracy has been derived, and shown to be tight.

Whereas multi-group multicast transmit beamforming under SINR constraints is NP-hard in general, we have shown that, in the special case of Vandermonde steering vectors it is in fact a semidefinite program, which can be efficiently solved. We have also considered robust beamforming solutions under channel uncertainty for the case of a single multicast group. For general steering vectors, we have shown that exact solutions of the robust and non-robust versions of the problem are related via a simple one-to-one scaling transformation. Since both problems are NP-hard, this suggests an algorithm to generate a quasi-optimal solution for one given a quasi-optimal solution for the other. In the important special case of Vandermonde steering vectors,

we have shown that the robust version of the problem is convex as well. This robust solution can be extended to the multi-group Vandermonde case.

B. Multiple Input Multiple Output Decoding

We have presented a two-stage hybrid PDA-SD algorithm for signal detection in MIMO systems. The basic idea is dimensionality reduction via hard decoding and cancellation of those symbols that can be quickly and reliably detected via a single PDA stage. In the V-BLAST scenario considered, simulations show that the proposed hybrid algorithm attains performance close to SD, at a complexity close to PDA. The dimensionality reduction idea can also be applied in conjunction with other variants of SD or SDR.

We have also proposed a new SDR approach for MIMO detection of high-order QAM constellations. The new approach is the simplest one in the class of SDR detectors for high-order QAM: its worst-case complexity is nearly cubic in the dimension of the transmitted symbol vector, and independent of the constellation order for uniform QAM / affine in the constellation order for non-uniform QAM. Under certain conditions, the new approach affords significant improvements in SER over prior methods. Specifically, the Sphere Decoder (SD) family of detectors exhibits a threshold behavior: it either works very well (for low-enough symbol vector dimension, order of the individual symbol constellation, and high-enough SNR) or it freezes. The threshold between the two regimes depends on a combination of these three factors. When SD works, it outperforms SDR in terms of complexity and SER performance. In difficult scenarios, SDR offers an attractive alternative relative to earlier solutions.

C. Acquiring Channel State Information: Node Localization

We have proposed a hybrid two-stage node localization algorithm that offers better accuracy than existing alternatives of the same (and, in certain cases, even higher) complexity order. The new algorithm employs Fastmap, coupled with judicious selection of anchor nodes that double as pivots, to generate a computationally cheap yet sufficiently accurate initialization for gradient descent. The new algorithm is particularly attractive (in terms of the offered performance-complexity trade-off) in the case of dense networks.

We also proposed using our adaptation of Fastmap as initialization for Costa's algorithm. The latter combination appears useful for sparse networks, in which case it attains better estimation performance than Fastmap followed by steepest descent (SD), albeit at a higher complexity cost. Our simulations indicate that, in the context of our present application, Fastmap+SD uniformly outperforms the classical principal component analysis (PCA)-based multi-dimensional scaling (MDS) algorithm, both in terms of complexity and in terms of estimation accuracy. We have also derived the pertinent CRB for the log-normal multiplicative measurement noise model, which was adopted for most of our simulations.

D. Synchronization, Doppler, and Intercept Issues: Particle Filtering Tools

We have developed three new particle filtering algorithms for tracking a frequency-hopped complex sinusoid, based on a novel stochastic state-space formulation. The algorithms range from a plain-vanilla version that uses the prior importance function, to a more advanced version that employs the optimal importance function, and, finally, an improvement of the latter using a problem-specific outer rejection loop. The two latter algorithms afford considerably better performance - complexity trade-offs.

We also revisited the important problem of tracking a single time-varying harmonic, whose frequency and complex amplitude evolve according to a linear Gaussian separable AR(1) model. A key difficulty in treating this model comes from the nonlinear measurement equation. For this model, we derived the optimal importance function in closed form. This yields interesting insights and opens up the possibility of designing particle filters that are more efficient than generic ones. We also derived a procedure to sample from this optimal importance function, using rejection and the concept of a dominating density. Our numerical experiments comparing the resulting filter to standard particle filters and the CRB show that the proposed PF algorithm has merits, particularly in terms of reducing the number of particles, and therefore memory requirements as well.

VI. Selected Bibliography, Classified by Topic

A. *Non-degraded Gaussian broadcast channel, sum capacity*

References

[1] G. Caire and S. Shamai (Shitz), "On the Achievable Throughput of a Multi-Antenna Gaussian Broadcast Channel," in *IEEE Trans. on Info. Theory*, vol. 49, no. 7, July 2003, pp. 1691–1706

[2] M. H. M. Costa, "Writing on Dirty Paper," *IEEE Trans. on Info. Theory,* vol. IT-29, no. 3, May 1983.

[3] D. Gore, R. W. Heath Jr., and A. Paulraj, "Transmit Selection in Spatial Multiplexing Systems," *IEEE Comm. Letters,* vol. 6., no. 11, Nov. 2002, pp. 491–493

[4] N. Jindal, W. Rhee, S. Vishwanath, S. Jafar and A. Goldsmith, "Sum Power Iterative Waterfilling for Gaussian Vector Broadcast Channels," *submitted to IEEE Trans. on Info. Theory,* July 2004; see also conference version in *Proc. ISIT2003*, Yokohama, Japan, July 2003.

[5] C.B. Peel, "On Dirty Paper Coding", *Signal Processing Magazine*, May 2003, pp. 112-113

[6] W. Rhee and J. M. Cioffi, "On the Capacity of Multiuser Wireless Channels With Multiple Antennas," *IEEE Trans. on Info. Theory.*, vol. 49, no. 10, Oct. 2003, pp. 2580–2595

[7] D. Samardzija and N. Mandayam, "Multiple Antenna Transmitter Optimization Schemes for Multiuser Systems," *Proc. of the IEEE Vehic. Tech. Conference,* Orlando, 2003, pp. 399–403

[8] Q. Spencer and M. Haardt, "Capacity and Downlink Transmission Algorithms for a Multi-user MIMO Channel," in *Proc. Of the 36th Asilomar Conf. On Sign. Syst. And Comp.,* Pacific Grove, CA, Nov. 2002.

[9] I. Emre Telatar, "Capacity of Multi-antenna Gaussian Channels," *European Trans. Telecomm.,* vol. 10, no. 6, Nov.-Dec. 1999, pp. 585–596

[10] Z. Tu and R. S. Blum, "Multiuser Diversity for a Dirty Paper Approach," *IEEE Comm. Letters,* vol. 7, no. 8, Aug. 2003, pp. 370–372

[11] S. Vishwanath, N. Jindal and A. Goldsmith, "Duality, Achievable Rates and Sum-Rate Capacity of Gaussian MIMO Broadcast Channels," *IEEE Trans. on Info. Theory*, vol. 49, no. 10, Oct. 2003, pp. 2658–2668

[12] P. Viswanath and D. Tse, "Sum Capacity of the Vector Gaussian Broadcast Channel and Uplink-Downlink Duality," *IEEE Trans. on Info. Theory*, vol. 49, no. 8, Aug. 2003, pp. 1912–1921

[13] H. Viswanathan, S. Venkatesan and H. Huang, "Downlink Capacity Evaluation of Cellular Networks with Known Interference Cancellation," *IEEE J. on Sel. Areas in Comm.*, vol. 21, no. 5, June 2003, pp. 802–811

[14] J. H. Winters, J. Salz, and R. D. Gitlin, "The Impact of Antenna Diversity on the Capacity of Wireless Communication Systems," *IEEE Trans. on Comm.,* vol. 42, n0. 2/3/4, Feb/Mar/Apr 1994, pp. 1740–1751.

[15] W. Yu and J. M. Cioffi, "Trellis Precoding for the Broadcast Channel," in *Proc. of Globecom 2001,* San Antonio, TX, November 2001, pp. 1344–1348.

[16] W. Yu and J. M. Cioffi, "Sum Capacity of a Gaussian Broadcast Channel," in *Proc. of IEEE Int. Symp. on Inform. Theory,* ISIT 2002, Lausanne, Switzerland, July 2002.

[17] R. Zamir, S. Shamai (Shitz), and U. Erez, "Nested Linear/Lattice Codes for Structured Multiterminal Binning," *IEEE Trans. on Inform. Theory*, vol. 48, no 6., June 2002, pp. 1250–1276

B. Transmit beamforming, multicasting, semidefinite relaxation

REFERENCES

[1] M. Bengtsson, and B. Ottersten, "Optimal and Suboptimal Transmit Beamforming", ch. 18 in *Handbook of Antennas in Wireless Communications*, L. C. Godara, Ed., CRC Press, Aug. 2001.

[2] S. Boyd, and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004; see also http://www.stanford.edu/∼boyd/cvxbook.html

[3] F.-R. Farrokhi, K.J.R. Liu, and L. Tassiulas, "Downlink Power Control and Base Station Assignment", *IEEE Communications Letters*, vol. 1, no. 4, pp. 102–104, July 1997.

[4] E. Karipidis, N.D. Sidiropoulos, Z.-Q. Luo, "Transmit Beamforming to Multiple Co-channel Multicast Groups", in *Proc. IEEE CAMSAP 2005*, Dec. 12-14, Puerto Vallarta, Mexico.

[5] E. Karipidis, N.D. Sidiropoulos, Z.-Q. Luo, "Convex Transmit Beamforming for Downlink Multicasting to Multiple Co-channel Groups", in *Proc. IEEE ICASSP 2006*, May 14-19, Toulouse, France.

[6] Z.-Q. Luo, N. D. Sidiropoulos, P. Tseng, and S. Zhang, "Approximation Bounds for Quadratic Optimization with Homogeneous Quadratic Constraints", *SIAM J. Optim.*, to appear.

[7] M. J. Lopez, "Multiplexing, scheduling, and multicasting strategies for antenna arrays in wireless networks", Ph.D. thesis, Dept. of Elect. Eng. and Comp. Sci., MIT, Cambridge, MA, 2002.

[8] N.D. Sidiropoulos, T.N. Davidson, "Broadcasting with Channel State Information", in *Proc. IEEE SAM 2004 Workshop*, vol. 1, pp. 489-493, Sitges, Barcelona, Spain, July 1821, 2004.

[9] N.D. Sidiropoulos, T.N. Davidson, and Z.-Q. Luo, "Transmit Beamforming for Physical Layer Multicasting", *IEEE Trans. on Signal Processing*, vol. 54, no. 6, pp. 2239-2251, June 2006.

[10] H. Wolkowicz, "Relaxations of Q2P", ch. 13.4 in *Handbook of Semidefinite Programming: Theory, Algorithms, and Applications*, H. Wolkowicz, R. Saigal, L. Vandenberghe, Eds., Kluwer Academic Publishers, 2000.

C. MIMO decoding, integer least squares

REFERENCES

[1] E. Agrell, T. Eriksson, A. Vardy, and K. Zeger, "Closest Point Search in Lattices," *IEEE Trans. Information Theory*, vol. 48, pp. 2201-2214, Aug. 2002.

[2] A. Chan and I. Lee, "A New Reduced-Complexity Sphere Decoder for Multiple Antenna Systems," in *Proc. of ICC 2002*, vol. 1, pp. 460-464, New York City, N.Y., April 28 - May 2, 2002.

[3] A. Duel-Hallen, "A family of multiuser decision-feedback detectors for asynchronous code-division multiple-access channels," *IEEE Trans. on Communications*, vol. 43, issue 234, pp. 421-434, Feb. 1995.

[4] J. Jaldén, B. Ottersten, "An Exponential Lower Bound on the Expected Complexity of Sphere Decoding," in *Proc. ICASSP 2004*, May 17-21, Montreal, Quebec, Canada.

[5] Z. Q. Luo, X. Luo, and M. Kisialiou, "An Efficient Quasi-Maximum Likelihood Decoder for PSK Signals," in *Proc. ICASSP2003*.

[6] W. K. Ma, P. C. Ching, and Z. Ding, "Semidefinite relaxation based multiuser detection for M-ary PSK multiuser systems," IEEE Trans. Signal Processing, vol. 52, no. 10, pp. 28622872, Oct. 2004.

[7] W.-K. Ma, T.N. Davidson, K.M. Wong, Z-Q Luo, P.-C. Ching, "Quasi-ML Multiuser Detection Using Semi-Definite Relaxation with Application to Synchronous CDMA," *IEEE Trans. on Signal Processing*, vol. 50, no. 4, pp. 912 -922, Apr. 2002.

[8] A. Mobasher, M. Taherzadeh, R. Sotirov, A.K. Khandani, "A near maximum likelihood decoding algorithm for MIMO systems based on semi-definite programming," in *Proc. 2005 Int. Symp. on Information Theory (ISIT 2005)*, pp. 1686-1690, Sep. 4-9, 2005.

[9] A. Stamoulis, G. B. Giannakis, A. Scaglione, "Block FIR decision-feedback equalizers for filterbank precoded transmissions with blind channel estimation capabilities," *IEEE Trans. on Communications*, vol. 49, no. 1, pp. 69-83, Jan. 2001.

[10] E. Viterbo and J. Boutros, "A Universal Lattice Code Decoder for Fading Channels," *IEEE Trans. Information Theory*, vol. 45, pp. 1639-1642, July 1999.

[11] R. Wang and G. B. Giannakis, "Approaching MIMO Capacity with Reduced-Complexity Soft Sphere-Decoding," in *Proc. WCNC 2004*, Atlanta, GA, March 21-25, 2004.

[12] A. Wiesel, Y. Eldar, and S. Shamai, "Semidefinite Relaxation for Detection of 16-QAM Signaling in MIMO Channels,", *IEEE Signal Processing Letters*, vol. 12, no. 9, pp. 653-656, Sep. 2005.

[13] W. Zhao, and G. B. Giannakis, "Sphere decoding algorithms with improved radius search," in *Proc. WCNC 2004*, Atlanta, GA, March 21-25, 2004.

D. Multidimensional scaling, node localization

REFERENCES

[1] P. Biswas, T.-C. Liang, T.-C. Wang and Y. Ye, "Semidefinite Programming Based Algorithms for Sensor Network Localization," to appear in *ACM Trans. on Sensor Networks*, 2006. See also http://www.stanford.edu/∼yyye/

[2] J. A. Costa, N. Patwari, A. O. Hero, "Distributed Multidimensional Scaling with Adaptive Weighting for Node Localization in Sensor Networks," *ACM Trans. on Sensor Networks*, submitted.

[3] C. Faloutsos, K. Lin, "FastMap: A Fast Algorithm for Indexing, Data-Mining and Visualization of Traditional and Multimedia Datasets," in *Proc. ACM SIGMOD*, vol. 24, no. 2, pp. 163-174, 1995.

[4] X. Ji, H. Zha "Sensor Positioning in Wireless Ad-hoc Sensor Networks Using Multidimensional Scaling," in *Proc. Infocom*, pp. 2652-2661, 2004.

[5] N. Patwari, A. Hero, M. Perkins, N. Correal, R. O'Dea, "Relative Location Estimation in Wireless Sensor Networks," *IEEE Trans. on Signal Processing*, vol. 51, no. 8, pp. 2137-2148, Aug. 2003.

[6] Y. Shang, W. Ruml, Y. Zhang, M. Fromherz, "Localization from Connectivity in Sensor Networks," *IEEE Trans. on Parallel and Distr. Systems*, vol. 15, no. 11, pp. 961-974, Nov. 2004.

[7] W.S. Torgerson, "Multidimensional Scaling: I. Theory and method," *Psychometrika*, vol. 17, pp. 401-419, 1952.

[8] W.S. Torgerson, "Multidimensional Scaling of Similarity," *Psychometrika*, vol. 30, pp. 379-393, 1965.

E. Tracking of slowly time-varying complex sinusoids, particle filtering

REFERENCES

[1] C. Andrieu, M. Davy, A. Doucet, "Improved Auxiliary Particle Filtering: Applications to Time-Varying Spectral Analysis", in *Proc. IEEE SSP 2001* Workshop, Singapore, Aug. 2001.

[2] M.S. Arulampalam, S. Maskell, N. Gordon, T. Clapp, "A tutorial on particle filters for nonlinear/non-Gaussian Bayesian tracking," *IEEE Trans. Signal Processing*, vol. 50, no. 2, pp. 174–188, Feb. 2002.

[3] P. Djuric, J.H. Kotecha, J. Zhang, Y. Huang, T. Ghirmai, M. Bugallo, J. Miguez, "Particle Filtering,", *IEEE Signal Processing Magazine*, pp. 19–38, Sep. 2003.

[4] A. Doucet, X. Wang, "Monte Carlo Methods for Signal Processing: A Review in the Statistical Signal Processing Context", *IEEE Signal Processing Magazine*, vol. 22, no. 6, pp. 152-170, Nov. 2005.

[5] C. Dubois, M. Davy, J. Idier, "Tracking of Time-Frequency Components Using Particle Filtering", in *Proc. IEEE ICASSP 2005*, March 18-23, 2005, Philadelphia, PA, U.S.A.

F. Tracking of frequency-hopped signals

REFERENCES

[1] L. Aydin and A. Polydoros, "Hop-timing estimation for FH signals using a coarsely channelized receiver," *IEEE Trans. Communcations*, vol. 44, no. 4, pp. 516–526, Apr. 1996.

[2] S. Barbarossa and A. Scaglione, "Parameter estimation of spread spectrum frequency-hopping signals using time-frequency distributions," in *Proc. Signal Proc. Advances in Wireless Communications*, pp. 213–216, Apr. 1997.

[3] X. Liu, N. D. Sidiropoulos, and A. Swami, "Blind high resolution localization and tracking of multiple frequency hopped signals," *IEEE Trans. Signal Processing*, vol. 50, no. 4, pp. 889–901, Apr. 2002.

[4] X. Liu, N. D. Sidiropoulos, and A. Swami, "Joint Hop Timing and Frequency Estimation for Collision Resolution in Frequency Hopped Networks," *IEEE Trans. Wireless Communications*, to appear, Nov. 2005.

[5] M. K. Simon, U. Cheng, L. Aydin, A. Polydoros, and B. K. Levitt, "Hop timing estimation for noncoherent frequency-hopped M-FSK intercept receivers," *IEEE Trans. Communications*, vol. 43, no. 2/3/4, pp. 1144–1154, Feb./Mar./Apr. 1995.

VII. Annex: Publications

All journal and conference papers are included in the following pages.

Quality of Service and Max-min-fair Transmit Beamforming to Multiple Co-channel Multicast Groups

Eleftherios Karipidis[1], *Student Member, IEEE,* Nicholas D. Sidiropoulos[2], *Senior Member, IEEE,* and Zhi-Quan Luo[3], *Senior Member, IEEE*

Abstract

The problem of transmit beamforming to multiple co-channel multicast groups is considered, from two viewpoints: minimizing total transmission power while guaranteeing a prescribed minimum signal-to-interference-plus-noise-ratio (SINR) at each receiver; and a "fair" approach maximizing the overall minimum SINR under a total power budget. The core problem is a multicast generalization of the multiuser downlink beamforming problem; the difference is that each transmitted stream is directed to multiple receivers, each with its own channel. Such generalization is relevant and timely, e.g., in the context of 802.16e wireless networks. The joint problem also contains single group multicast beamforming as a special case. The latter (and therefore also the former) is NP-hard. This motivates the pursuit of computationally efficient quasi-optimal solutions. It is shown that Lagrangian relaxation coupled with suitable randomization / co-channel multicast power control loops yield computationally efficient high-quality approximate solutions. For a significant fraction of problem instances, the solutions generated this way are exactly optimal. Extensive numerical results using both simulated and measured wireless channels are presented to corroborate our main findings.

Index Terms

Broadcasting, multicasting, downlink beamforming, semidefinite relaxation, convex optimization; TSP EDICS: SAM-BEAM (Beamforming)

Submitted to *IEEE Trans. on Signal Processing*, July 4, 2006. Earlier version of part of this work appears in conference form in the *Proc. of IEEE CAMSAP*, Dec. 12-14, 2005, Puerto Vallarta, Mexico. E. Karipidis was supported in part by the GSRT under a PENED grant, and the E.U. under the U-BROAD FP6 project #506790. N. Sidiropoulos was supported in part by the U.S. ARO under ERO Contract N62558-03-C-0012, and the E.U. under FP6 project NEWCOM. Z.-Q. Luo was supported in part by the National Science Foundation, Grant No. DMS-0312416.

[1] The author is with the Department of Electronic and Computer Engineering, Technical University of Crete, 73100 Chania - Crete, Greece; Phone: +30-28210-37324, E-mail: karipidis@telecom.tuc.gr

[2] The author is with the Department of Electronic and Computer Engineering, Technical University of Crete, 73100 Chania - Crete, Greece; Fax: +30-28210-37542, Phone: +30-28210-37227, E-mail: nikos@telecom.tuc.gr

[3] The author is with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455, U.S.A.; E-mail: luozq@ece.umn.edu

I. INTRODUCTION

The proliferation of streaming media (digital audio, video, IP radio), peer-to-peer services, large-scale software updates, and profiled newscasts over the wireline Internet has brought renewed interest in multicast routing protocols. These protocols were originally conceived and have since evolved under the "wireline premise": the physical network is a graph comprising point-to-point links that do not interfere with each other at the physical layer. Today, multicast routing protocols operate at the network or application layer, using either controlled flooding or minimum spanning tree access.

As wireless networks become ever more ubiquitous, and wireless becomes the choice for not only the "last hop" but also suburban- and metropolitan-area backbones, wireless multicasting solutions are needed to account for and exploit the idiosyncracies of the wireless medium. Wireless is inherently a broadcast medium, where it is possible to reach multiple destinations with a single transmission; different co-channel transmissions are interfering with one-another at the intended destination(s); and links are subject to fading and shadowing, in addition to co-channel interference.

The broadcast advantage of wireless has of course been exploited since the early days of radio. The interference problem was dealt with by allocating different frequency bands to the different stations, and transmission was mostly isotropic or focused towards a specific service area.

Today, the situation with wireless networks is much different. First, transmissions need not be "blind". Many wireless network standards provision the use of transmit antenna arrays. Using baseband beamforming, it is possible to steer energy in the direction(s) of the intended users, whose locations (or, more generally, channels) can often be accurately estimated. Second, the push towards higher capacity and end-user rates necessitates co-channel transmission which exploits the spatial diversity in the user population (*spatial multiplexing*). Third, quality of service is an important consideration, especially in wireless backhaul solutions like 802.16e. Finally, due to co-channel interference, wireless multicasting cannot be dealt with in isolation, one group at a time; a joint solution is needed.

The problem of transmit beamforming towards a (single) group of users was first considered in the Ph.D. thesis of Lopez [9], using the averaged (over all users in the group) received Signal to Noise Ratio (SNR) as the design criterion. The solution boils down to a relatively simple eigenvalue problem, but no SNR guarantee is provided this way: some users may get really poor SNR [11]. This is not acceptable in multicasting applications, because it is the worst SNR that determines the *common* information rate. Quality of service (providing a guaranteed minimum received SNR to every user) and max-min-fair (maximizing the smallest received SNR) designs were first proposed in [10], [11], where it was shown that the core problem is NP-hard, yet high-quality approximate solutions can be obtained using relaxation techniques based on semidefinite programming (SDP). The latter is a class of convex optimization problems which can be solved in polynomial time by powerful interior point

methods.

As already mentioned, designing a transmit beamformer separately for each multicast group can be far from optimal, due to inter-group interference. In this paper, we consider the joint design problem under quality of service and max-min-fair criteria. In addition to semidefinite relaxation ideas, our solutions entail a co-channel multicast power control component, which can be viewed as a generalization of multiuser power control ideas for the cellular downlink (e.g., see [3] and references therein). The multiuser downlink beamforming problem (e.g., see [1] and references therein) can be viewed as a special case of our formulation, where each multicast group consists of a single receiver.

A carefully designed suite of numerical results is used to demonstrate the efficacy of our designs, including extensive results using measured wireless channel data.

II. Data Model and Problem Statement

Consider a wireless scenario incorporating a single transmitter with N antenna elements and M receivers, each with a single antenna. Let $\mathbf{h}_i$ denote the $N \times 1$ complex vector that models the propagation loss and phase shift of the frequency-flat quasi-static channel from each transmit antenna to the receive antenna of user $i \in \{1, \ldots, M\}$. Let there be a total of $1 \leq G \leq M$ multicast groups, $\{\mathcal{G}_1, \ldots, \mathcal{G}_G\}$, where $\mathcal{G}_k$ contains the indices of receivers participating in multicast group k, and $k \in \{1, \ldots, G\}$. Each receiver listens to a single multicast; thus $\mathcal{G}_k \cap \mathcal{G}_l = \emptyset$, $l \neq k$, $\cup_k \mathcal{G}_k = \{1, \ldots, M\}$, and, denoting $G_k := |\mathcal{G}_k|$, $\sum_{k=1}^{G} G_k = M$.

Let $\mathbf{w}_k$ denote the beamforming weight vector applied to the N transmitting antenna elements to generate the spatial channel for transmission to group k (see Fig. 1). Then the signal transmitted by the antenna array is equal to $\sum_{k=1}^{G} \mathbf{w}_k^H s_k(t)$, where $s_k(t)$ is the temporal information-bearing signal directed to receivers in multicast group k. Note that the above setup includes the case of *broadcasting* (a *single* multicast group, $G = 1$) [11], as well as the case of individual information transmission to each receiver ($G = M$) by means of spatial multiplexing (see, e.g., [1]). If each $s_k(t)$ is zero-mean, temporally white with unit variance, and the waveforms $\{s_k(t)\}_{k=1}^{G}$ are mutually uncorrelated, then the total power radiated by the transmitting antenna array is equal to $\sum_{k=1}^{G} \|\mathbf{w}_k\|_2^2$.

The joint design of transmit beamformers can then be posed as the problem of minimizing the total radiated power subject to meeting prescribed SINR constraints γ_i at each of the M receivers

$$\mathcal{Q}: \quad \min_{\{\mathbf{w}_k \in \mathbb{C}^N\}_{k=1}^{G}} \sum_{k=1}^{G} \|\mathbf{w}_k\|_2^2$$

$$\textit{s.t.}: \quad \frac{|\mathbf{w}_k^H \mathbf{h}_i|^2}{\sum_{l \neq k} |\mathbf{w}_l^H \mathbf{h}_i|^2 + \sigma_i^2} \geq \gamma_i, \ \forall i \in \mathcal{G}_k, \ \forall k \in \{1, \ldots, G\}.$$

Problem $\mathcal{Q}$ contains the associated broadcasting problem ($G = 1$) as a special case; from this and [11], it immediately follows that

Claim 1: Problem $\mathcal{Q}$ is NP-hard.

This motivates (cf. [4]) the pursuit of sensible approximate solutions to problem $\mathcal{Q}$. [1]

III. Relaxation

Towards this end, let us define $\mathbf{Q}_i := \mathbf{h}_i\mathbf{h}_i^H$ and $\mathbf{X}_k := \mathbf{w}_k\mathbf{w}_k^H$, and note that $|\mathbf{w}_k^H\mathbf{h}_i|^2 = \mathbf{h}_i^H\mathbf{w}_k\mathbf{w}_k^H\mathbf{h}_i = \mathrm{trace}(\mathbf{h}_i^H\mathbf{w}_k\mathbf{w}_k^H\mathbf{h}_i) = \mathrm{trace}(\mathbf{h}_i\mathbf{h}_i^H\mathbf{w}_k\mathbf{w}_k^H) = \mathrm{trace}(\mathbf{Q}_i\mathbf{X}_k)$. Note that $\mathbf{X}_k = \mathbf{w}_k\mathbf{w}_k^H$ for some $\mathbf{w}_k \in \mathbb{C}^N$ if and only if $\mathbf{X}_k \succeq \mathbf{0}$ and $\mathrm{rank}(\mathbf{X}_k) = 1$. It follows that problem $\mathcal{Q}$ can be *equivalently* reformulated as

$$\min_{\{\mathbf{X}_k \in \mathbb{C}^{N\times N}\}_{k=1}^G} \sum_{k=1}^G \mathrm{trace}(\mathbf{X}_k)$$

$$\textit{s.t.} \;:\quad \mathrm{trace}(\mathbf{Q}_i\mathbf{X}_k) \geq \gamma_i \sum_{l\neq k} \mathrm{trace}(\mathbf{Q}_i\mathbf{X}_l) + \gamma_i\sigma_i^2,$$

$$\forall i \in \mathcal{G}_k,\ \forall k \in \{1,\ldots,G\},$$

$$\mathbf{X}_k \succeq \mathbf{0},\ \forall k \in \{1,\ldots,G\},$$

$$\mathrm{rank}(\mathbf{X}_k) = 1,\ \forall k \in \{1,\ldots,G\},$$

where the fact that the terms in the denominator are all nonnegative has also been taken into account. Dropping the last G rank-one constraints, which are nonconvex, we arrive at the following relaxation of problem $\mathcal{Q}$

$\mathcal{Q}_r$:

$$\min_{\{\mathbf{X}_k \in \mathbb{C}^{N\times N}\}_{k=1}^G,\ \{s_i\in\mathbb{R}\}_{i=1}^M} \sum_{k=1}^G \mathrm{trace}(\mathbf{X}_k)$$

$$\textit{s.t.} \;:\quad \mathrm{trace}(\mathbf{Q}_i\mathbf{X}_k) - \gamma_i \sum_{l\neq k} \mathrm{trace}(\mathbf{Q}_i\mathbf{X}_l) - s_i = \gamma_i\sigma_i^2,$$

$$\forall i \in \mathcal{G}_k,\ \forall k \in \{1,\ldots,G\},$$

$$s_i \geq 0,\ \forall i \in \{1,\ldots,M\},$$

$$\mathbf{X}_k \succeq \mathbf{0},\ \forall k \in \{1,\ldots,G\},$$

where M nonnegative real "slack" variables s_i have been introduced, in order to convert the first M linear inequality constraints to M linear equality constraints, plus M nonnegativity constraints.

[1] Note that other special cases of problem $\mathcal{Q}$ are not NP-hard: e.g., the multiuser downlink beamforming problem ($G = M$) is a Second Order Cone Program (SOCP) [1]; see also [7] for a restriction on the channel vectors that enables convex reformulation and thereby efficient solution of the problem.

Next, we seek to express the equality constraints as linear combinations of the unknown vector $\mathbf{x} = [\mathrm{vec}(\mathbf{X}_1)^T \cdots \mathrm{vec}(\mathbf{X}_G)^T]^T$, which is formed by stacking the columns of the $\mathbf{X}_k$ matrices. Towards this end, the $G \times 1$ vectors

$$\mathbf{a}_i = (\gamma_i + 1)\mathbf{e}_{k(i)} - \gamma_i \mathbf{1}_G, \quad \forall i \in \{1, \ldots, M\},$$

are introduced, whose $k(i)$-th element is equal to one, whereas all others are set to $-\gamma_i$. Here, $\mathbf{e}_{k(i)}$ is the $G \times 1$ vector indicating the multicast group $k(i)$ that user i belongs to, and $\mathbf{1}_G$ is the $G \times 1$ all-ones vector. Using $\mathbf{a}_i$ we can recast the equality constraints as

$$[\mathbf{a}_i \otimes \mathrm{vec}(\mathbf{Q}_i^T)]^T \mathbf{x} - s_i = \gamma_i \sigma_i^2, \quad \forall i \in \{1, \ldots, M\},$$

where $\otimes$ denotes the Kronecker product. Finally, the relaxed problem $\mathcal{Q}_r$ is written as

$$\begin{aligned} \mathcal{Q}_r: \quad & \min_{\mathbf{x} \in \mathbb{C}^{GN^2},\, \{s_i \in \mathbb{R}\}_{i=1}^M} [\mathbf{1}_G \otimes \mathrm{vec}(\mathbf{I}_N)]^T \mathrm{vec}(\mathbf{x}) \\ \textit{s.t.}: \quad & [\mathbf{a}_i \otimes \mathrm{vec}(\mathbf{Q}_i^T)]^T \mathbf{x} - s_i = \gamma_i \sigma_i^2, \\ & \forall i \in \mathcal{G}_k,\ \forall k \in \{1, \ldots, G\}, \\ & s_i \geq 0,\ \forall i \in \{1, \ldots, M\} \\ & \mathbf{X}_k \succeq \mathbf{0},\ \forall k \in \{1, \ldots, G\}. \end{aligned}$$

Here, $\mathbf{I}_N$ is the identity matrix of size $N \times N$. Problem $\mathcal{Q}_r$ is a *Semi-Definite Program* (SDP), expressed in the primal standard form used by SDP solvers, such as SeDuMi [12]. This SDP has G matrix variables of size $N \times N$, and M linear constraints. Interior point methods will take $O(\sqrt{GN}\log(1/\epsilon))$ iterations, with each iteration requiring at most $O(G^3N^6 + MGN^2)$ arithmetic operations, where the parameter ϵ represents the solution accuracy at the algorithm's termination. SeDuMi uses interior point methods to solve such SDP problems efficiently. Actual runtime complexity will usually scale far slower with G, N, M than this worst-case bound.

IV. Obtaining an Approximate Solution to Problem $\mathcal{Q}$

Problem $\mathcal{Q}$ may not admit a feasible solution (counter-examples may be easily constructed), but if it does, the aforementioned approach will yield a solution to problem $\mathcal{Q}_r$. Due to relaxation, this solution will not, in general, consist of rank-one blocks. In order to obtain a high-quality approximate solution of problem $\mathcal{Q}$, the concept of *randomization* can be employed to generate candidate beamforming vectors in the span of the respective transmit covariance matrices. The main difference relative to the simpler broadcast case ($G = 1$) considered in [11], is that here we cannot simply "scale up" the candidate beamforming vectors generated during randomization to satisfy the SINR constraints

of problem $\mathcal{Q}$. The reason is that, in contrast to [11], we herein deal with an interference scenario, and boosting one group's beamforming vector also increases interference to nodes in other groups. Whether it is feasible to satisfy the constraints for a given set of candidate beamforming vectors is also an issue here. Let $a_{k,i} := |\mathbf{w}_k^H \mathbf{h}_i|^2$ denote the signal power received at receiver i from the stream directed towards users in multicast group k. Let $\beta_k := \|\mathbf{w}_k\|_2^2$, and p_k denote the power boost (or reduction) factor for multicast group k. Then the following *Multi-Group Power Control* ($\mathcal{MGPC}$) problem emerges in converting candidate beamforming vectors to a candidate solution of problem $\mathcal{Q}$.

$$
\begin{aligned}
\mathcal{MGPC}: \quad & \\
& \min_{\{p_k \in \mathbb{R}\}_{k=1}^{G}} \sum_{k=1}^{G} \beta_k p_k \\
\textit{s.t.}: \quad & \frac{p_k a_{k,i}}{\sum_{l \neq k} p_l a_{l,i} + \sigma_i^2} \geq \gamma_i, \\
& \forall i \in \mathcal{G}_k, \ \forall k \in \{1, \ldots, G\}, \\
& p_k \geq 0, \ \forall k \in \{1, \ldots, G\}.
\end{aligned}
$$

As in Section III, taking advantage of the fact that the terms in the denominator are all nonnegative and introducing M nonnegative real "slack" variables s_i, problem $\mathcal{MGPC}$ can be equivalently reformulated as

$$
\begin{aligned}
\mathcal{MGPC}: \quad & \\
& \min_{\{p_k \in \mathbb{R}\}_{k=1}^{G}, \ \{s_i \in \mathbb{R}\}_{i=1}^{M}} \sum_{k=1}^{G} \beta_k p_k \\
\textit{s.t.}: \quad & p_k a_{k,i} - \gamma_i \sum_{l \neq k} p_l a_{l,i} - s_i = \gamma_i \sigma_i^2, \\
& \forall i \in \mathcal{G}_k, \ \forall k \in \{1, \ldots, G\}, \\
& p_k \geq 0, \ \forall k \in \{1, \ldots, G\}, \\
& s_i \geq 0, \ \forall i \in \{1, \ldots, M\}.
\end{aligned}
$$

Towards transforming the $\mathcal{MGPC}$ problem formulation to the primal standard form used by convex optimization problem solvers, such as SeDuMi, we denote $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_G]^T$, $\mathbf{p} = [p_1, \ldots, p_G]^T$, and $\boldsymbol{\alpha}_i = [\alpha_{1,i}, \ldots, \alpha_{G,i}]^T$. We can now recast problem $\mathcal{MGPC}$ as

$$
\begin{array}{|ll|}
\hline
\mathcal{MGPC}: & \\
 & \min\limits_{\mathbf{p}\in\mathbb{R}^G,\,\{s_i\in\mathbb{R}\}_{i=1}^M} \boldsymbol{\beta}^T\mathbf{p} \\
\textit{s.t.}: & [\mathbf{a}_i \odot \boldsymbol{\alpha}_i]^T\mathbf{p} - s_i = \gamma_i\sigma_i^2, \\
 & \forall i \in \mathcal{G}_k,\ \forall k \in \{1,\ldots,G\}, \\
 & p_k \geq 0,\ \forall k \in \{1,\ldots,G\}, \\
 & s_i \geq 0,\ \forall i \in \{1,\ldots,M\}, \\
\hline
\end{array}
$$

where $\mathbf{a}_i$ are the $G \times 1$ vectors introduced in Section III and $\odot$ stands for element-wise multiplication (the Hadamard product). Problem $\mathcal{MGPC}$ is a *Linear Program* (LP) with G nonnegative variables and M linear inequality constraints. Interior point methods can either find the problem infeasible or generate an ϵ-optimal solution in $O(\sqrt{G}\log(1/\epsilon))$ iterations, each requiring at most $O(G^3 + MG)$ arithmetic operations. SeDuMi can be used to find its optimum solution. Note that SeDuMi will also yield an infeasibility certificate in case the $\mathcal{MGPC}$ problem is not solvable for a particular beamforming configuration. This is useful to determine the feasibility of a candidate beamforming configuration.

For $G = M$ (independent information transmission to each receiver), problem $\mathcal{Q}_r$ is in fact equivalent to (not a relaxation of) problem $\mathcal{Q}$, see [1]; likewise, problem $\mathcal{MGPC}$ reduces to the well-known multiuser downlink power control problem, which can be solved using simpler means (e.g., [3]): matrix inversion and iterative descent algorithms. In this special case, (in)feasibility can be determined from the spectral radius of a certain "connectivity" matrix. Similar simplifications for the general instance of $\mathcal{MGPC}$ are perhaps possible, but appear highly non-trivial. At any rate, interior point LP routines are very efficient, hence this is not a major issue. The overall algorithm for obtaining an approximate solution to problem $\mathcal{Q}$ can be summarized as follows:

1) Relaxation: Solve problem $\mathcal{Q}_r$, using a SDP solver (e.g. SeDuMi). Denote the solution $\{\mathbf{X}_k\}_{k=1}^G$.
2) Randomization / Scaling Loop: For each k, generate a vector in the span of $\mathbf{X}_k$, using the Gaussian randomization technique (randC) in [11]. If, for some k, $\mathrm{rank}(\mathbf{X}_k) = 1$, then use the principal component instead. Next, feed the resulting set of candidate beamforming vectors $\{\mathbf{w}_k\}_{k=1}^G$ into problem $\mathcal{MGPC}$ and solve it using SeDuMi. If the particular instance of $\mathcal{MGPC}$ is infeasible or yields a larger $\mathcal{MGPC}$ objective than previously checked candidates, discard the proposed set of candidate beamforming vectors; else, record the solution and associated objective value.

The *quality* of approximate solutions to problem $\mathcal{Q}$ generated this way can be checked against the lower bound on transmit power obtained in solving problem $\mathcal{Q}_r$. This bound can be further motivated

from a duality perspective, as in [11]; that is, the aforementioned relaxation lower bound is in fact the tightest lower bound on the optimum of problem $\mathcal{Q}$ attainable via Lagrangian duality [2]. This follows from arguments in [13] (see also the single-group case in [11]), due to the fact that problem $\mathcal{Q}$ is a quadratically constrained quadratic program.

V. Joint Max-Min Fair Beamforming

In this section, we consider the related problem of maximizing the minimum SINR, received by any of the M intended users irrespective of the multicast group they belong to, subject to an upper bound P on the total transmitted power. This problem formulation is a generalization of the respective max-min fair transmit beamforming problem towards a single multicast group, which was considered in [11]. The key difference is that here we seek to maximize a SINR, instead of a SNR; that is, the beamforming vectors, which are to be optimized, appear in the numerator as well as in the denominator of the objective function. Specifically, the joint max-min fair (JMMF) transmit beamforming design is formulated as

$$\mathcal{F}:$$

$$\max_{\{\mathbf{w}_k\in\mathbb{C}^N\}_{k=1}^G} \min_{k\in\{1,\ldots,G\}} \min_{i\in\mathcal{G}_k} \frac{|\mathbf{w}_k^H\mathbf{h}_i|^2}{\sum_{\ell\neq k}|\mathbf{w}_\ell^H\mathbf{h}_i|^2+\sigma_i^2}$$

$$\textit{s.t.}: \quad \sum_{k=1}^{G}\|\mathbf{w}_k\|_2^2 \leq P.$$

Since problem $\mathcal{F}$ contains as a special case the associated broadcasting problem ($G=1$), it follows from [11] that

Claim 2: Problem $\mathcal{F}$ is NP-hard.

The inequality constraint on the total transmit power will be met with equality at an optimum. Otherwise, one could multiply all beam vectors by a constant $c>1$, thereby increasing the minimum SINR (note that $\sigma_i^2>0$). We may therefore focus on the equality constrained problem and denote this as $\mathcal{F}$ from now on.

Claim 2 motivates the pursuit of sensible approximate solutions to the JMMF problem. Towards this end, we introduce an auxiliary positive real variable t and rewrite the (equality constrained)

JMMF downlink beamforming problem $\mathcal{F}$ as follows

$$\max_{\{\mathbf{w}_k\in\mathbb{C}^N\}_{k=1}^G,\ t\in\mathbb{R}} \quad t$$

$$s.t.\ : \quad \frac{|\mathbf{w}_k^H\mathbf{h}_i|^2}{\sum_{\ell\neq k}|\mathbf{w}_\ell^H\mathbf{h}_i|^2+\sigma_i^2}\geq t,$$

$$\forall k\in\{1,\ldots,G\},\ \forall i\in\mathcal{G}_k,$$

$$\sum_{k=1}^G\|\mathbf{w}_k\|_2^2=P,\quad \text{and}\quad t\geq 0.$$

Then, using the matrices $\mathbf{Q}_i$ and $\mathbf{X}_k$ introduced in Section III, we can further recast problem $\mathcal{F}$ as

$$\max_{\{\mathbf{X}_k\in\mathbb{C}^{N\times N}\}_{k=1}^G,\ t\in\mathbb{R}} \quad t$$

$$s.t.\ : \frac{\mathrm{trace}(\mathbf{Q}_i\mathbf{X}_k)}{\sum_{\ell\neq k}\mathrm{trace}(\mathbf{Q}_i\mathbf{X}_\ell)+\sigma_i^2}\geq t,$$

$$\forall k\in\{1,\ldots,G\},\ \forall i\in\mathcal{G}_k,$$

$$\sum_{k=1}^G\mathrm{trace}(\mathbf{X}_k)=P,$$

$$\mathrm{rank}(\mathbf{X}_k)=1,\ \forall k\in\{1,\ldots,G\},$$

$$\mathbf{X}_k\succeq\mathbf{0},\ \forall k\in\{1,\ldots,G\},\quad \text{and}\quad t\geq 0.$$

Finally, dropping the nonconvex rank constraints we obtain the following relaxation of the original problem $\mathcal{F}$

$\mathcal{F}_r$:

$$\max_{\{\mathbf{X}_k\in\mathbb{C}^{N\times N}\}_{k=1}^G,\ t\in\mathbb{R}} \quad t$$

$$s.t.\ : \quad \mathrm{trace}(\mathbf{Q}_i\mathbf{X}_k)-t\left(\sum_{\ell\neq k}\mathrm{trace}(\mathbf{Q}_i\mathbf{X}_\ell)+\sigma_i^2\right)\geq 0,$$

$$\forall k\in\{1,\ldots,G\},\ \forall i\in\mathcal{G}_k,$$

$$\sum_{k=1}^G\mathrm{trace}(\mathbf{X}_k)=P,$$

$$\mathbf{X}_k\succeq\mathbf{0},\ \forall k\in\{1,\ldots,G\},\quad \text{and}\quad t\geq 0,$$

where we have also taken into account the fact that the terms in the denominators of the first M inequality constraints are all nonnegative. Problem $\mathcal{F}_r$ has a linear objective function, 1 linear equality constraint, G positive semidefinite constraints, and 1 nonnegativity constraint; however, it is nonconvex, due to the first M nonlinear inequality constraints.

A solution to the relaxed problem $\mathcal{F}_r$ can be found by means of bisection over SDP problems, as explained next. Let t^* be the optimum value of problem $\mathcal{F}_r$. A feasible solution of $\mathcal{F}_r$ that is at most $\epsilon > 0$ away from t^* can be generated as follows. Let $[L, U]$ be an interval containing t^*. We begin by setting $L = 0$, $U = P \min\limits_{i\in\{1,\dots,M\}} \dfrac{\|\mathbf{h}_i\|_2^2}{\sigma_i^2}$, where the lower bound follows from non-negativity of t^* and the upper bound follows from the Cauchy-Schwartz inequality. Given $[L, U]$, the convex feasibility problem $\mathcal{FP}$, shown in the box below, is solved at the midpoint $t = (L + U)/2$ of the interval. If problem $\mathcal{FP}$ is feasible for the given choice of t, we set $L := t$; otherwise $U := t$. Thus, in each iteration the interval is halved. Repeating until $U - L \leq \epsilon$ requires only $N_{\text{iter}} = \lceil \log_2((U - L)/\epsilon) \rceil$ iterations. In practice, 10-12 iterations are usually enough for typical problem setups.

The convex feasibility problem $\mathcal{FP}$ is formulated, for any choice of the positive real t, as

$$
\begin{aligned}
&\mathcal{FP}: \\
&\qquad \text{find } \mathbf{v} \\
&\textit{s.t.}: \quad \operatorname{trace}(\mathbf{Q}_i\mathbf{X}_k) - t\sum_{\ell\neq k}\operatorname{trace}(\mathbf{Q}_i\mathbf{X}_\ell) - s_i = t\sigma_i^2, \\
&\qquad \forall k \in \{1,\dots,G\},\ \forall i \in \mathcal{G}_k, \\
&\qquad \sum_{k=1}^{G}\operatorname{trace}(\mathbf{X}_k) = P, \\
&\qquad \mathbf{X}_k \succeq \mathbf{0},\ \forall k \in \{1,\dots,G\}, \\
&\qquad s_i \geq 0,\ \forall i \in \{1,\dots,M\},
\end{aligned}
$$

where M nonnegative real "slack" variables s_i have been introduced to convert the linear inequality constraints to linear equality plus nonnegativity constraints. Here, $\mathbf{v} \in \mathbb{R}^M \times \mathbb{C}^{GN^2}$ denotes the variable vector $\mathbf{v} = [\mathbf{s}^T\ \operatorname{vec}(\mathbf{X}_1)^T\ \cdots\ \operatorname{vec}(\mathbf{X}_G)^T]^T$, where the vector $\mathbf{s} = [s_1 \cdots s_M]^T$ contains the "slack" variables. The feasibility problem $\mathcal{FP}$ is comprised of an objective function, set to zero, and $M+1$ linear equality constraints, G positive semidefinite constraints, and M nonnegativity constraints. Hence, it is a SDP problem expressed in the standard primal form. Thus, for each iteration of the aforementioned bisection algorithm, problem $\mathcal{FP}$ can be efficiently solved by SDP solvers. Similar to problem $\mathcal{Q}_r$, this SDP feasibility problem has G matrix variables of size $N \times N$, and $M + 1$ linear constraints. So computing an ϵ-feasible solution by an interior point method will have an overall

iteration count of $O(\sqrt{GN}\log(1/\epsilon))$, while each iteration has a complexity of $O(G^3N^6 + MGN^2)$. The use of SeDuMi in the algorithm is convenient, because it does not only yield a solution to problem $\mathcal{FP}$ when the latter is feasible, but it also provides a certificate of infeasibility otherwise. As with problem $\mathcal{Q}_r$, actual runtime complexity will usually scale far slower with G, N, M than this worst-case bound.

When the algorithm terminates, the solution vector $\mathbf{v}$, obtained by the last feasible iteration, contains the approximate solution to the relaxed problem $\mathcal{F}_r$, namely the blocks $\{\mathbf{X}_k\}_{k=1}^G$. The corresponding (approximate) optimal value of problem $\mathcal{F}_r$ is an upper bound on the guaranteed received SINR by all users, that can be achieved with total transmit power P. This bound can only be met in the case when all blocks $\mathbf{X}_k$ are rank-one, so that their principal components can be chosen as optimum beamforming vectors $\mathbf{w}_k$. Due to the relaxation of the rank constraints, this is generally not true. Thus, post-processing of the relaxed solution is needed when the solution matrices $\{\mathbf{X}_k\}_{k=1}^G$ are not all rank-one, so as to yield an approximate solution to the original joint max-min fair problem $\mathcal{F}$. This can be accomplished by using a combined randomization - joint power control procedure, similar to the one described in Section IV. Specifically, Gaussian randomization (e.g., see [11]) may be used in a first step to create candidate sets of beamforming vectors $\{\mathbf{w}_k\}_{k=1}^G$ in the span of the respective transmit covariance matrices. In a second step, the available transmit power P is allocated to the candidate beamforming vectors, by adjusting the power boost (or back-off) factors p_k for each multicast group. The set of $(p_k, \mathbf{w}_k)$ pairs which maximizes the minimum received SINR is then chosen among all feasible solutions generated this way. Given a candidate set of beamforming vectors, the transmit power can be optimally allocated by solving the following problem

$$\mathcal{MGPC}' :$$

$$\max_{\{p_k\in\mathbb{R}\}_{k=1}^G} \min_{k\in\{1,\ldots,G\}} \min_{i\in\mathcal{G}_k} \frac{p_k\alpha_{k,i}}{\sum_{\ell\neq k} p_\ell\alpha_{\ell,i} + \sigma_i^2}$$

$$\textit{s.t.}: \quad \sum_{k=1}^G \beta_k p_k = P,$$

$$p_k \geq 0, \ \forall k \in \{1,\ldots,G\},$$

where, as introduced in Section IV, $\beta_k = \|\mathbf{w}_k\|_2^2$ and $\alpha_{k,i}$ denotes the signal power received by user i from the stream directed towards users in multicast group k. Introducing a real positive auxiliary

variable γ, we can recast problem $\mathcal{MGPC}'$ as

$$\max_{\{p_k \in \mathbb{R}\}_{k=1}^{G},\ \gamma \in \mathbb{R}} \quad \gamma$$

$$s.t.\ : \frac{p_k \alpha_{k,i}}{\sum_{\ell \neq k} p_\ell \alpha_{\ell,i} + \sigma_i^2} \geq \gamma,$$

$$\forall k \in \{1, \ldots, G\},\ \forall i \in \mathcal{G}_k,$$

$$\sum_{k=1}^{G} \beta_k p_k = P,$$

$$p_k \geq 0,\ \forall k \in \{1, \ldots, G\}, \quad \text{and} \quad \gamma \geq 0.$$

The bisection algorithm, described earlier in this section, can be used again to obtain a solution to problem $\mathcal{MGPC}'$. The search interval is bounded below by $L = 0$, as before. However, we may now further restrict the upper bound U to the optimal value obtained for the relaxed problem $\mathcal{F}_r$. The convex feasibility problem, which is to be solved in each iteration for a given choice of the positive real γ, is

$\mathcal{FP}'$:

find $\mathbf{v}'$

s.t. : $\alpha_{k,i} p_k - \gamma \sum_{\ell \neq k} \alpha_{\ell,i} p_\ell - s_i = \gamma \sigma_i^2,$

$\forall k \in \{1, \ldots, G\},\ \forall i \in \mathcal{G}_k,$

$\sum_{k=1}^{G} \beta_k p_k = P,$

$p_k \geq 0,\ \forall k \in \{1, \ldots, G\},$

$s_i \geq 0,\ \forall i \in \{1, \ldots, M\},$

where $\mathbf{v}' \in \mathbb{R}^{M+G}$ denotes the variable vector $\mathbf{v}' = [\mathbf{s}^T\ \mathbf{p}^T]^T$. Problem $\mathcal{FP}'$ is a linear feasibility problem with G nonnegative variables and $M + 1$ linear inequality constraints. An interior point method can generate either an ϵ-feasible solution in $O(\sqrt{G}\log(1/\epsilon))$ iterations, each requiring at most $O(G^3 + MG)$ arithmetic operations, or return a dual certificate showing the problem is infeasible. When the bisection algorithm terminates, the solution vector $\mathbf{v}'$ obtained in the last feasible iteration contains the boost / attenuation factors which optimally allocate the available transmit power among the G multicast groups, for the given set of candidate beamforming vectors. If this set of $(p_k, \mathbf{w}_k)$ pairs

yields larger worst-case received SINR than previously checked sets, then it is recorded; otherwise it is discarded.

Using the algorithm described so far, the cost of finding an approximate solution to the joint max-min fair beamforming problem is that of solving N_{iter} SDP and $N_{\text{rand}}N'_{\text{iter}}$ LP feasibility problems, where N'_{iter} are the iterations of the bisection executed for the solution of the $\mathcal{MGPC}'$ problem.

The quality of the final approximate solution can be measured by the ratio of the optimal value of problem $\mathcal{F}_r$ (which, as mentioned already, is actually an upper bound) to the maximum attained optimal value of problem $\mathcal{MGPC}'$.

VI. Numerical results

A. QoS Approach

In Sections III and IV we have derived a two-step algorithm to yield, in polynomial time, an approximate solution to the joint QoS multicast beamforming problem $\mathcal{Q}$. The first step of the proposed algorithm consists of a relaxation of the original problem $\mathcal{Q}$ to problem $\mathcal{Q}_r$. The original problem $\mathcal{Q}$ may or may not be feasible; if it is, then so is problem $\mathcal{Q}_r$. If $\mathcal{Q}_r$ is infeasible, then so is $\mathcal{Q}$. The converse is generally not true; i.e., if $\mathcal{Q}_r$ is feasible, $\mathcal{Q}$ need not be feasible. In order to establish feasibility of $\mathcal{Q}$ in this case, the randomization - $\mathcal{MGPC}$ loop should yield at least one feasible solution. This is most often the case, as will be verified in the sequel. If the randomization - $\mathcal{MGPC}$ loop fails to return at least one feasible solution, then the (in)feasibility of $\mathcal{Q}$ cannot be determined. There is, therefore, a relatively small proportion of problem instances for which (in)feasibility of $\mathcal{Q}$ cannot be decided using the proposed approach. It is evident from the above discussion that feasibility is a key aspect of problem $\mathcal{Q}$ and its proposed solution via problem $\mathcal{Q}_r$ and the randomization - $\mathcal{MGPC}$ loop. Feasibility depends on a number of factors; namely, the number of transmit antenna elements N, the number and the populations of the multicast groups, G and G_k respectively, the channel characteristics $\mathbf{h}_i$, the channel noise variances σ_i^2, and finally the desired receive SINR constraints γ_i.

Beyond feasibility, there are two key issues of interest. The first has to do with cases for which the solution to the relaxed problem $\mathcal{Q}_r$ yields an exact optimum of the original problem $\mathcal{Q}$. This happens when the $N \times N$ solution blocks $\mathbf{X}_k$, $k \in \{1, \cdots, G\}$, turn out all being rank-one. In this case, the associated principal components solve optimally the original problem $\mathcal{Q}$, i.e., in such a case $\mathcal{Q}_r$ is not a relaxation after all. It is interesting to find the frequency of occurrence of such an event, whose benefit is twofold: the problem is solved not only optimally, but also at a smaller complexity, since the randomization step and the repeated solution of the ensuing $\mathcal{MGPC}$ problem is avoided. The second issue of interest has to do with the quality of the final approximate solution to problem $\mathcal{Q}$, in those cases where a feasible solution can be found using the proposed two-step algorithm. As in [11], a

practical figure of merit for the quality of the final approximate solution (set of beamforming vectors and power scaling factors) is the ratio of the total transmitted power corresponding to the approximate solution over $\sum_{k=1}^{G} \mathrm{trace}(\mathbf{X}_k)$ - the lower bound generated from the solution of the relaxed problem $\mathcal{Q}_r$.

We first consider the standard i.i.d. Rayleigh fading model, i.e., the elements of the $N \times 1$ channel vectors $\mathbf{h}_i,\ \forall i \in \{1, \ldots, M\}$ are i.i.d. circularly symmetric complex Gaussian random variables of variance 1. For each scenario considered, 300 different channel snapshots are randomly created according to the aforementioned model and fed to the proposed algorithm. The results presented in this subsection are obtained by averaging over 300 Monte-Carlo runs, using 300 Gaussian randomization samples in each run. Tables I, II, and III summarize these results, for N (number of transmit antenna elements) set to 4, 6, and 8 respectively. The proposed algorithm is tested for a variety of choices for M (the total number of single-antenna receivers) and G (the number of multicast groups), which index the rows in the tables (columns 1 and 2, respectively). The users are considered to be evenly distributed among the multicast groups, i.e., $G_k = M/G,\ \forall k \in \{1, \ldots, G\}$. For each such configuration, the QoS downlink beamforming problem is solved for increasing values (in the 6-20 dB range, see column 3) of the received SINR constraints (same for all users), provided that there exist channel instances for which problem $\mathcal{Q}_r$ is feasible. The noise variance is set to $\sigma^2 = 1$ for all channels. The percentage of the 300 Monte-Carlo runs for which $\mathcal{Q}_r$ is feasible is shown in column 4. Columns 5 reports the percentage of feasible solutions to problem $\mathcal{Q}_r$, which yield exact solutions to problem $\mathcal{Q}$. This is calculated as the percentage of problem instances for which all $\mathbf{X}_k$ in the solution of $\mathcal{Q}_r$ turn out having rank (essentially) equal to one (defined by the second largest eigenvalue being smaller than 10^{-3} of the sum of all eigenvalues). Column 6 reports the percentage of problem instances for which, once a feasible solution to problem $\mathcal{Q}_r$ is found, the proposed algorithm's second step, i.e., the ensuing randomization - $\mathcal{MGPC}$ loop, yields at least one feasible solution for the original problem $\mathcal{Q}$. The next two columns (7 and 8), hold the mean and standard deviation of the quality measure, defined in Section IV as the ratio of transmitted power corresponding to the final approximate solution over the lower bound obtained from the SDR solution. This ratio equals one when rank relaxation is exact (not a relaxation after all), and the reported statistics depend on the frequency (see column 5) of this event. In order to obtain additional insight in the quality of the approximation step, conditional statistics are also reported in the last two columns (9 and 10) after excluding exact optimum solutions from the calculation.

An initial comment, regarding the feasibility of the relaxed problem $\mathcal{Q}_r$, is that in all configurations considered, the higher the target SINR, the less likely it is that $\mathcal{Q}_r$ is feasible, which is intuitive. Furthermore, $\mathcal{Q}_r$ is getting more difficult to solve as the number G of multicast groups increases and/or

as more (randomly generated) users per multicast group are added, since in either case interference is higher. Finally, it is seen that increasing the number of transmit antennas (N) improves service, as expected: higher receive SINR can be attained by more users in more multicast groups.

The most interesting observation, concerning the percentage of problems $\mathcal{Q}_r$ for which the relaxation is tight, is that it increases as the number of users per multicast group decreases; percentages are significant especially when the number of users per group is smaller or equal to the number of transmit antennas. This can be seen in two ways: either by holding the number of groups fixed while decreasing their populations, or by fixing the total number of users and distributing them in more multicast groups. Trying to interpret this fact, note that in both cases the problem is pushed towards the multiuser (independent information) downlink problem, where each user forms a multicast group. The latter is known to be convex, and the associated SDP relaxation has been shown to be tight [1]. In addition, the $\mathcal{Q}_r$ optimality percentage also increases with target SINR. It seems as if rank-one solutions are more likely when operating close to the infeasibility boundary. In some scenarios, $\mathcal{Q}_r$ *consistently* yields an exact solution of $\mathcal{Q}$. That is, the $\mathbf{X}_k$ blocks are all consistently rank-one. In this case, no further randomization is needed - the principal components of the extracted blocks are the optimal beamformers. More on this can be found in [7].

As far as the approximation step of the proposed algorithm is concerned, we can distinguish two cases. In most of the scenarios considered, the number of users per multicast group was kept smaller or equal to the number of transmit antenna elements, so that a realistic value of the receive SINR could be guaranteed, for a significant fraction of the different channel instances. There, the randomization - $\mathcal{MGPC}$ loop yields a feasible solution with a probability higher than 90% in most cases where $\mathcal{Q}_r$ is feasible; this solution entails transmission power that is under two times (3 dB from) the possibly unattainable lower bound, on average. The actual numbers for each configuration depend on the number of the Gaussian randomization samples; 300 have proved adequate for most configurations. However, when a relatively low target SINR is to be guaranteed to a number of users per group larger than the number of antennas, the feasibility of the approximation decreases and the power penalty increases. This can be appreciated by looking at the lowest sub-matrices of Tables I–III. Simulations are repeated for these configurations using 1000 Gaussian random samples. The results are summarized in Table IV, where an extra column has been added at the front, indicating the number N of transmit antenna elements. A small improvement is observed in the quality of the approximation; but it is still inadequate for the last configuration.

B. Max-min-fair Approach

In this subsection we assess the performance of the algorithm derived in Section V for the JMMF downlink multicast beamforming problem. As in the previous subsection, the standard i.i.d. Rayleigh fading model is used for Monte-Carlo simulations. Table V summarizes the results obtained using the proposed algorithm for 300 Monte-Carlo runs and 1000 Gaussian randomization samples each. The value of the available transmit power P is set to 1000 for all the reported simulation results. Note at this point that, contrary to the single-group multicasting scenario [11], the optimization problem in the general case of multiple multicast groups is interference-limited; hence, it depends on the value of P. Specifically, if the same problem is solved for two different values of P, the designed beams will have the same shape, but the power allocation, i.e. the solution of the $\mathcal{MGPC}'$ problem will differ.

Simulations are conducted for three different choices (4, 6, and 8) of the number N of transmit antenna elements and a variety of choices for the number of receiving single-antenna mobile users M, shown respectively in the first and the second column of the table. The users are considered to be evenly distributed among the G multicast groups; their number is stored in the third column. The fourth column reports the percentage of the Monte-Carlo runs for which all solution blocks $\mathbf{X}_k$ of the relaxed problem $\mathcal{F}_r$ are essentially rank-one. As mentioned already, when this is the case, the principal components of the blocks optimally solve the original joint max-min-fair problem, i.e. problem $\mathcal{F}_r$ is equivalent to and not a relaxation of problem $\mathcal{F}$; hence, there is no need for the algorithm's second step (randomization - $\mathcal{MGPC}'$ loop). It is evident that this case occurs quite frequently, with a frequency which drops as the number of users and the number of multicast groups increases.

The next two columns (fifth and sixth) of Table V hold the average value (over all Monte-Carlo runs) and the standard deviation, respectively, of the ratio of the optimal value of problem $\mathcal{F}_r$ to the maximum attained optimal value of problem $\mathcal{MGPC}'$. This is a measure of the quality of the overall solution obtained using our proposed approach. The final two columns (seventh and eighth) report the same statistics, but only for the Monte-Carlo runs for which the relaxation is not essentially tight, for additional insight on the quality of the approximation step. It is observed that the minimum achieved SINR is usually very close (in the mean) to the upper bound calculated by the relaxed SDP problem $\mathcal{F}$; thus, the approximation step yields high quality solutions. Compared with the respective results for the single multicast group case [11], the multi-group algorithm consistently performs better. In addition, the quality of the approximation becomes better (i.e., the mean of the ratio drops) as a given number of users is distributed among a larger number of multicast groups (e.g., see the case of 12 users, divided into 2, 3, and 4 groups). The interpretation given for the QoS formulation, that the

problem is pushed towards the (convex) multiuser downlink problem, applies here, too.

Regarding practical execution time, the SDP feasibility problem $\mathcal{F}_r$ is solved in about 0.1 sec, on a typical desktop PC, for the cases considered. The variation of this execution time is almost negligible for the tested variation of the values M and G (users and groups, respectively). However, an approximately linear dependence of the execution time on the number of transmit antennas N has been observed. The LP feasibility problem is solved in approximately 0.05 sec, irrespective of the scenario considered. Thus, in practice the algorithm needs approximately $1 + 0.5N_{\text{rand}}$ sec (for $N_{\text{iter}} \approx N'_{\text{iter}} \approx 10$), when the relaxation is not tight.

C. Experiments with Measured Channel Data

The performance of the proposed multicast beamforming algorithms was also tested on measured channel data courtesy of iCORE HCDC Lab, University of Alberta in Edmonton, Canada. Measurements were carried out using a portable 4×4 multiple-input multiple-output (MIMO) testbed that operates in the 902–928 MHz (ISM) band. The transmitter (Tx) and the receiver (Rx) were equipped with antenna arrays, each comprising four vertically polarized dipole antennas spaced $\lambda/2$ (≈ 16 cm) apart. The chip rate used for sounding was low enough to safely assume that the channel is not frequency selective. More details on the testbed configuration and the procedure used to estimate the channel gains of the MIMO channel matrix can be found in [5]. Datasets and a detailed description of many measurement campaigns in typical propagation environments are available at the iCORE HCDC Lab website (http://www.ece.ualberta.ca/∼mimo/). The most pertinent scenario for our purposes is the stationary outdoor one, called Quad and illustrated in Figure 2. Quad is a 150 by 60 meters lawn surrounded by buildings with heights from approximately 15 to 30 meters. The Tx location was fixed, whereas the Rx was placed in 6 different locations (no measurements are actually provided for location 4) as indicated in Figure 2. For every Rx location, 9 different measurements were taken by shifting the Rx antenna array on a 3×3 square grid with $\lambda/4$ spacing. Each measurement contains about 100 4×4 channel snapshots, recorded 3 per second; thus for each location there are about 900 MIMO channel gain matrices available. We form multicast groups by considering each receive antenna at each location as a separate terminal, and grouping terminals in 1–3 locations into one multicast group. The results reported in Tables VI–XII and XIII, for the QoS and the JMMF problem formulations, respectively, are obtained by averaging over the 900 channel instances. Channel gains are normalized before use, dividing by the average channel amplitude for the respective configuration. 300 Gaussian samples are employed in the randomization / MGPC loop. The main findings regarding performance of our algorithms applied to the measured channel data can be summarized as follows:

• For 2 multicast groups and number of users per group equal to the number of Tx antennas ($N = 4$),

the relaxation $\mathcal{Q} \rightarrow \mathcal{Q}_r$ is tight very frequently (70–100%) and the power penalty paid by the approximation step very small. These hold irrespective of the distribution of each group's users in 1, 2, or even 3 locations (see Tables VI, VII, and VIII, respectively).

• For 2 multicast groups of 6 (or 8) users each, evenly distributed in 2 locations, the relaxation $\mathcal{Q} \rightarrow \mathcal{Q}_r$ is tight for more than half of the occasions (see Tables IX, and XI). There exist channel instances for which SINR up to 14 (or 12) dB can be guaranteed; such high SINR values are unattainable under the corresponding i.i.d. Rayleigh fading scenario. The quality of approximation is good, even though the number of user per group is larger than the number of transmit antenna elements.

• When 6 users in each of 2 multicast groups are evenly distributed in 3 locations, the relaxation $\mathcal{Q} \rightarrow \mathcal{Q}_r$ is tight less frequently $(< 80\%)$, and the problem is feasible only up to about 10 dB (see Table X). The feasibility of the approximation step can drop $< 80\%$.

• For 3 multicast groups (see Table XII) of 3 co-located users each, the relaxation $\mathcal{Q} \rightarrow \mathcal{Q}_r$ is almost always tight $(> 90\%)$ and feasible up to 10 dB of prescribed SINR. For 4 users per group it becomes infeasible for SINR values larger than about 8 dB.

• When the number of users per multicast group is small, the relaxation $\mathcal{F} \rightarrow \mathcal{F}_r$ in the JMMF formulation is tight in a high percentage of cases (see Table XIII). This percentage drops as the number of users per multicast group increases. In all scenarios considered, the proposed algorithm yields high quality approximate solutions.

VII. Conclusions

The downlink beamforming problem was considered for the general case of multiple co-channel multicast groups, under two design criteria: QoS, in which we seek to minimize the total transmitted power while guaranteeing a prescribed minimum SINR at all receivers; and a fair objective, in which we seek to maximize the minimum received SINR under a total power constraint. Both formulations contain single group multicast beamforming as a special case, and are therefore NP-hard. Computationally efficient quasi-optimal solutions were proposed by means of SDR and a combined randomization - MGPC loop. Extensive numerical results have been presented, using both simulated (i.i.d. Rayleigh) and measured stationary outdoor wireless channel data, showing that the proposed algorithms yield high quality approximate solutions at a moderate complexity cost. Interestingly, our numerical findings indicate that the solutions generated by our algorithms are often exactly optimal, especially in the case of measured channels. In certain cases this optimality can be proven beforehand, and alternative convex reformulations of lower complexity can be constructed; see [7] for further details. In other cases, a theoretical worst-case bound on approximation accuracy can be derived, and shown to be tight; on this issue, see [8].

REFERENCES

[1] M. Bengtsson, and B. Ottersten, "Optimal and Suboptimal Transmit Beamforming", ch. 18 in *Handbook of Antennas in Wireless Communications*, L. C. Godara, Ed., CRC Press, Aug. 2001.

[2] S. Boyd, and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004; see also http://www.stanford.edu/∼boyd/cvxbook.html

[3] F.-R. Farrokhi, K.J.R. Liu, and L. Tassiulas, "Downlink Power Control and Base Station Assignment", *IEEE Communications Letters*, vol. 1, no. 4, pp. 102–104, July 1997.

[4] M.R. Garey, and D.S. Johnson, *Computers and Intractability. A Guide to the Theory of NP-Completeness*, W.H. Freeman and Company, 1979.

[5] P. Goud, and R. Hang, D. Truhachev, and C. Schlegel, "A Portable MIMO Testbed and Selected Channel Measurements", *EURASIP JASP*, to appear.

[6] E. Karipidis, N.D. Sidiropoulos, Z.-Q. Luo, "Transmit Beamforming to Multiple Co-channel Multicast Groups", in *Proc. IEEE CAMSAP 2005*, Dec. 12-14, Puerto Vallarta, Mexico.

[7] E. Karipidis, N.D. Sidiropoulos, Z.-Q. Luo, "Convex Transmit Beamforming for Downlink Multicasting to Multiple Co-channel Groups", in *Proc. IEEE ICASSP 2006*, May 14-19, Toulouse, France.

[8] Z.-Q. Luo, N. D. Sidiropoulos, P. Tseng, and S. Zhang, "Approximation Bounds for Quadratic Optimization with Homogeneous Quadratic Constraints", *SIAM J. Optim.*, accepted subject to minor revision.

[9] M. J. Lopez, "Multiplexing, scheduling, and multicasting strategies for antenna arrays in wireless networks", Ph.D. thesis, Dept. of Elect. Eng. and Comp. Sci., MIT, Cambridge, MA, 2002.

[10] N.D. Sidiropoulos, T.N. Davidson, "Broadcasting with Channel State Information", in *Proc. IEEE SAM 2004 Workshop*, vol. 1, pp. 489-493, Sitges, Barcelona, Spain, July 1821, 2004.

[11] N.D. Sidiropoulos, T.N. Davidson, and Z.-Q. Luo, "Transmit Beamforming for Physical Layer Multicasting", *IEEE Trans. on Signal Processing*, vol. 54, no. 6, pp. 2239-2251, June 2006.

[12] J.F. Sturm, "Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones", *Optimization Methods and Software*, vol. 11-12, pp. 625–653, 1999; see also http:///sedumi.mcmaster.ca

[13] H. Wolkowicz, "Relaxations of Q2P", ch. 13.4 in *Handbook of Semidefinite Programming: Theory, Algorithms, and Applications*, H. Wolkowicz, R. Saigal, L. Vandenberghe, Eds., Kluwer Academic Publishers, 2000.

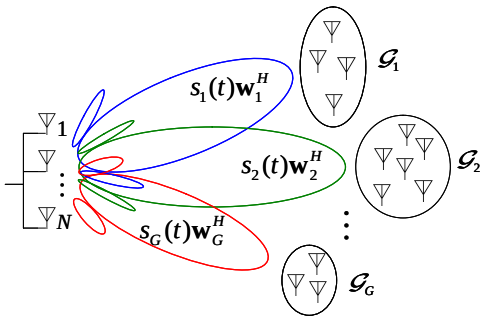


Fig. 1. Co-channel multicast beamforming concept (note that groups need not be spatially clustered).

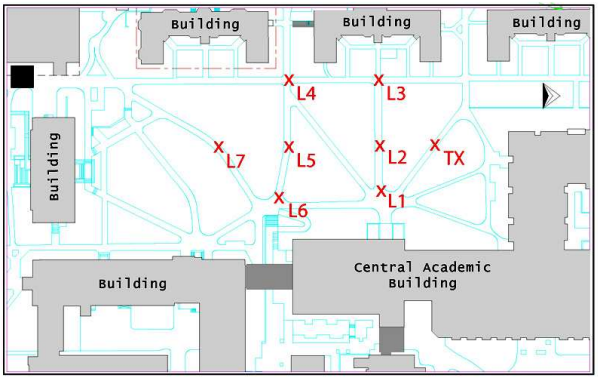


Fig. 2. Sample wireless channel measurement scenario from http://www.ece.ualberta.ca/∼mimo/

TABLE I

MC SIMULATION RESULTS (RAYLEIGH); QOS TX BEAMFORMING; $N = 4$ TX ANTENNAS, 300 RANDOMIZATIONS

M	G	SINR	feas. $\mathcal{Q}_r$	opt. $\mathcal{Q}_r$	feas. appr.	all solutions mean	all solutions std	appr. solutions mean	appr. solutions std
6	3	6	89.67	99.63	100	1.0000	0.0005	1.0079	0
6	3	8	70.33	100	-	1	0	-	-
6	3	10	45.33	100	-	1	0	-	-
6	3	12	27	100	-	1	0	-	-
6	3	14	14	100	-	1	0	-	-
6	3	16	7	100	-	1	0	-	-
8	2	6	98.33	79.66	98.31	1.0550	0.1710	1.2902	0.2950
8	2	8	90.67	83.46	98.90	1.0838	0.3788	1.5366	0.8301
8	2	10	73.33	83.18	98.18	1.1935	1.8118	2.2668	4.5446
8	2	12	52	85.90	98.72	1.2018	2.1247	2.5542	5.8430
8	2	14	32	88.54	100	1.0128	0.0593	1.1113	0.1462
8	2	16	16.33	89.80	95.92	1.0426	0.1892	1.6679	0.4433
8	2	18	9.33	92.86	100	1.0154	0.0669	1.2162	0.1847
8	2	20	3	88.89	100	1.0543	0.1628	1.4884	0
9	3	6	5.67	100	-	1	0	-	-
12	2	6	42	49.21	79.37	1.6927	1.8918	2.8228	2.7314
12	2	8	10.33	80.65	93.55	1.1921	0.5123	2.3929	0.4689
12	2	10	1.33	100	-	1	0	-	-
16	2	6	6.33	26.32	68.42	1.5619	1.2120	1.9131	1.4669

TABLE II

MC SIMULATION RESULTS (RAYLEIGH); QOS TX BEAMFORMING; $N = 6$ TX ANTENNAS, 300 RANDOMIZATIONS

			feas.	opt.	feas.	all solutions		appr. solutions	
M	G	SINR	$\mathcal{Q}_r$	$\mathcal{Q}_r$	appr.	mean	std	mean	std
8	2	6	100	80.67	99	1.0228	0.0734	1.1233	0.1301
8	2	8	100	82.33	98.67	1.0162	0.0514	1.0979	0.0900
8	2	10	100	87.67	97.67	1.0118	0.0485	1.1150	0.1067
8	2	12	100	88	98	1.0102	0.0396	1.1004	0.0803
8	2	14	100	89.33	98.67	1.0099	0.0478	1.1050	0.1211
8	2	16	100	90.33	98	1.0089	0.0483	1.1143	0.1359
8	2	18	100	92.67	99.33	1.0071	0.0409	1.1052	0.1235
8	2	20	100	92	99.33	1.0064	0.0328	1.0862	0.0894
12	2	6	100	35.33	94	1.2782	0.5915	1.4458	0.6977
12	2	8	100	39	95	1.505	3.3075	1.8661	4.2771
12	2	10	97	50.52	92.44	1.2513	0.5905	1.5542	0.7766
12	2	12	86.67	56.92	94.23	1.2172	0.5583	1.5487	0.7800
12	2	14	68.67	63.59	94.66	1.2330	0.8614	1.7098	1.3993
12	2	16	47	69.50	92.91	1.2031	1.3485	1.8064	2.6241
12	2	18	27.33	69.51	95.12	1.1972	0.9655	1.7324	1.7825
12	2	20	17	82.35	100	1.0734	0.2537	1.4157	0.4921
12	3	6	72	76.85	93.06	1.2440	1.1504	2.4010	2.4730
12	3	8	19.67	83.05	94.92	1.0433	0.2193	1.3460	0.5645
12	3	10	2.33	100	-	1	0	-	-
12	4	6	4	100	-	1	0	-	-
16	2	6	98.33	11.19	74.58	3.1376	4.9208	3.5149	5.2495
16	2	8	75	15.11	63.56	2.4204	2.2800	2.8635	2.4499
16	2	10	26.67	31.25	58.75	1.5876	1.0146	2.2554	1.1734
16	2	12	4.33	38.46	84.62	4.5305	7.4475	7.4725	9.3851

TABLE III

MC SIMULATION RESULTS (RAYLEIGH); QOS TX BEAMFORMING; $N = 8$ TX ANTENNAS, 300 RANDOMIZATIONS

			feas.	opt.	feas.	all solutions		appr. solutions	
M	G	SINR	$\mathcal{Q}_r$	$\mathcal{Q}_r$	appr.	mean	std	mean	std
12	2	6	100	37	95.33	1.1814	0.2527	1.2964	0.2651
12	2	8	100	35.67	96.33	1.1733	0.2409	1.2752	0.2532
12	2	10	100	34.67	95	1.1734	0.2329	1.2730	0.2413
12	2	12	100	41.33	96	1.1485	0.2099	1.2607	0.2194
12	2	14	100	43	95	1.1478	0.2157	1.2700	0.2281
12	2	16	100	45	94.33	1.1316	0.1956	1.2516	0.2074
12	2	18	100	48.33	95.67	1.1226	0.2297	1.2477	0.2754
12	2	20	100	53.33	95.33	1.0993	0.1765	1.2253	0.2059
12	3	6	100	78.67	98.33	1.0372	0.1065	1.1862	0.1711
12	3	8	100	79	98	1.0367	0.1081	1.1892	0.1783
12	3	10	98.67	81.42	98.99	1.0452	0.1406	1.2545	0.2425
12	3	12	94.67	85.21	97.54	1.0393	0.1464	1.3112	0.2945
12	3	14	78.67	88.14	98.73	1.0559	0.2863	1.5207	0.7351
12	3	16	52.33	92.99	99.36	1.0241	0.1114	1.3766	0.2571
12	3	18	30.67	93.48	98.91	1.0291	0.1444	1.5303	0.3705
12	3	20	17.67	98.11	100	1.0054	0.0396	1.2881	0
12	4	6	100	93.33	99.67	1.0072	0.0342	1.1130	0.0822
12	4	8	87.33	98.09	99.62	1.0037	0.0353	1.2439	0.1731
12	4	10	42.33	97.64	100	1.0075	0.0635	1.3166	0.3272
12	4	12	12	97.22	100	1.0099	0.0595	1.3568	0
12	4	14	3.33	100	-	1	0	-	-
16	2	6	100	9.67	93	1.8833	1.6259	1.9858	1.6882
16	2	8	100	11.67	91	1.9955	2.2743	2.1419	2.4018
16	2	10	99.67	15.05	86.62	1.8757	1.3169	2.0598	1.3800
16	2	12	98.67	22.64	88.18	1.6957	1.5693	1.9359	1.7582
16	2	14	94.67	31.69	88.38	1.7937	2.2986	2.2374	2.7755
16	2	16	72.67	46.33	92.20	1.7075	3.8635	2.4220	5.3971
16	2	18	54	59.26	93.21	1.3305	1.0350	1.9073	1.5628
16	2	20	33.33	65	94	1.2662	0.8232	1.8630	1.3105
24	2	6	99.33	0.34	43.96	6.7920	8.7394	6.8366	8.7582
24	2	8	60.67	4.40	30.22	4.8706	6.2300	5.5294	6.5203
24	2	10	11.67	14.29	34.29	3.6415	5.1995	5.5284	6.2925

TABLE IV

MC SIMULATION RESULTS (RAYLEIGH); QOS TX BEAMFORMING; 1000 RANDOMIZATIONS

				feas.	opt.	feas.	all solutions		appr. solutions	
N	M	G	SINR	$\mathcal{Q}_r$	$\mathcal{Q}_r$	appr.	mean	std	mean	std
4	12	2	6	42	49.21	85.71	1.7778	3.0955	2.8261	4.5637
4	12	2	8	10.33	80.65	96.77	1.1785	0.4303	2.0710	0.3838
4	16	2	6	6.33	26.32	73.68	1.5490	1.1660	1.8540	1.3843
6	16	2	6	98.33	11.19	85.08	3.5662	9.0980	3.9547	9.7061
6	16	2	8	75	15.11	68.89	2.6867	3.6744	3.1607	4.0366
6	16	2	10	26.67	31.25	65	1.6679	1.5570	2.2863	1.9823
6	16	2	12	4.33	38.46	92.31	2.9157	4.1172	4.2840	5.0828
8	16	2	6	100	9.67	96.67	1.8918	3.8850	1.9909	4.0839
8	16	2	8	100	11.67	95.33	1.8077	1.5349	1.9203	1.6067
8	16	2	10	99.67	15.05	93.31	1.7600	1.9039	1.9061	2.0475
8	16	2	12	98.67	22.64	92.23	1.6557	1.7557	1.8689	1.9758
8	16	2	14	94.67	31.69	94.01	1.6887	2.7756	2.0389	3.3582
8	16	2	16	72.67	46.33	95.87	1.3992	1.0721	1.7725	1.3939
8	16	2	18	54	59.26	96.30	1.3021	0.8748	1.7856	1.2744
8	16	2	20	33.33	65	94	1.2416	0.7279	1.7832	1.1492
8	24	2	6	99.33	0.34	52.01	5.8627	7.2044	5.8942	7.2142
8	24	2	8	60.67	4.40	35.71	5.7322	10.6103	6.3963	11.1810
8	24	2	10	11.67	14.29	37.14	2.6311	3.7697	3.6505	4.6122

TABLE V

MC SIMULATION RESULTS (RAYLEIGH); JMMF TX BEAMFORMING; $P = 1000$, 1000 RANDOMIZATIONS

			opt.	all solutions		appr. solutions	
N	M	G	$\mathcal{F}_r$	mean	std	mean	std
4	8	2	75.33	1.011	0.047	1.043	0.087
4	12	2	28.00	1.086	0.121	1.119	0.129
4	16	2	5.67	1.214	0.192	1.227	0.190
4	24	2	0	1.528	0.311	1.528	0.311
4	12	3	11.67	1.053	0.074	1.060	0.076
4	18	3	0	1.200	0.145	1.200	0.145
4	12	4	16.00	1.022	0.037	1.026	0.039
4	16	4	3.00	1.082	0.081	1.084	0.081
6	12	2	61.33	1.046	0.104	1.119	0.139
6	16	2	17.33	1.146	0.162	1.176	0.163
6	24	2	0.67	1.557	0.324	1.561	0.322
6	12	3	42.67	1.022	0.051	1.039	0.063
6	18	3	5.00	1.178	0.150	1.187	0.149
6	12	4	36.33	1.009	0.026	1.013	0.032
6	16	4	9.00	1.053	0.071	1.058	0.072
8	12	2	29.00	1.066	0.116	1.093	0.128
8	16	2	50.67	1.066	0.133	1.133	0.164
8	24	2	2.33	1.490	0.332	1.502	0.327
8	32	2	0	1.996	0.447	1.996	0.447
8	12	3	78.33	1.007	0.026	1.033	0.048
8	18	3	17.67	1.098	0.126	1.118	0.130
8	12	4	66.00	1.001	0.011	1.005	0.019
8	16	4	25.33	1.025	0.054	1.037	0.060

TABLE VI

2 MULTICAST GROUPS; 4 USERS PER GROUP IN 1 LOCATION

SINR	feas. $\mathcal{Q}_r$	opt. $\mathcal{Q}_r$	feas. appr.	all solutions mean	all solutions std	appr. solutions mean	appr. solutions std
Group 1 (4 at L1) & Group 2 (4 at L2)							
6–18	100	100	-	1	0	-	-
20	99.89	100	-	1	0	-	-
22	97.37	100	-	1	0	-	-
24	84.82	100	-	1	0	-	-
26	65.53	100	-	1	0	-	-
28	45.21	100	-	1	0	-	-
30	24.43	100	-	1	0	-	-
Group 1 (4 at L1) & Group 2 (4 at L3)							
6	100	99.89	100	1.0001	0.0035	1.1026	0
8	100	99.89	100	1.0000	0.0014	1.0421	0
10	100	99.77	100	1.0001	0.0014	1.0247	0.0216
12	100	99.89	100	1.0001	0.0018	1.0536	0
14	100	99.89	100	1.0001	0.0030	1.0876	0
16	100	99.77	100	1.0002	0.0047	1.0805	0.0777
18	93.16	97.92	99.88	1.0068	0.0761	1.3493	0.4321
20	82.10	98.61	100	1.0024	0.0246	1.1715	0.1262
22	72.98	99.69	100	1.0040	0.0934	2.2920	1.4971
24	60.78	99.44	100	1.0102	0.2047	2.8121	2.4993
26	35.01	99.02	100	1.0006	0.0086	1.0623	0.0752
28	18.02	100	-	1	0	-	-
30	9.46	100	-	1	0	-	-
Group 1 (4 at L5) & Group 2 (4 at L2)							
6	100	98.63	100	1.0009	0.0091	1.0636	0.0470
8	100	98.63	99.89	1.0011	0.0115	1.0884	0.0549
10	99.77	96.22	99.66	1.0083	0.0691	1.2398	0.2927
12	96.80	93.28	98.94	1.0217	0.1365	1.3789	0.4407
14	85.27	92.91	98.66	1.0364	0.4837	1.6236	1.9302
16	64.61	96.64	99.65	1.0105	0.0824	1.3491	0.3363
18	40.87	97.77	99.44	1.0029	0.0249	1.1709	0.0972
20	23.74	99.52	100	1.0002	0.0030	1.0433	0
22	10.05	97.73	100	1.0017	0.0148	1.0731	0.0930
24	4.22	100	-	1	0	-	-
Group 1 (4 at L5) & Group 2 (4 at L7)							
6	97.72	82.15	97.78	1.0475	0.1900	1.2968	0.3908
8	91.11	83.48	87.87	1.0486	0.2323	1.3306	0.5251
10	81.64	87.29	98.05	1.0436	0.2147	1.3978	0.5314
12	49.49	91.94	98.85	1.0427	0.2919	1.6110	0.9480
14	18.93	90.97	98.19	1.0172	0.0782	1.2339	0.1868
16	8.78	93.51	100	1.0101	0.0637	1.1551	0.2212
18	2.85	92	96	1.0003	0.0017	1.0081	0

TABLE VII

2 MULTICAST GROUPS; 4 USERS PER GROUP IN 2 LOCATIONS

SINR	feas. $\mathcal{Q}_r$	opt. $\mathcal{Q}_r$	feas. appr.	all solutions mean	all solutions std	appr. solutions mean	appr. solutions std
Group 1 (2 at L2 & 2 at L3) & Group 2 (2 at L1 & 2 at L6)							
6	100	95.09	99.54	1.0112	0.0716	1.2507	0.2367
8	100	92.01	99.77	1.0264	0.1277	1.3388	0.3242
10	99.66	94.96	99.66	1.0059	0.0421	1.1258	0.1519
12	95.21	96.28	99.40	1.0121	0.1395	1.3863	0.7031
14	82.08	97.22	99.72	1.0107	0.0989	1.4243	0.4757
16	64.16	97.86	99.82	1.0092	0.0937	1.4705	0.5024
18	43.84	98.44	100	1.0040	0.0395	1.2534	0.2096
20	24.54	99.07	100	1.0019	0.0194	1.2014	0.0238
Group 1 (2 at L1 & 2 at L3) & Group 2 (2 at L2 & 2 at L6)							
6	99.89	95.20	99.89	1.0207	0.2014	1.4405	0.8339
8	99.32	91.72	99.20	1.0313	0.1918	1.4162	0.5770
10	95.89	94.64	99.52	1.0287	0.3514	1.5852	1.4980
12	83.79	95.37	99.73	1.0168	0.1532	1.3848	0.6380
14	60.16	98.29	100	1.0059	0.0700	1.3431	0.4384
16	32.08	99.64	100	1.0030	0.0511	1.8562	0
18	13.01	100	-	1	0	-	-
20	4.00	100	-	1	0	-	-
Group 1 (2 at L2 & 2 at L6) & Group 2 (2 at L5 & 2 at L7)							
6	100	81.39	99.43	1.0518	0.1817	1.2854	0.3404
8	99.66	81.21	98.51	1.0529	0.1808	1.3014	0.3345
10	96.12	85.87	98.81	1.0413	0.1690	1.3150	0.3642
12	82.76	90.21	98.90	1.0374	0.2634	1.4262	0.7954
14	57.42	93.24	98.82	1.1537	2.8925	3.7284	12.1003
16	30.48	92.51	98.88	1.0230	0.1203	1.3573	0.3329
18	13.47	94.92	99.15	1.0132	0.0740	1.3091	0.2106
20	5.94	92.31	100	1.0342	0.1755	1.4442	0.5297

TABLE VIII

2 MULTICAST GROUPS; 4 USERS PER GROUP IN 3 LOCATIONS

SINR	feas. $\mathcal{Q}_r$	opt. $\mathcal{Q}_r$	feas. appr.	all solutions mean	all solutions std	appr. solutions mean	appr. solutions std
Group 1 (1 at L1, 1 at L2 & 2 at L3) & Group 2 (1 at L5, 2 at L6 & 1 at L7)							
6	99.89	86.40	99.31	1.0320	0.2806	1.2459	0.7464
8	98.17	88.26	98.72	1.0319	0.2202	1.3006	0.6166
10	93.61	88.29	98.78	1.0307	0.1664	1.2890	0.4335
12	72.95	92.02	99.06	1.0276	0.1933	1.3888	0.6271
14	47.49	94.95	99.04	1.0116	0.0736	1.2800	0.2436
16	24.43	97.67	100	1.0333	0.4264	2.4273	2.6821
18	12.10	98.11	100	1.0017	0.0147	1.0897	0.0824
Group 1 (1 at L1, 1 at L3 & 2 at L6) & Group 2 (1 at L2, 1 at L5 & 2 at L7)							
6	100	72.37	98.06	1.1180	0.5397	1.4503	0.9824
8	99.43	74.97	97.70	1.0897	0.3115	1.3856	0.5513
10	93.38	80.32	97.31	1.1802	2.7113	2.0320	6.4391
12	72.60	87.11	97.33	1.0465	0.2447	1.4429	0.6323
14	44.06	88.60	97.93	1.0741	0.7053	1.7781	2.1897
16	22.60	92.93	98.48	1.0292	0.1804	1.5183	0.5936

TABLE IX

2 MULTICAST GROUPS; 6 USERS PER GROUP IN 2 LOCATIONS

SINR	feas. $\mathcal{Q}_r$	opt. $\mathcal{Q}_r$	feas. appr.	all solutions mean	all solutions std	appr. solutions mean	appr. solutions std
Group 1 (3 at L2 & 3 at L3) & Group 2 (3 at L1 & 3 at L6)							
6	100	87.10	98.74	1.0465	0.2543	1.3943	0.6438
8	99.77	82.95	97.71	1.0690	0.6592	1.4569	1.6483
10	84.13	83.45	95.79	1.1520	1.4584	2.1791	3.9289
12	32.53	90.18	97.19	1.1019	0.9403	2.4109	3.3016
14	8.90	92.31	97.44	1.0269	0.1923	1.5115	0.7708
Group 1 (3 at L1 & 3 at L3) & Group 2 (3 at L2 & 3 at L6)							
6	100	73.17	97.72	1.1900	1.4948	1.7566	2.9149
8	90.41	68.06	94.44	1.3882	2.4513	2.3839	4.4924
10	60.84	65.85	92.31	1.3287	1.0586	2.1469	1.7277
12	17.69	72.26	91.61	1.3294	1.1546	2.5589	2.1211
Group 1 (3 at L2 & 3 at L3) & Group 2 (3 at L5 & 3 at L7)							
6	79.57	50.22	85.80	1.7201	3.1468	2.7365	4.7076
8	36.99	60.19	85.80	1.9784	6.2140	4.2771	11.0821
10	11.87	67.31	90.38	1.6219	1.7062	3.4356	2.6761

TABLE X

2 MULTICAST GROUPS; 6 USERS PER GROUP IN 3 LOCATIONS

SINR	feas. $\mathcal{Q}_r$	opt. $\mathcal{Q}_r$	feas. appr.	all solutions mean	all solutions std	appr. solutions mean	appr. solutions std
Group 1 (2 at L1, 2 at L2 & 2 at L3) & Group 2 (2 at L5, 2 at L6 & 2 at L7)							
6	92.69	59.73	91.38	1.4015	1.2621	2.1591	1.9312
8	61.99	67.04	88.40	1.3053	1.2901	2.2634	2.3898
10	13.47	83.05	94.92	1.3440	2.6684	3.7517	7.3235
Group 1 (2 at L1, 2 at L3 & 2 at L6) & Group 2 (2 at L2, 2 at L5 & 2 at L7)							
6	94.86	46.93	87.97	1.9805	4.1862	3.1019	5.9382
8	44.41	32.96	75.32	1.8251	2.9113	3.7787	4.8256
10	7.19	82.54	96.83	1.1688	0.8581	2.1441	2.0659
Group 1 (2 at L1, 2 at L2 & 2 at L6) & Group 2 (2 at L3, 2 at L5 & 2 at L7)							
6	70.21	24.07	81.63	2.2538	3.5065	2.7779	4.0640
8	32.53	41.40	81.40	2.0232	3.5807	3.0824	4.8974
10	7.42	47.6923	64.6154	1.1750	0.4518	1.6682	0.6888
Group 1 (2 at L1, 2 at L3 & 2 at L5) & Group 2 (2 at L2, 2 at L6 & 2 at L7)							
6	83.33	58.36	89.18	1.6009	2.7181	2.7385	4.4104
8	43.38	71.05	89.21	1.1648	0.6820	1.8097	1.3349
10	14.61	80.47	93.75	1.0962	0.3556	1.6790	0.7211
Group 1 (2 at L2, 2 at L3 & 2 at L5) & Group 2 (2 at L1, 2 at L6 & 2 at L7)							
6	97.26	47.30	92.49	1.4560	1.4843	1.9334	2.0170
8	68.95	63.91	88.91	1.2968	0.9642	2.0553	1.5861
10	10.89	86.34	93.44	1.0478	0.2604	1.6284	0.7519

TABLE XI

2 MULTICAST GROUPS; 8 USERS PER GROUP IN 2 LOCATIONS

SINR	feas. $\mathcal{Q}_r$	opt. $\mathcal{Q}_r$	feas. appr.	all solutions mean	all solutions std	appr. solutions mean	appr. solutions std
Group 1 (4 at L2 & 4 at L3) & Group 2 (4 at L1 & 4 at L6)							
6	100	80.37	97.72	1.0984	0.4743	1.5540	1.0098
8	96.23	83.87	97.03	1.0713	0.3764	1.5256	0.9007
10	48.86	82.71	96.50	1.1279	0.7362	1.8951	1.7752
12	5.82	72.55	92.16	1.5911	3.1056	3.7783	6.5226
Group 1 (4 at L1 & 4 at L3) & Group 2 (4 at L2 & 4 at L6)							
6	85.96	56.57	89.51	1.7763	3.1965	3.1098	5.0015
8	45.89	51.74	83.83	2.2882	6.8722	4.3653	10.8132
10	15.18	73.68	92.48	1.3903	2.1492	2.9202	4.5190
Group 1 (4 at L2 & 4 at L3) & Group 2 (4 at L5 & 4 at L7)							
6	50.46	39.14	66.74	1.8379	2.1140	3.0261	2.9037
8	2.86	64	68	1.2478	1.0216	5.21233	0

TABLE XII

3 MULTICAST GROUPS; 3–4 USERS PER GROUP IN 1 LOCATION

SINR	feas. $\mathcal{Q}_r$	opt. $\mathcal{Q}_r$	feas. appr.	all solutions mean	all solutions std	appr. solutions mean	appr. solutions std
Group 1 (3 at L1), Group 2 (3 at L2) & Group 3 (3 at L3)							
6	72.15	97.94	99.84	1.0069	0.1004	1.3638	0.6604
8	36.87	99.38	100	1.0006	0.0085	1.0961	0.0712
10	13.93	100	-	1	0	-	-
Group 1 (4 at L1), Group 2 (4 at L2) & Group 3 (4 at L3)							
6	29.11	94.90	99.22	1.0155	0.1085	1.3556	0.4043
8	7.65	100	-	1	0	-	-
Group 1 (4 at L3), Group 2 (4 at L6) & Group 3 (4 at L7)							
6	9.46	96.39	100	1.0174	0.1191	1.4820	0.4954

TABLE XIII

MEASURED CHANNELS; JMMF TX BEAMFORMING; $P = 1000$, 300 RANDOMIZATIONS

Group 1	Group 2	opt. $\mathcal{F}_r$	all solutions mean	all solutions std	appr. solutions mean	appr. solutions std
4@L1	4@L3	75.26	1.0009	0.0096	1.0035	0.0191
2@L1, 2@L3	2@L2, 2@L6	87.56	1.0040	0.0230	1.0321	0.0580
2@L5, 2@L7	2@L2, 2@L6	82.42	1.0065	0.0308	1.0372	0.0655
1@L1, 1@L2, 2@L3	1@L5, 2@L6, 1@L7	77.72	1.0179	0.0570	1.0653	0.0937
3@L2, 3@L3	3@L5, 3@L7	20.89	1.0771	0.1126	1.0974	0.1185
3@L2, 3@L3	3@L1, 3@L6	45.89	1.0154	0.0516	1.0285	0.0674
2@L1, 2@L2, 2@L3	2@L5, 2@L6, 2@L7	37.56	1.0449	0.0915	1.0720	0.1071
2@L1, 2@L2, 2@L6	2@L3, 2@L5, 2@L7	27.85	1.0957	0.1315	1.1327	0.1381
4@L2, 4@L3	4@L1, 2@L6	38.13	1.0157	0.0437	1.0254	0.0533
4@L2, 4@L3	4@L5, 2@L5	9.13	1.1084	0.1380	1.1193	0.1402

Fast and Effective Hybrid Multidimensional Scaling Approach for Node Localization in Wireless Sensor Networks

Georgios Latsoudas, Nicholas D. Sidiropoulos, *Senior Member, IEEE*

Abstract

Given a set of pairwise distance estimates between nodes, it is often of interest to generate a map of node locations. This is an old nonlinear estimation problem that has recently drawn interest in the signal processing community, due to the emergence of wireless sensor networks. Sensor maps are useful for estimating the spatial distribution of measured phenomena, and for routing purposes. We propose a two-stage algorithm that combines algebraic initialization and gradient descent. In particular, we borrow an algebraic solution known as *Fastmap* from the database literature and adapt it to the sensor network context, using a specific choice of anchor/pivot nodes. The resulting estimates are fed to a gradient descent iteration. The overall algorithm offers very competitive performance at significantly lower complexity than existing solutions with similar estimation performance. For a certain multiplicative measurement noise model that is often adopted in the literature, we also derive the pertinent Cramér-Rao bound (CRB). Simulations indicate that the performance of our algorithm is close to the CRB when the network is (close to) fully connected, in the sense that every node can estimate its distance from all (most) other nodes. Our adaptation of Fastmap also turns out to make a big difference when used to initialize other iterative distributed estimation algorithms that have been developed specifically for sparse networks.

I. Introduction

The problem of node localization from pairwise distance estimates has recently attracted interest in the signal processing and communications literature (e.g., [1], [2], [4], [6]), owing to the recent interest in wireless sensor networks. Given a matrix of pairwise distances (usually estimated using received signal strength measurements and a path loss model), the localization problem aims to

Submitted to *IEEE Trans. on Signal Processing*, July 28, 2006. Earlier version of part of this work appeared in conference form in the *Proc. of IEEE CAMSAP*, Dec. 12-14, 2005, Puerto Vallarta, Mexico. Supported by the U.S. ARO under ERO Contract N62558-03-C-0012. The authors are with the Department of Electronic and Computer Engineering, Technical University of Crete, 73100 Chania - Crete, Greece; Fax: +30-28210-37542, Phone: +30-28210-37227, E-mail: (latsoud,nikos)@telecom.tuc.gr

determine the (*relative*) node locations that generate these distances. In other words, one seeks a map of sensor locations with a given (approximate) distance structure. This is a classic problem originating in psychometrics [7], [8], known as *Multi-Dimensional Scaling* (MDS). There are many MDS flavors and variants; perhaps the single most important version is *metric MDS*. The classical approach to solving MDS is based on computing the principal components of a double-centered version of the distance matrix. This works reasonably well (albeit not optimally, due to the double centering), but its complexity is cubic in the number of nodes, and thus does not scale well with network size. A popular alternative to principal component analysis (PCA) is the use of gradient descent or other numerical optimization tools that aim to optimize a *stress function*. The stress function measures the error between the given distances and those reproduced by a given configuration of points. The drawback of gradient descent and related approaches is that they require accurate initialization.

We propose a two-stage MDS algorithm that employs an algebraic initialization procedure followed by gradient descent. The algebraic initialization is based on the Fastmap [3] algorithm, borrowed from the database literature. Fastmap is a linear-complexity mapping tool, which is, however, sensitive to range measurement errors.

Due to the fact that distances are invariant to coordinate frame transformations (rotation, reflection, shift), there is a need to employ three so-called *anchor nodes*, whose position is accurately known (e.g., via GPS) in order to fix a desired coordinate frame. Unfortunately, Fastmap is very sensitive to coordinate alignment, because the estimated position of every node (and thus anchor nodes as well) is only based on distances to selected *pivot* nodes - there is no averaging. In order to mitigate this problem, we advocate a judicious choice of anchor/pivot nodes, placed at the outer edges of the network. This placement bypasses the need for alignment and thus alignment errors, thereby providing a high-quality initialization to the gradient descent. The overall algorithm affords better localization accuracy than PCA-based MDS, at substantially lower complexity cost (quadratic in the number of nodes). Our algorithm is also competitive with respect to recent low-complexity solutions (e.g., [2]), especially when the network is (close to) fully connected. Finally, our adaptation of Fastmap also makes a big difference when used to initialize other iterative distributed estimation algorithms, specifically developed for sparse networks.

The rest of this paper is structured as follows. In section II we explain in detail the PCA-based MDS algorithm, and the standard gradient descent-based MDS. The Fastmap algorithm is briefly reviewed in section III. In section IV we describe the proposed hybrid algorithm, while in section V we summarize Costa's distributed MDS algorithm [2]. Section VI presents the CRB for a certain multiplicative measurement noise model that is often adopted in the literature on node localization in sensor networks [1], [6]. Section VII contains simulation results illustrating the performance of the

above algorithms, and the CRB. We remark that there are other algorithms in the recent literature that assume a different measurement model (e.g., 0-1 node connectivity only, as in [6]), or propose solutions of considerably higher complexity (e.g., as in [1]). We aim for the low-complexity regime, for simplicity and scalability considerations. Conclusions are drawn in section VIII.

II. Multidimensional Scaling

MDS [7], [8] has its origins in psychometrics and psychophysics. MDS postulates that perceptual or objective "dissimilarities" or "distances" between pairs of abstract "objects" can be be generated by points in m-dimensional space. Any set of distances obeying the triangle inequality can be reproduced (or closely approximated) by choosing m to be sufficiently large; but usually $m = 2$ or $m = 3$ is chosen to retain the systematic variation, and also for ease of visualization. Thus, MDS aims to find a geometric representation of the data in 2-D or 3-D space, such that the distances between data points fit as well as possible the given dissimilarity information.

We denote the dissimilarity measure (the estimated distances in our case), between objects i and j as d_{ij}. The set of dissimilarities yields a measured distance matrix $\mathbf{D}$. We also let $\hat{d}_{ij}$ denote the Euclidean distance between (generated by) two points $X_i = (x_{i1}, x_{i2}, ..., x_{im})$ and $X_j = (x_{j1}, x_{j2}, ..., x_{jm})$, i.e.

$$\hat{d}_{ij} = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2}. \tag{1}$$

In classical metric MDS, we estimate the node coordinates $\mathbf{X}$ by computing the m principal components of a double-centered and element-wise squared version of the matrix $\mathbf{D}$, denoted by $\mathbf{B}$:

$$B = -\frac{1}{2}\mathbf{JPJ}, \tag{2}$$

where $\mathbf{P} = D \odot D$ is the matrix of squared distances ($\odot$ denotes the element-wise matrix product), and $\mathbf{J}$ is the centering operator,

$$\mathbf{J} = \mathbf{I} - \mathbf{e}\mathbf{e}^T/N, \tag{3}$$

with N denoting the number of objects (sensor nodes), and $\mathbf{e}$ denoting the $N \times 1$ vector of all 1's . For an $N \times N$ matrix $\mathbf{D}$ and for m dimensions, it can be shown that

$$-\frac{1}{2}(d_{ij}^2 - \frac{1}{N}\sum_{j=1}^{N} d_{ij}^2 - \frac{1}{N}\sum_{i=1}^{N} d_{ij}^2 + \frac{1}{N^2}\sum_{j=1}^{N}\sum_{i=1}^{N} d_{ij}^2) = \sum_{k=1}^{m} x_{ik}x_{jk}, \tag{4}$$

thus the estimated node coordinates are given by the m principal eigenvectors of the matrix $\mathbf{B}$, scaled by the square roots of the corresponding eigenvalues. That is , with $\mathbf{U}_r$ containing the m principal eigenvectors and $\mathbf{V}_r$ diagonal containing the corresponding eigenvalues, $\mathbf{B}_r = \mathbf{U}_r\mathbf{V}_r\mathbf{U}_r^T$ is an optimal least squares approximation of B, and $\mathbf{X}_r = \mathbf{U}_r\mathbf{V}_r^{1/2}$ is an approximation of the node coordinates in m-dimensional space, up to a common coordinate rotation, reflection, and shift. An

alignment procedure is necessary to transform the estimated node locations to a desired frame of reference.

It is important to note that, due to the preprocessing steps prior to PCA, this approach is not equivalent to nonlinear least-squares parameter fitting using the original measurements.

Direct minimization of a suitable *stress function* is an alternative to PCA-based MDS [7]. A common[1] stress function is

$$stress^2 = \sum_{i,j} w_{ij}(\hat{d}_{ij} - d_{ij})^2. \tag{5}$$

Where $[w_{ij}]$ is the weight matrix, whose elements are equal to 1 if node j is in the measurement range of node i and 0 otherwise. Minimization starts with an initial guess of the node positions (often random), followed by gradient descent iterations. Initialization matters a lot in this context, because the stress function is multi-modal. Furthermore, the number of iterations required for convergence depends heavily on the quality of the initialization.

III. FASTMAP

The basic element of Fastmap [3] is the projection of the nodes on a properly selected line. This is achieved by selecting two objects O_a, O_b, called *pivots*, and projecting all other objects on the line that passes through them. A pair of pivots is chosen for each of the m dimensions. The coordinates, (i.e. projections on the pivot line) of the objects can be found by employing the *cosine law* [3]. Thus, the first coordinate for object O_i is given by:

$$x_i = \frac{d_{ai}^2 + d_{ab}^2 - d_{bi}^2}{2d_{ab}}, \tag{6}$$

where d_{ij} is the dissimilarity measure between nodes i and j and a, b are the pivot objects. After computing these coordinates for each object O_i, we consider a hyperplane which is orthogonal to the pivot line. We then project the objects on this hyperplane, and repeat the process, this time using

$$\tilde{d}_{ij}^2 = d_{ij}^2 - (x_i - x_j)^2, \quad i, j = 1, ..., N. \tag{7}$$

A heuristic method is proposed in [3] for choosing the pivots as far as possible from one another.

In database applications there is no "natural" or preferred coordinate frame of reference, thus the final alignment step is not used, and anchors are not needed. In the context of sensor networks,

[1]The negative log-likelihood of the observed data under a suitable measurement noise model would seem to be the natural choice of stress function. This is not fortuitus in our context, however, because the resulting function is not only multi-modal, but also leads to numerical difficulties. For this reason, a least squares criterion is preferred. While still multi-modal, the adopted least squares criterion is much more benign from a numerical optimization viewpoint, and it often yields performance close to the pertinent CRB, as will be seen in the simulations.

however, obtaining absolute position estimates is important. Unfortunately, Fastmap is very sensitive to coordinate alignment, because the estimated position of every node (and thus anchor nodes as well) is only based on distances to the chosen pivot nodes - there is no averaging. In order to mitigate this problem, we advocate a particular choice of anchor/pivot nodes, placed at the outer edges of the network. In particular, we assume that the sensor nodes are spread over a square, and place the anchor nodes, which will also serve as pivots, at three vertices (see Fig. 1). This placement bypasses the need for alignment and thus alignment errors, thereby providing a high-quality initialization to the gradient descent. Anchors #1 and #2 also serve as pivots for determining the coordinates in the first dimension, while anchors #2 and #3 double as pivots for the second dimension.

We assume that the anchor/pivot nodes which are used by the Fastmap can take distance measurements from all the sensor nodes, (even if we don't have full connectivity information for the rest of the nodes). This is reasonable if the anchor/pivot nodes are airborne or in higher ground.

IV. A Two-stage Approach

Fastmap is a fast algebraic mapping method that is rather sensitive to measurement errors, particularly so in the final alignment step. In our context, this sensitivity can be mitigated by the proposed choice of anchor/pivot nodes. The resulting estimates can be used as initialization for gradient descent. Each step of gradient descent costs $\mathcal{O}(N^2)$. Assuming good-enough initialization, only a few gradient descent steps will be needed. This suggests that a substantial complexity reduction relative to PCA and other techniques is possible. Interestingly, estimation accuracy can be improved as well, as we will see.

The basic steps of the two-stage algorithm are shown in Table I. Denoting by (x_i, y_i) the estimated position of node i, the partial derivative of the stress function in (5) is given by

$$\frac{\partial stress}{\partial x_i} = \sum_{j \neq i} w_{ij} \frac{(\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} - d_{ij})(x_i - x_j)}{\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}}. \tag{8}$$

with a similar expression for the partial derivative with respect to y_i. For simplicity, but also to bound complexity, a fixed number $p = 10$ of gradient descent steps is used in our simulations.

V. Costa's Algorithm

An iterative distributed estimation algorithm for MDS has been recently proposed in [2], using the principle of *majorization*. The idea behind majorization is simple. Instead of directly minimizing a complicated cost/stress function, majorization uses a simpler (usually quadratic) *majorizing* function that lies over the said cost/stress function and is equal to it at the current parameter estimate. Minimizing the majorizing function thus yields a new parameter estimate whose cost/stress is lower

than or equal to that of the previous one. Continuing in this fashion yields a sequence of parameter estimates of decreasing cost/stress values. Specializing to the present context [2] yields the following update

$$\mathbf{x_i}^k = a_i \mathbf{X}^{k-1} \mathbf{b}_i^{k-1}, \tag{9}$$

where $\mathbf{X}^k$ is the matrix which contains the position estimates for all the sensor nodes in the kth iteration of the algorithm, and a is a parameter given by

$$a_i^{-1} = \sum_{j=1, j\neq i}^{N-M} w_{ij} + \sum_{j=N-M+1}^{N} 2w_{ij}, \tag{10}$$

where M is the number of anchor nodes ($M = 3$ in the 2-D case). The entries of the $N \times 1$ vector $\mathbf{b}_i$ are given by

$$\begin{aligned} \mathbf{b}_i(j) &= w_{ij}(1 - d_{ij}/\hat{d}_{ij}), \quad j \leq N - M, j \neq i, \\ \mathbf{b}_i(j) &= 2w_{ij}(1 - d_{ij}/\hat{d}_{ij}), \quad j > N - M, j \neq i \\ \mathbf{b}_i(i) &= \sum_{i=1}^{N-M} w_{ij} d_{ij}/\hat{d}_{ij} + \sum_{j=N-M+1}^{N} 2w_{ij} d_{ij}/\hat{d}_{ij}, \end{aligned} \tag{11}$$

where $\hat{d}_{ij}$ is the reproduced distance computed from the coordinate estimates at iteration k. The algorithm runs iteratively and the requisite computation can be performed at each node in a distributed function (every node computes its own position coordinates and the corresponding part of the cost function). The iterations continue until the associated sequence of costs converges within ϵ in the Cauchy sense. The cost function which the authors in [2] propose is:

$$S = \sum_{i=1}^{N-M} S_i, \tag{12}$$

where the local cost functions S_i are given by:

$$S_i = \sum_{j=1, j\neq i}^{N-M} w_{ij}(d_{ij} - \hat{d}_{ij})^2 + \sum_{j=N-M+1}^{N} 2w_{ij}(d_{ij} - \hat{d}_{ij})^2 \tag{13}$$

When the difference between the previous and the current cost values becomes smaller than a threshold ϵ the algorithm terminates. This is guaranteed due to the fact that a single iteration can reduce or maintain, but cannot increase the cost, which is also bounded from below.

VI. Measurement Noise Model and Cramér-Rao Bound

Pairwise distance estimates will inevitably contain measurement errors, which are generally amplified with increasing distance between nodes. The choice of measurement noise model depends on many factors, and is application-specific. We shall adopt a certain multiplicative noise model from the recent literature on node localization in wireless sensor networks [1], [6], in which the distance

measurement error is proportional to the actual distance between the pair of nodes. Thus the measured distance d_{ij} between nodes i, j is assumed to be drawn from

$$d_{ij} \sim \delta_{ij} + \delta_{ij}\mathcal{N}(0, e_r^2), \tag{14}$$

where δ_{ij} is the actual distance between nodes i, j and e_r^2 is the range error variance. We also assume that the measurements are reciprocal (or symmetrized by averaging prior to further processing); i.e., $d_{ij} = d_{ji}$.

In this section, we derive the Cramér-Rao Bound (CRB) for node localization using the above multiplicative noise model. Analogous derivations for different noise models employed in [2], [5] can be found in [5]. An explanation of the difference between the RSS noise model described therein and our multiplicative noise model can be found in the appendix.

Define the vector of sensor parameters $\boldsymbol{\gamma} = (\boldsymbol{\gamma_1}\boldsymbol{\gamma_2}...\boldsymbol{\gamma_N})$. Each $\boldsymbol{\gamma_i}$ contains the location coordinates for node i, i.e., $\boldsymbol{\gamma}_i = (x_i,\ y_i])$ in the 2-D case. The unknown parameter vector for the $N-3$ sensors whose locations are unknown[2] is defined as $\boldsymbol{\theta} = (\boldsymbol{\theta}_x\ \boldsymbol{\theta}_y)$, with $\boldsymbol{\theta}_x = (x_1, x_2, ..., x_{N-3})$ and $\boldsymbol{\theta}_y = (y_1, y_2, ..., y_{N-3})$. This is the vector we wish to estimate. Sensors i, j perform pairwise observations d_{ij}. We assume that the observations d_{ij} are statistically independent for $i < j$. The density function of the observations d_{ij} given the locations of nodes i, j is denoted by $f(d_{ij}|\gamma_i,\ \gamma_j)$. Thus the joint log-likelihood is

$$\begin{aligned} l(\mathbf{D}, \boldsymbol{\gamma}) &= \sum_{i=1}^{N} \sum_{j \in H(i),\ j<i} l_{i,j}, \\ l_{i,j} &= log f(d_{ij}|\gamma_i, \gamma_j) \end{aligned} \tag{15}$$

where $H(i)$ is the set of nodes which are in the range of node i.

The CRB for coordinate θ_i is $\text{cov}(\theta_i) \geq [{\mathbf{F}_\theta}^{-1}]_{ii}$, where $\mathbf{F}_\theta$ is the Fisher Information Matrix (FIM), given by

$$\mathbf{F}_\theta = \begin{bmatrix} \mathbf{F_{xx}} & \mathbf{F_{xy}} \\ \mathbf{F_{xy}^T} & \mathbf{F_{yy}} \end{bmatrix}. \tag{16}$$

The elements for the sub-matrix $\mathbf{F_{xx}}$ are given by

$$\mathbf{F_{xx}}(k,l) = \begin{cases} -\sum_{j \in H(k)} E[\frac{\partial^2}{\partial x_k^2} l_{k,j}], & k = l \\ -I_{H(k)}(l) E[\frac{\partial^2}{\partial x_k \partial x_l} l_{k,l}], & k \neq l \end{cases}, \tag{17}$$

[2]In the 2-D case we need 3 anchor nodes.

where $I_{H(k)}(l)$ is the indicator function (1 if l is in the range of k, 0 otherwise). Similar expressions hold for the $\mathbf{F_{xy}}, \mathbf{F_{yy}}$ sub-matrices. For full connectivity, the elements of the above matrices are

$$\mathbf{F_{xx}}(k,l) = \begin{cases} -\sum_j \frac{2(x_k-x_j)^2}{\delta_{kj}^4} - \frac{1}{\delta_{kj}^2} - \frac{e_r^2+1}{e_r^2}(-\frac{1}{\delta_{kj}^2} + 4\frac{(x_k-x_j)^2}{\delta_{kj}^4}) - \frac{1}{e_r^2}(\frac{1}{\delta_{kj}^2} - 3\frac{(x_k-x_j)^2}{\delta_{kj}^4}), & k = l, \\ -(\frac{1}{d_{kl}^2} - 2\frac{(x_k-x_l)^2}{\delta_{kl}^4} + \frac{1+e_r^2}{e_r^2}(4\frac{(x_k-x_l)^2}{\delta_{kl}^4} - \frac{1}{\delta_{kl}^2}) - \frac{1}{e_r^2}(3\frac{(x_k-x_l)^2}{\delta_{kl}^2} - \frac{1}{\delta_{kl}^2})), & k \neq l \end{cases} \tag{18}$$

(similar expressions can be obtained for the elements of $\mathbf{Fyy}$) and

$$\mathbf{F_{xx}}(k,l) = \begin{cases} -\sum_j 2(x_k - x_j)(y_k - y_j)\frac{1}{\delta_{kj}^4} - 4\frac{1+e_r^2}{e_r^2}\frac{(x_k-x_j)(y_k-y_j)}{\delta_{kj}^4} + \frac{3}{e_r^2}\frac{(x_k-x_j)(y_k-y_j)}{\delta_{kj}^4}, & k = l, \\ -(-2(x_k - x_l)(y_k - y_l)\frac{1}{\delta_{kl}^4} + 4\frac{1+e_r^2}{e_r^2}\frac{(x_k-x_l)(y_k-y_j)}{\delta_{kj}^4} - \frac{3}{e_r^2}\frac{(x_k-x_j)(y_k-y_j)}{\delta_{kj}^4}), & k \neq l \end{cases} \tag{19}$$

VII. Simulation Results

In this section, we compare the aforementioned algorithms in the context of node localization in sensor networks. Network nodes are considered to be uniformly distributed in a square with area equal to 1, i.e., the x and y coordinates of the sensor nodes are uniformly distributed in $[0, 1]$. We employ the alignment procedure described in [4], when necessary, in order to estimate the absolute coordinates, and adopt root mean squared error as our estimation performance metric:

$$RMSE := \frac{\sum_{i=1}^{N}\sqrt{(x_{ri}-x_{ei})^2 + (y_{ri}-y_{ei})^2}}{N}, \tag{20}$$

where x_{ei}, y_{ei} are the estimated coordinates, and x_{ri}, y_{ri} are the actual coordinates of sensor i. The computational complexity orders of the various algorithms under consideration are listed in Tables II and IV, for the case of full and partial connectivity, respectively.

The baseline[3] MDS algorithm is based on PCA of the doubly-centered matrix of squared distances, and henceforth referred to as PCA-based MDS. We also implemented Costa's iterative majorization algorithm. We tried both a random initialization and the alternative initialization strategy suggested in [2]. The latter strategy often yields complex coordinates when the triangle inequality fails due to measurement errors, whereas the former (random) yields unsatisfactory results that do not improve with decreasing error variance. It is clear that Costa's algorithm is sensitive with respect to initialization, and could benefit from a better "warm start". For this reason, we also tried using our adaptation of Fastmap to initialize Costa's iteration.

Fig. 2 shows the RMSE performance of the various algorithms (PCA, Fastmap, Fastmap+SD, Fastmap+Costa, and Costa with random initialization) for a sensor network with 80 sensors, as a

[3]PCA-based MDS is not directly applicable in the case of partial connectivity.

function of e_r^2. Distance measurements are drawn from the multiplicative noise model in (14). The corresponding Cramér-Rao Bound (CRB) is also plotted as a benchmark. For the SD step of the proposed algorithm (Fastmap+SD), a step-size of $\lambda = 0.01$ and $p = 10$ SD iterations were used. The convergence threshold in Costa's algorithm was set to $\epsilon = 0.1$. From Fig. 2, we observe that stand-alone Fastmap exhibits poor performance, which quickly degrades with increasing range error variance. When randomly initialized, Costa's algorithm also performs poorly in this setup, and its performance does not improve with decreasing error variance. Fastmap+SD and Fastmap+Costa are the best options from the viewpoint of RMSE performance, and remain relatively close to the CRB, especially for low range error variance. Interestingly, the proposed algorithm is not only less complex, but also more accurate than PCA. This is partially attributed to the fact that PCA uses double centering, which colors the noise, whereas the proposed algorithm directly minimizes the stress function.

Fig. 3 shows corresponding results for a network with 200 nodes ($\lambda = 0.005$; the remaining setup is the same as Fig. 2). The estimation accuracy of PCA, Fastmap+SD, and Fastmap+Costa, is improved relative to Fig. 2, as expected. Fastmap does not benefit, due to the lack of (implicit or explicit) averaging, while Costa's algorithm with random initialization actually does quite the same as in Fig. 2.

We also tried an additive measurement noise model, i.e., the measurements are drawn from

$$d_{ij} \sim \delta_{ij} + \mathcal{N}(0, e_r^2), \tag{21}$$

where the variance of the measurement error is independent of the distance between the two nodes. The results are shown in Fig. 4 for the case of 80 nodes, and in Fig. 5 for the case of 200 nodes. We observe again that Fastmap+SD and Fastmap+Costa yield approximately the same RMSE performance, significantly outperforming stand-alone Fastmap and PCA.

One might also wonder whether the RMSE comparison of the various algorithms is sensitive with respect to the statistics of the multiplicative noise (normal versus log-normal, see also the appendix). Fig. 6 presents simulation results for the log-normal multiplicative noise model employed in [2]. We observe that the relative performance ordering of the different algorithms is the same as in Fig. 2.

Fig. 7 shows the average computational cost in floating point operations (FLOPS) of Fastmap+SD and Fastmap+Costa, as a function of the number of nodes, N. We observe that Fastmap+SD exhibits significantly lower complexity (almost five times lower) than Fastmap+Costa. The values of the step-size λ used for the different values of N are listed in Table III.

In all simulation results presented so far, the network was assumed to be fully connected, i.e., distance measurements were available for each pair of nodes in the network. We now switch to partially connected scenarios. We assume that nodes which are further apart than a certain threshold (radio

range) cannot hear each other, the corresponding distance measurement is marked as unavailable, and the associated weight in the stress function is set to zero. An exception is that every node is assumed to be within range from each of the three anchor/pivot nodes. We adopt the multiplicative noise model in 14, and consider two cases: in the first the measurement range is 0.14 and in the second it is 0.3. Fig. 8 and Fig. 9 show the RMSE performance of Fastmap+SD, Fastmap+Costa, and the CRB (which accounts for the missing data) for the two cases, as a function of range error variance, for $N = 80$ nodes. Table V lists the values of λ used in the SD iteration for the three different connectivity scenarios (fully connected, partially connected with measurement range equal to 0.3, or 0.14) and $N = 80$. For Fastmap+SD, we tried two different values for the number of SD iterations: $p = 10$ and $p = 30$. From Fig. 8 and Fig. 9, we observe that Fastmap+Costa outperforms Fastmap+SD in terms of RMSE, even when $p = 30$ is used in SD. This is in contrast to the case of full connectivity. The corresponding FLOP counts in Fig. 10 and Fig. 11 show that Fastmap+SD with $p = 10$ maintains its computational complexity advantage compared to Fastmap+Costa. Increasing p improves the RMSE performance of Fastmap+SD, but at the cost of computational complexity, which is brought closer to that of Fastmap+Costa. We conclude that while Fastmap+SD offers lower complexity for the same RMSE performance as Fastmap+Costa in the fully connected case, there is a performance penalty for the reduced complexity in the partially connected case, wherein Fastmap+Costa may be preferable.

VIII. Conclusions

We have proposed a hybrid two-stage node localization algorithm that offers better accuracy than existing alternatives of the same (and, in certain cases, even higher) complexity order. The new algorithm employs Fastmap, coupled with judicious selection of anchor nodes that double as pivots, to generate a computationally cheap yet sufficiently accurate initialization for gradient descent. The new algorithm is particularly attractive (in terms of the offered performance-complexity trade-off) in the case of dense networks.

We also proposed using our adaptation of Fastmap as initialization for Costa's algorithm. The latter combination appears useful for sparse networks, in which case it attains better estimation performance than Fastmap+SD, albeit at a higher complexity cost. Our simulations indicate that, in the context of our present application, Fastmap+SD uniformly outperforms PCA-based MDS, both in terms of complexity and in terms of estimation accuracy. We have also derived the pertinent CRB for the multiplicative noise model in [1], [6], which was adopted for most of our simulations.

Appendix

Normal vs. log-normal multiplicative noise modelling: In [2], [5], the power received at node i from node j, measured in decibel (dB), is modelled as $P_{ij} = \bar{P}_{ij} + v$, where $\bar{P}_{ij}$ is the mean power,

and v is a zero-mean Gaussian random variable of standard deviation σ. The mean power is modelled as $\bar{P}_{ij} = P_0 - 10n_p log_{10}\frac{\delta_{ij}}{\delta_0}$, where P_0 is the mean power for a reference distance, δ_0, and n_p is the path loss exponent. It follows that

$$P_0 - P_{ij} = P_0 - \bar{P}_{ij} - v = 10n_p log_{10}\frac{\delta_{ij}}{\delta_0} - v, \tag{22}$$

and the associated distance estimate is given by [2]

$$d_{i,j} = \delta_0 10^{(P_0 - P_{ij})/10n_p}. \tag{23}$$

Substituting $P_{ij} = \bar{P}_{ij} + v$ and $\bar{P}_{ij} = P_0 - 10n_p log_{10}\frac{\delta_{ij}}{\delta_0}$ yields

$$d_{ij} = \delta_{i,j} 10^{-v/10n_p}. \tag{24}$$

Notice that the noise factor is *log-normal*, whereas in the model of [1], [6] (also adopted herein) the noise factor is normally distributed.

REFERENCES

[1] P. Biswas, T.-C. Liang, T.-C. Wang and Y. Ye, "Semidefinite Programming Based Algorithms for Sensor Network Localization," to appear in *ACM Trans. on Sensor Networks*, 2006. See also http://www.stanford.edu/∼yyye/

[2] J. A. Costa, N. Patwari, A. O. Hero, "Distributed Multidimensional Scaling with Adaptive Weighting for Node Localization in Sensor Networks," *ACM Trans. on Sensor Networks*, submitted.

[3] C. Faloutsos, K. Lin, "FastMap: A Fast Algorithm for Indexing, Data-Mining and Visualization of Traditional and Multimedia Datasets," in *Proc. ACM SIGMOD*, vol. 24, no. 2, pp. 163-174, 1995.

[4] X. Ji, H. Zha "Sensor Positioning in Wireless Ad-hoc Sensor Networks Using Multidimensional Scaling," in *Proc. Infocom*, pp. 2652-2661, 2004.

[5] N. Patwari, A. Hero, M. Perkins, N. Correal, R. O'Dea, "Relative Location Estimation in Wireless Sensor Networks," *IEEE Trans. on Signal Processing*, vol. 51, no. 8, pp. 2137-2148, Aug. 2003.

[6] Y. Shang, W. Ruml, Y. Zhang, M. Fromherz, "Localization from Connectivity in Sensor Networks," *IEEE Trans. on Parallel and Distr. Systems*, vol. 15, no. 11, pp. 961-974, Nov. 2004.

[7] W.S. Torgerson, "Multidimensional Scaling: I. Theory and method," *Psychometrika*, vol. 17, pp. 401-419, 1952.

[8] W.S. Torgerson, "Multidimensional Scaling of Similarity," *Psychometrika*, vol. 30, pp. 379-393, 1965.

TABLE I

TWO-STAGE FASTMAP+SD ALGORITHM

Input: $\mathbf{D}$

1) Run Fastmap using as pivots three anchor nodes, judiciously placed on the three vertices of the square distribution area. Let X be the vector containing the resulting estimated node coordinates.
2) For $i = 1$ to p

   begin

   - evaluate $\nabla stress$ at the point X
   - $X = X - \lambda \nabla stress$

   end

TABLE II

COMPUTATIONAL COMPLEXITIES FOR FULL CONNECTIVITY (N IS NUMBER OF NODES, m IS NUMBER OF SPATIAL DIMENSIONS)

Algorithm	Complexity
Fastmap	$\mathcal{O}(mN)$
Fastmap+SD	$\mathcal{O}(pmN^2)$, $p << N$
PCA	$\mathcal{O}(N^3)$
Costa's	$\mathcal{O}(kmN^2)$, $k << N$

TABLE III

CHOICE OF STEP-SIZE λ AS A FUNCTION OF THE NUMBER OF NODES N

N	λ
80	0.01
110	0.0075
140	0.007
170	0.006
200	0.005

TABLE IV

COMPUTATIONAL COMPLEXITIES FOR PARTIAL CONNECTIVITY (s IS THE AVERAGE NUMBER OF DISTANCE MEASUREMENTS COLLECTED BY A NODE)

Algorithm	Complexity
Fastmap	$\mathcal{O}(mN)$
Fastmap+SD	$\mathcal{O}(pmsN)$, $p << N$
Costa's	$\mathcal{O}(kmsN)$, $k << N$

TABLE V

CHOICE OF STEP-SIZE λ AS A FUNCTION OF MEASUREMENT RANGE

Measurement Range	λ
Infinite	0.01
0.3	0.0125
0.14	0.015

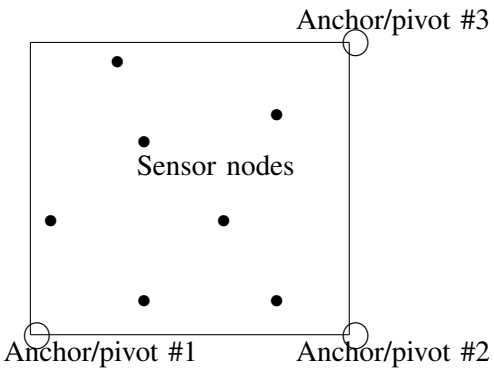


Fig. 1. Anchor/pivot node placement

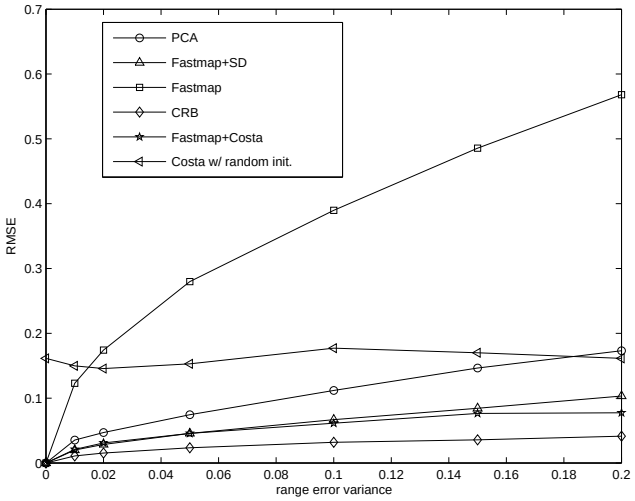


Fig. 2. RMSE performance vs. measurement range error variance. $N = 80$, all pairwise distance estimates collected. Measurement error proportional to the actual distance. 100 Monte Carlo runs.

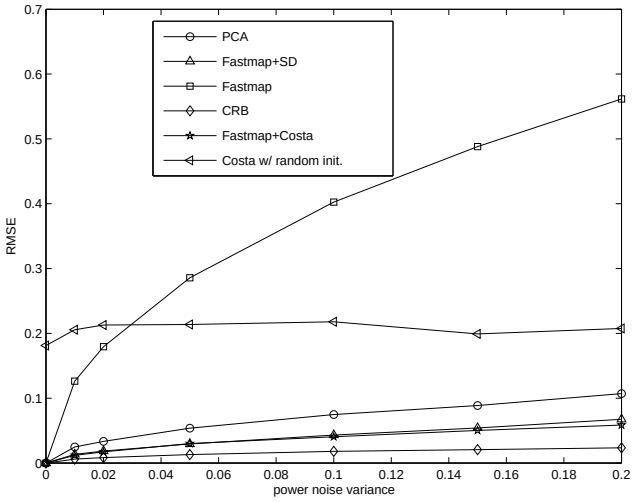


Fig. 3. RMSE performance vs. measurement range error variance. $N = 200$, all pairwise distance estimates collected. Measurement error proportional to the actual distance. 100 Monte Carlo runs.

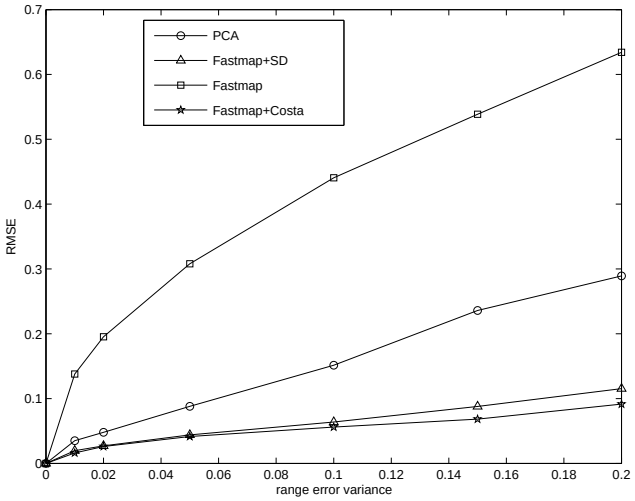


Fig. 4. RMSE performance vs. measurement range error variance. $N = 80$, additive noise measurement model, all pairwise distance estimates collected. 100 Monte Carlo runs.

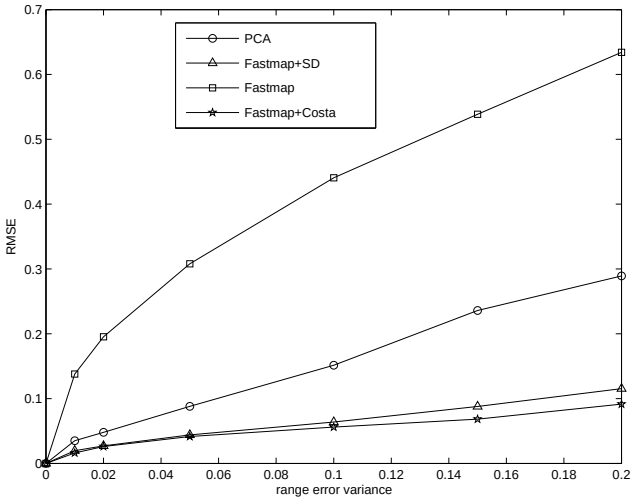


Fig. 5. RMSE performance vs. measurement range error variance. $N = 200$ sensor nodes, all pairwise distance estimates collected. Additive noise measurement model. 100 Monte Carlo runs.

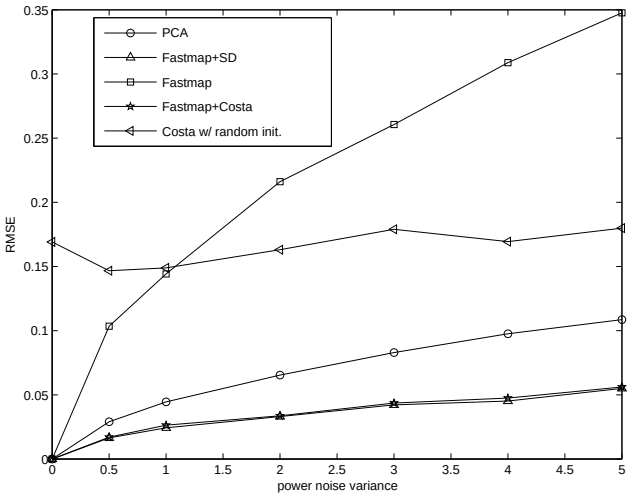


Fig. 6. RMSE performance vs. power noise variance σ^2. $N = 80$ sensor nodes, all pairwise distance estimates collected. Log-normal noise measurement model. 100 Monte Carlo runs.

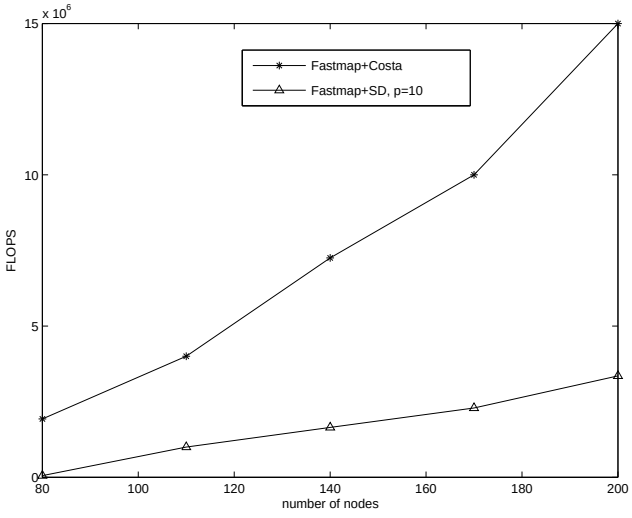


Fig. 7. Computational cost in FLOPS vs. number of nodes. All pairwise distance estimates collected, $e_r^2 = 0.1$, $\epsilon = 0.1$. Multiplicative noise measurement model. 50 Monte Carlo runs.

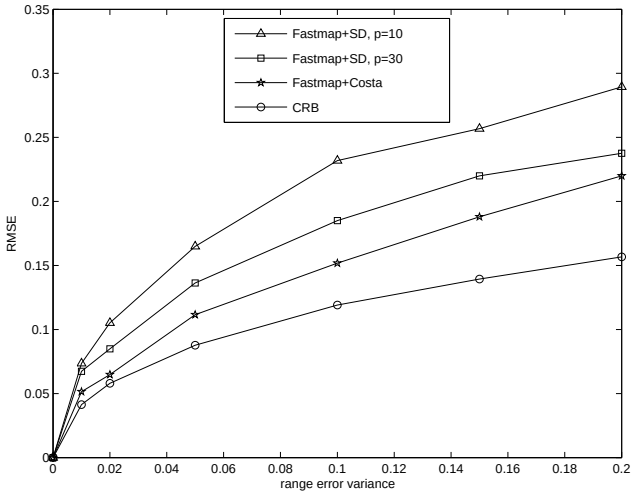


Fig. 8. RMSE performances and CRB for limited measurement range $= 0.14$ (the weights which correspond to distances greater than this limit are set to zero). $\epsilon = 0.1$, $\lambda = 0.015$, $N = 80$. 100 Monte Carlo runs.

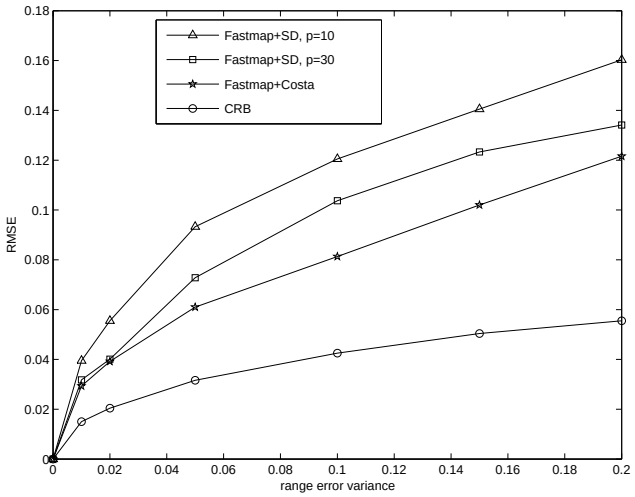


Fig. 9. RMSE performances and CRB for limited measurement range $= 0.3$. $\epsilon = 0.1$, $\lambda = 0.013$, $N = 80$. 100 Monte Carlo runs.

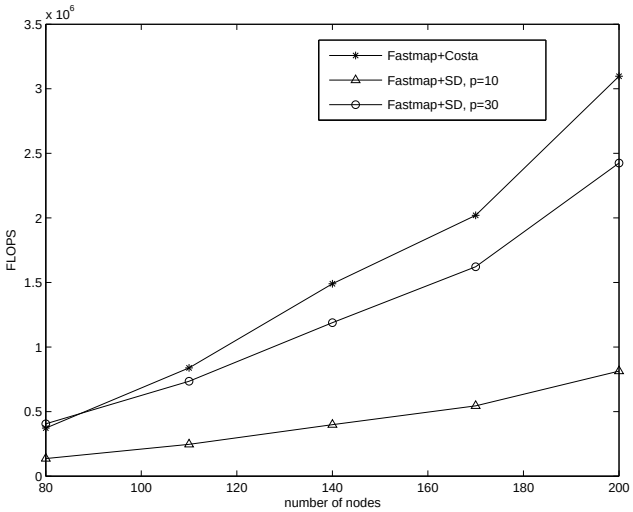


Fig. 10. Computational cost in FLOPS vs. number of nodes. Pairwise distances collected only for nodes with actual distance smaller than 0.3. $e_r^2 = 0.1$. Multiplicative measurement noise model. 50 Monte Carlo runs.

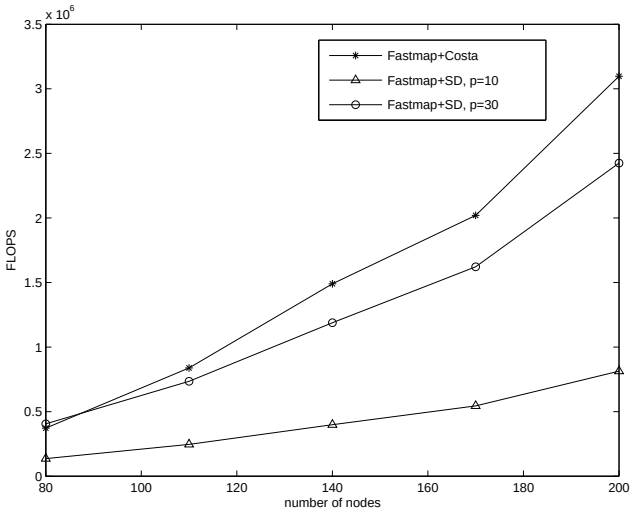


Fig. 11. Computational cost in FLOPS vs number of nodes. Pairwise distances collected only for nodes with actual distance smaller than 0.14. $e_r^2 = 0.1$. Multiplicative measurement noise model. 50 Monte Carlo runs.

A Semidefinite Relaxation Approach to MIMO Detection for High-order QAM Constellations†

Nicholas D. Sidiropoulos[1], *Zhi-Quan Luo*[2]

Abstract

A new and conceptually simple semidefinite relaxation approach is proposed for MIMO detection in communication systems employing high-order QAM constellations. The new approach affords improved detection performance compared to existing solutions of comparable worst-case complexity order, which is nearly cubic in the dimension of the transmitted symbol vector and independent of the constellation order for uniform QAM, or affine in the constellation order for non-uniform QAM.

SPL EDICS: Primary: COM-ESTI; Secondary: COM-MIMO

† Original manuscript submitted to *IEEE Signal Processing Letters*, Nov. 10 2005; revised Feb. 25, 2006. The work of the first author was supported in part by the U.S. ARO under ERO Contract N62558-03-C-0012, the E.U. under FP6 U-BROAD STREP # 506790, and the GSRT. The work of the second author was supported in part by the National Science Foundation, Grant No. DMS-0312416.

[1] The (corresponding) author is with the Department of Electronic and Computer Engineering, Technical University of Crete, 73100 Chania - Crete, Greece; Fax: +30-28210-37542, Phone: +30-28210-37227, E-mail: nikos@telecom.tuc.gr

[2] The author is with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455, U.S.A.; E-mail: luozq@ece.umn.edu

I. INTRODUCTION

Maximum likelihood (ML) detection in memoryless Multiple-Input Multiple-Output (MIMO) communication systems with Gaussian noise is equivalent to a least-squares lattice search problem which is NP-hard. For this reason, several computationally efficient approximate solutions have been developed. The current state-of-art includes two main families of high-performance MIMO detectors: those based on *Sphere Decoding* (SD) [11], [1], [2], [12], [14] and those based on *Semidefinite Relaxation* (SDR) [7], [6], [5], [13]. SD detectors can provide the exact ML solution at low computational cost, provided that the Signal to Noise Ratio (SNR) is relatively high, and the aggregate transmission rate is relatively low. However, SD cannot efficiently handle high problem dimensions (long symbol vectors) or high-order symbol constellations, especially at low SNR, and it has recently been shown that its *expected complexity* is exponential [4], under certain conditions that are relatively mild and general in our context. Worst-case complexity of computing the exact ML solution is generically exponential, due to NP-hardness.

In contrast, SDR approaches feature polynomial *worst-case* complexity and very competitive performance. Initially, SDR multiuser / MIMO detection was developed for Binary Phase-Shift Keying (BPSK) constellations, but the ideas were later extended to M-PSK [7], [6], [5], and, very recently, to 16- Quadrature Amplitude Modulation (16-QAM) [13] and general QAM constellations [8]. While [13] deals exclusively with 16-QAM, the approach can, in principle, be extended to higher-order QAM alphabets. This, however, entails the introduction of additional slack variables, and complexity becomes $O(K^{6.5}N^{6.5})$, where $N = O(M)$, M is the number of symbols, and K is the square root of the order of the constellation. The idea in [13] is fruitful for 16-QAM, but impractical for higher orders. Likewise, the complexity of the methods in [8] ranges from $O(K^{6.5}N^4)$ to $O(K^{6.5}N^{6.5})$.

In this contribution, we propose a different, $O(N^{3.5})$ relaxation for high-order QAM alphabets. Our approach can be viewed as further relaxation of [13], only utilizing upper and lower bounds on the symbol energy in the relaxation step. The key features of our approach are that i) it provides significant performance improvements relative to existing solutions of comparable worst-case complexity order; and ii) its complexity is independent of the constellation order for uniform QAM, and affine in the constellation order for non-uniform QAM. For BPSK and 4-QAM, our approach reduces to the one in [7].

II. Problem Statement and Preliminaries

For any separable QAM constellation[1], ML detection in memoryless MIMO communication systems with Gaussian noise can be formulated as the following optimization problem (possibly after noise pre-whitening):

$$\min ||\mathbf{d} - \mathbf{M}\mathbf{s}||_2^2 \tag{1}$$

$$\text{subject to: } Re\left\{\mathbf{s}(i)\right\} \in \mathcal{A}_{real},\ Im\left\{\mathbf{s}(i)\right\} \in \mathcal{A}_{imag},\ \forall i. \tag{2}$$

For brevity of exposition, we will assume that $\mathcal{A}_{real} = \mathcal{A}_{imag} = \mathcal{A}$ in the sequel, although our approach generalizes trivially to different alphabets for the real and imaginary parts. We thus consider

$$\min ||\mathbf{d} - \mathbf{M}\mathbf{s}||_2^2 \tag{3}$$

$$\text{subject to: } Re\left\{\mathbf{s}(i)\right\} \in \mathcal{A},\ Im\left\{\mathbf{s}(i)\right\} \in \mathcal{A},\ \forall i, \tag{4}$$

where $\mathbf{d}$ is the complex baseband received vector, $\mathbf{M}$ is a known baseband-equivalent channel matrix, and $\mathbf{s}$ is the symbol vector. Upon defining

$$\mathbf{z} := \left[Re\left\{\mathbf{d}\right\}^T\ Im\left\{\mathbf{d}\right\}^T\right]^T, \tag{5}$$

$$\mathbf{H} := \begin{bmatrix} Re\left\{\mathbf{M}\right\} & -Im\left\{\mathbf{M}\right\} \\ Im\left\{\mathbf{M}\right\} & Re\left\{\mathbf{M}\right\} \end{bmatrix} \tag{6}$$

$$\mathbf{r} := \left[Re\left\{\mathbf{s}\right\}^T\ Im\left\{\mathbf{s}\right\}^T\right]^T, \tag{7}$$

we may convert the problem to real-valued form

$$\min ||\mathbf{z} - \mathbf{H}\mathbf{r}||_2^2 \tag{8}$$

$$\text{subject to: } \mathbf{r}(i) \in \mathcal{A},\ \forall i. \tag{9}$$

III. Proposed solution

Assume that $\mathcal{A}$ is symmetric about the origin (always the case for QAM constellations). In this case, if $\mathbf{r}$ satisfies the finite alphabet constraints in (9), then so does $t\mathbf{r}$, for $t \in \{-1, 1\}$. Furthermore,

$$||\mathbf{z} - \mathbf{H}\mathbf{r}||_2^2 = \mathbf{r}^T\mathbf{H}^T\mathbf{H}\mathbf{r} - 2\mathbf{z}^T\mathbf{H}\mathbf{r} + \mathbf{z}^T\mathbf{z}. \tag{10}$$

[1]Separable constellations are almost always adopted for ease of decoding, even in the single-input single-output case.

It follows that (8)-(9) is equivalent to

$$\min \left(\mathbf{r}^T \mathbf{H}^T \mathbf{H} \mathbf{r} - 2 \mathbf{z}^T \mathbf{H} t \mathbf{r} \right) \tag{11}$$

$$\text{subject to: } \mathbf{r}(i) \in \mathcal{A},\ \forall i,\ t \in \{-1, 1\}. \tag{12}$$

Further defining $\mathbf{x} := \left[\begin{array}{cc} \mathbf{r}^T & t \end{array} \right]^T \in \mathbb{R}^N$ and

$$\mathbf{Q} := \left[\begin{array}{cc} \mathbf{H}^T \mathbf{H} & -\mathbf{H}^T \mathbf{z} \\ -\mathbf{z}^T \mathbf{H} & 0 \end{array} \right], \tag{13}$$

problem (11)-(12) can be put in homogeneous quadratic form

$$\min \mathbf{x}^T \mathbf{Q} \mathbf{x} \tag{14}$$

$$\text{subject to: } \mathbf{x}(i) \in \mathcal{A},\ \forall i \in \{1, \cdots, N-1\},\ \mathbf{x}(N) \in \{-1, 1\}. \tag{15}$$

Using $\mathbf{x}^T \mathbf{Q} \mathbf{x} = \text{Trace}(\mathbf{x}^T \mathbf{Q} \mathbf{x}) = \text{Trace}(\mathbf{Q} \mathbf{x} \mathbf{x}^T)$, and denoting $\mathbf{X} := \mathbf{x} \mathbf{x}^T$, we can rewrite problem (14)-(15) *equivalently* as:

$$\min \text{Trace}(\mathbf{Q} \mathbf{X}) \tag{16}$$

$$\text{subject to: } \mathbf{X} \geq \mathbf{0},\ \text{rank}(\mathbf{X}) = 1, \tag{17}$$

$$\mathbf{X}(i,i) \in \mathcal{A}^2,\ \forall i \in \{1, \cdots, N-1\},\ \mathbf{X}(N,N) = 1. \tag{18}$$

Problem (16)-(18) entails nonconvex constraints: the $\text{rank}(\mathbf{X}) = 1$ constraint, as well as the finite (squared) alphabet constraints $\mathbf{X}(i,i) \in \mathcal{A}^2,\ \forall i \in \{1, \cdots, N-1\}$. Dropping the rank-one constraint, and relaxing the constraints $\mathbf{X}(i,i) \in \mathcal{A}^2,\ \forall i \in \{1, \cdots, N-1\}$ to the convex half-space constraints $L := \min_{a \in \mathcal{A}} a^2 \leq \mathbf{X}(i,i) \leq \max_{a \in \mathcal{A}} a^2 =: U,\ \forall i \in \{1, \cdots, N-1\}$, we obtain the following convex relaxation:

$$\min \text{Trace}(\mathbf{Q} \mathbf{X}) \tag{19}$$

$$\text{subject to: } \mathbf{X} \geq \mathbf{0}, \tag{20}$$

$$L \leq \mathbf{X}(i,i) \leq U,\ \forall i \in \{1, \cdots, N-1\},\ \mathbf{X}(N,N) = 1. \tag{21}$$

Note that (19)-(21) is not a Lagrangian relaxation of (16)-(18), because, in addition to the rank-one constraint, we have relaxed the alphabet constraints. This means that the bi-dual interpretation does not hold for our relaxation in (19)-(21). For a bi-dual relaxation see [13]. Our proposed relaxation in (19)-(21) can be viewed as further relaxation of [13], and it affords lower complexity for large $|\mathcal{A}|$ compared to [13].

The relaxed problem in (19)-(21) can be solved using any of the available modern SDP solvers, such as SeDuMi [10], based on interior point methods. After this step, an approximate solution to the original problem can be generated using *Gaussian randomization*: that is, drawing random vectors $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{X}_o)$, where $\mathbf{X}_o$ denotes the solution of (19)-(21), quantizing each element of $\mathbf{x}$ to the nearest point in $\mathcal{A}$, reconstructing $\mathbf{s}$ from the quantized $\mathbf{x}$, and picking the $\mathbf{s}$ that yields the smallest cost in (3).

A. Complexity

The worst-case complexity of solving a generic SDP problem involving a matrix variable of size $N \times N$ and $O(N)$ linear constraints is $O(N^{6.5})$. That would imply a complexity of $O(N^{6.5})$ for problem (19)-(21). However, exploiting the fact that the constraints in (21) are separable and only apply to the diagonal elements of $\mathbf{X}$, that figure can be reduced to $O(N^{3.5})$, which is very competitive ($N = 2M + 1$, where M is the number of QAM symbols). The complexity of the randomization step is $O(N^2)$ per draw. We emphasize that, unlike [13], the complexity of the overall algorithm is independent of the constellation order for uniform QAM, and affine in the constellation order for non-uniform QAM. This is because the quantization step in the randomization loop amounts to simple scaling and rounding for uniform constellations, but may require a linear search for non-uniform constellations.

IV. Simulations

We conducted Monte-Carlo (MC) simulation experiments for two indicative MIMO transmission scenarios: a 16×16 system using 64-QAM, and an 8×8 system using 16-QAM. In both cases, the channel matrix comprised i.i.d. elements drawn from a circularly symmetric zero-mean complex normal distribution of unit variance ($\mathcal{CN}(0, 1)$), and a new channel realization was drawn for each vector transmission (MC trial). The signal to noise ratio is defined as $SNR := 10log_{10}\frac{ME_s}{N_o}$, where M is the length of the transmitted QAM symbol vector $\mathbf{s}$, E_s is the mean symbol energy of the QAM constellation, and the noise vector is i.i.d. $\mathcal{CN}(0, N_o)$.

In order to gauge performance as a function of the number of randomizations, we tested our SDR algorithm with 100, 300, and 1000 randomization samples per decoded vector. As baselines for comparison, we employed i) the Schnor-Euchner variant of SD (SE-SD) with an infinite radius so that the optimal solution is always obtained; and ii) two commonly used suboptimal solutions of complexity $O(M^3)$: the quantized output of the zero-forcing linear receiver (QZF), and the (nonlinear) block MMSE-DFE (BMMSE-DFE) [3], [9]. Two performance metrics were used: Symbol Error Rate (SER), and worst-case execution[2] time. SE-SD was implemented as a Matlab executable (mex) compiled from optimized C

[2]On an Intel Centrino 1.6GHz system, with 512M RAM.

code; SDR was implemented using the general-purpose SeDuMi toolbox [10]. As a result, execution time estimates are somewhat biased in favor of SE-SD. The reason for using a measure of worst-case (as opposed to average) complexity is that in on-line applications we have to decode within a specified time, and bad channels do happen with positive probability. The choice between execution time or number of floating point operations is debatable, especially because SE-SD was implemented in mex/C; but we are interested in order-of-magnitude estimates, and differences in execution time are easier to appreciate.

Figures 1 and 2 show the SER versus SNR and worst-case execution time versus SNR, respectively, for the 16×16 system using 64-QAM ($64^{16} \approx 8 \times 10^{28}$). From figure 2, it is evident that SE-SD is too complex for this configuration; very long runs are actually not atypical. Due to this, figure 2 actually shows a *lower bound* on the worst-case execution time of SE-SD, computed from far fewer realizations. The associated SER cannot be estimated in reasonable time, and is therefore not reported in figure 1. SDR provides a performance improvement of up to 7.5 dB over BMMSE-DFE. Note that the worst-case complexity of SDR is essentially independent of SNR. In fact the point-wise complexity of SDR is very stable and predictable for any problem realization. This is good at low to moderate SNR, but a drawback at high SNR where the detection problem becomes easier. Also note that the number of randomization samples used in SDR does not affect the *grosso modo* complexity order, as expected; and a moderate number of randomizations is sufficient.

Figures 3 and 4 show corresponding results for the 8×8 system using 16-QAM ($16^8 \approx 4.3 \times 10^9$). Notice that, in this (far) simpler scenario, SE-SD is much more efficient computationally than SDR, and it always yields the exact ML solution. SDR is up to 7.5 dB away from SE-SD, at a uniformly higher computational cost across the range of SNR of interest. It clearly makes no sense to use SDR in this case.

Summarizing, the SD family of detectors exhibits a threshold behavior: it either works very well (for low-enough symbol vector dimension, order of the individual symbol constellation, and high-enough SNR) or it "freezes". The threshold between the two regimes depends on a combination of these three factors. When SD works, it outperforms SDR in terms of complexity and SER performance. In difficult scenarios, SDR offers an attractive alternative relative to earlier solutions.

V. CONCLUSIONS

We have proposed a new SDR approach for MIMO detection of high-order QAM constellations. The new approach is the simplest one in the class of SDR detectors for high-order QAM: its worst-case complexity is nearly cubic in the dimension of the transmitted symbol vector, and independent of the

constellation order for uniform QAM / affine in the constellation order for non-uniform QAM. Under certain conditions, the new approach affords significant improvements in SER over prior methods.

REFERENCES

[1] E. Agrell, T. Eriksson, A. Vardy, and K. Zeger, "Closest Point Search in Lattices," *IEEE Trans. Information Theory*, vol. 48, pp. 2201-2214, Aug. 2002.

[2] A. Chan and I. Lee, "A New Reduced-Complexity Sphere Decoder for Multiple Antenna Systems," in *Proc. of ICC 2002*, vol. 1, pp. 460-464, New York City, N.Y., April 28 - May 2, 2002.

[3] A. Duel-Hallen, "A family of multiuser decision-feedback detectors for asynchronous code-division multiple-access channels," *IEEE Trans. on Communications*, vol. 43, issue 234, pp. 421-434, Feb. 1995.

[4] J. Jaldén, B. Ottersten, "An Exponential Lower Bound on the Expected Complexity of Sphere Decoding," in *Proc. ICASSP 2004*, May 17-21, Montreal, Quebec, Canada.

[5] Z. Q. Luo, X. Luo, and M. Kisialiou, "An Efficient Quasi-Maximum Likelihood Decoder for PSK Signals," in *Proc. ICASSP2003*.

[6] W. K. Ma, P. C. Ching, and Z. Ding, "Semidefinite relaxation based multiuser detection for M-ary PSK multiuser systems," IEEE Trans. Signal Processing, vol. 52, no. 10, pp. 28622872, Oct. 2004.

[7] W.-K. Ma, T.N. Davidson, K.M. Wong, Z-Q Luo, P.-C. Ching, "Quasi-ML Multiuser Detection Using Semi-Definite Relaxation with Application to Synchronous CDMA," *IEEE Trans. on Signal Processing*, vol. 50, no. 4, pp. 912 -922, Apr. 2002.

[8] A. Mobasher, M. Taherzadeh, R. Sotirov, A.K. Khandani, "A near maximum likelihood decoding algorithm for MIMO systems based on semi-definite programming," in *Proc. 2005 Int. Symp. on Information Theory (ISIT 2005)*, pp. 1686-1690, Sep. 4-9, 2005.

[9] A. Stamoulis, G. B. Giannakis, A. Scaglione, "Block FIR decision-feedback equalizers for filterbank precoded transmissions with blind channel estimation capabilities," *IEEE Trans. on Communications*, vol. 49, no. 1, pp. 69-83, Jan. 2001.

[10] J.F. Sturm, "Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones", *Optimization Methods and Software*, vol. 11-12, pp. 625–653, 1999. See also http://sedumi.mcmaster.ca

[11] E. Viterbo and J. Boutros, "A Universal Lattice Code Decoder for Fading Channels," *IEEE Trans. Information Theory*, vol. 45, pp. 1639-1642, July 1999.

[12] R. Wang and G. B. Giannakis, "Approaching MIMO Capacity with Reduced-Complexity Soft Sphere-Decoding," in *Proc. WCNC 2004*, Atlanta, GA, March 21-25, 2004.

[13] A. Wiesel, Y. Eldar, and S. Shamai, "Semidefinite Relaxation for Detection of 16-QAM Signaling in MIMO Channels,", *IEEE Signal Processing Letters*, vol. 12, no. 9, pp. 653-656, Sep. 2005.

[14] W. Zhao, and G. B. Giannakis, "Sphere decoding algorithms with improved radius search," in *Proc. WCNC 2004*, Atlanta, GA, March 21-25, 2004.

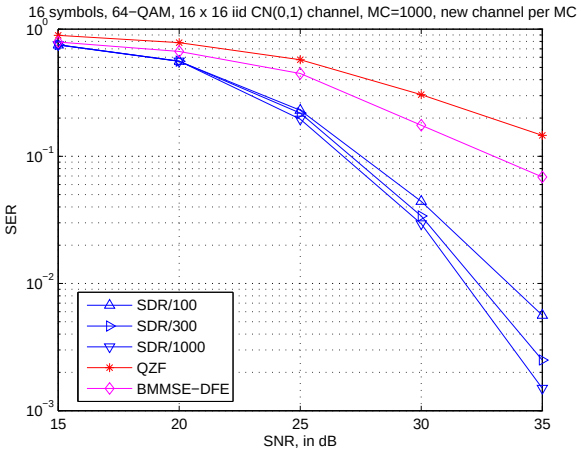


Fig. 1. SER versus SNR: 16×16 system, 64-QAM symbols.

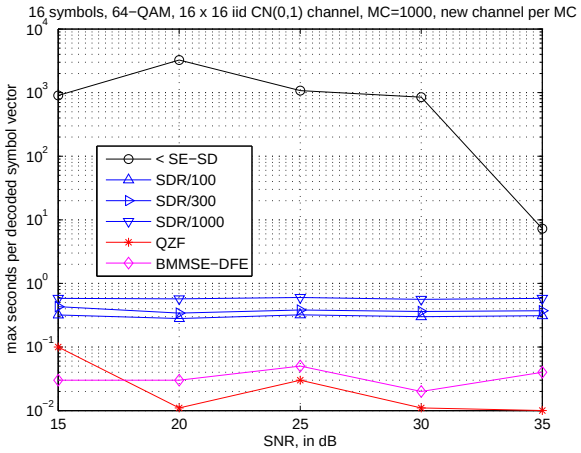


Fig. 2. Worst-case execution time versus SNR: 16×16 system, 64-QAM symbols.

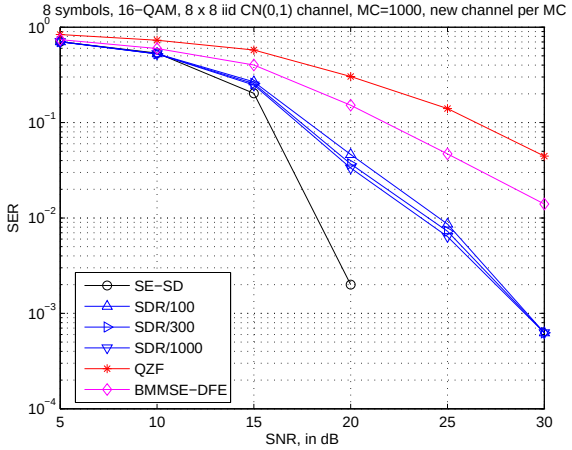


Fig. 3. SER versus SNR: 8×8 system, 16-QAM symbols.

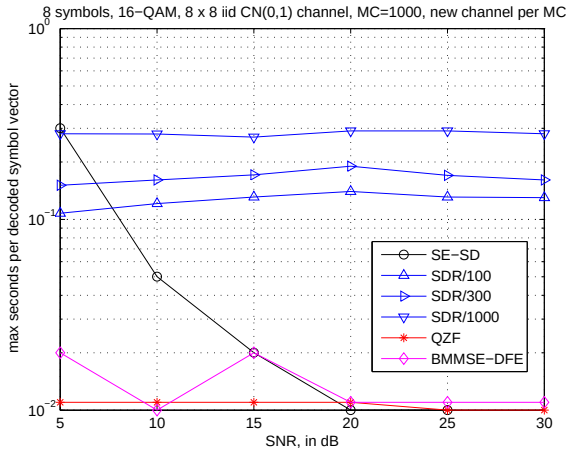


Fig. 4. Worst-case execution time versus SNR: 8×8 system, 16-QAM symbols.

APPROXIMATION BOUNDS FOR QUADRATIC OPTIMIZATION WITH HOMOGENEOUS QUADRATIC CONSTRAINTS

ZHI-QUAN LUO*, NICHOLAS D. SIDIROPOULOS†, PAUL TSENG‡, AND SHUZHONG ZHANG§

Abstract. We consider the NP-hard problem of finding a minimum norm vector in n-dimensional real or complex Euclidean space, subject to m concave homogeneous quadratic constraints. We show that a semidefinite programming (SDP) relaxation for this nonconvex quadratically constrained quadratic program (QP) provides an $O(m^2)$ approximation in the real case, and an $O(m)$ approximation in the complex case. Moreover, we show that these bounds are tight up to a constant factor. When the Hessian of each constraint function is of rank 1 (namely, outer products of some given so-called *steering* vectors) and the phase spread of the entries of these steering vectors are bounded away from $\pi/2$, we establish a certain "constant factor" approximation (depending on the phase spread but independent of m and n) for both the SDP relaxation and a convex QP restriction of the original NP-hard problem. Finally, we consider a related problem of finding a maximum norm vector subject to m convex homogeneous quadratic constraints. We show that a SDP relaxation for this nonconvex QP provides an $O(1/\ln(m))$ approximation, which is analogous to a result of Nemirovski, Roos and Terlaky [14] for the real case.

Key words. semidefinite programming relaxation, nonconvex quadratic optimization, approximation bound

AMS subject classifications. 90C22, 90C20, 90C59

1. Introduction.

Consider the quadratic optimization problem with concave homogeneous quadratic constraints:

$$\begin{array}{rll} v_{\rm qp} := \min & \|z\|^2 & \\ \text{s.t.} & \sum\limits_{\ell\in\mathcal{I}_i} |h_\ell^H z|^2 \geq 1, & i=1,...,m, \\ & z\in \mathbb{F}^{\rm n}, & \end{array} \tag{1.1}$$

where $\mathbb{F}$ is either $\mathbb{R}$ or $\mathbb{C}$, $\|\cdot\|$ denotes the Euclidean norm in $\mathbb{F}^{\rm n}$, $m\geq 1$, each h_ℓ is a given vector in $\mathbb{F}^{\rm n}$, and $\mathcal{I}_1,...,\mathcal{I}_m$ are nonempty, mutually disjoint index sets satisfying $\mathcal{I}_1\cup\cdots\cup\mathcal{I}_m=\{1,...,M\}$. Throughout, the superscript "H" will denote the complex Hermitian transpose, i.e., for $z=x+\mathbf{i}y$, where $x,y\in\mathbb{R}^{\rm n}$ and $\mathbf{i}^2=-1$, $z^H=x^T-\mathbf{i}y^T$. Geometrically, the above problem (1.1) corresponds

*Department of Electrical and Computer Engineering, University of Minnesota, 200 Union Street SE, Minneapolis, MN 55455 (luozq@ece.umn.edu). The work of this author is supported in part by the National Science Foundation, Grant No. DMS-0312416.

† Department of Electronic and Computer Engineering, Technical University of Crete, 73100 Chania - Crete, Greece.(nikos@telecom.tuc.gr). The work of this author is supported in part by the U.S. ARO under ERO, Contract No. N62558-03-C-0012, and the EU under U-BROAD STREP, Grant No. 506790.

‡Department of Mathematics, University of Washington, Seattle, Washington 98195 (tseng@math.washington.edu). The work of this author is supported by the National Science Foundation, Grant No. DMS-0511283.

§Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, Hong Kong. (zhang@se.cuhk.edu.hk). The work of this author is supported by Hong Kong RGC Earmarked Grant CUHK418505.

to finding a least norm vector in a region defined by the intersection of the exteriors of m co-centered ellipsoids. If the vectors $h_1, ..., h_M$ are linearly independent, then M equals the sum of the rank of the matrices defining these m ellipsoids. Notice that the problem (1.1) is easily solved for the case of $n = 1$, so we assume $n \geq 2$.

We assume that $\sum_{\ell \in \mathcal{I}_i} \|h_\ell\| \neq 0$ for all i, which is clearly a necessary condition for (1.1) to be feasible. This is also a sufficient condition (since $\bigcup_{i=1}^m \{z \mid \sum_{\ell \in \mathcal{I}_i} |h_\ell^H z|^2 = 0\}$ is a finite union of proper subspaces of $\mathbb{F}^n$, so its complement is nonempty and any point in its complement can be scaled to be feasible for (1.1)). Thus, the above problem (1.1) always has an optimal solution (not necessarily unique) since its objective function is coercive, continuous, and its feasible set is nonempty, closed. Notice, however, that the feasible set of (1.1) is typically nonconvex and disconnected, with an exponential number of connected components exhibiting little symmetry. This is in contrast to the quadratic problems with convex feasible set but nonconvex objective function considered in [13, 14, 22]. Furthermore, unlike the class of quadratic problems studied in [1, 7, 8, 15, 16, 21, 23, 24, 25, 26], the constraint functions in (1.1) do not depend on $z_1^2, ..., z_n^2$ only.

Our interest in the nonconvex QP (1.1) is motivated by the transmit beamforming problem for multicasting applications [20] and by the wireless sensor network localization problem [6]. In the transmit beamforming problem, a transmitter utilizes an array of n transmitting antennas to broadcast information within its service area to m radio receivers, with receiver $i \in \{1, ..., m\}$ equipped with $|\mathcal{I}_i|$ receiving antennas. Let h_ℓ, $\ell \in \mathcal{I}_i$, denote the $n \times 1$ complex *steering vector* modelling propagation loss and phase shift from the transmitting antennas to the ℓth receiving antenna of receiver i. Assuming that each receiver performs spatially matched filtering / maximum ratio combining, which is the optimal combining strategy under standard mild assumptions, then the constraint

$$\sum_{\ell \in \mathcal{I}_i} |h_\ell^H z|^2 \geq 1$$

models the requirement that the total received signal power at receiver i must be above a given threshold (normalized to 1). This constraint is also equivalent to a signal-to-noise ratio (SNR) condition commonly used in data communication. Thus, to minimize the total transmit power subject to individual SNR requirements (one at each receiver), we are led to the QP (1.1). In the special case where each radio receiver is equipped with a single receiving antenna, the problem reduces to [20]:

$$\begin{array}{ll} \min & \|z\|^2 \\ \text{s.t.} & |h_\ell^H z|^2 \geq 1, \quad \ell = 1, ..., m, \\ & z \in \mathbb{F}^n, \end{array} \tag{1.2}$$

This problem is a special case of (1.1) whereby each ellipsoid lies in $\mathbb{F}^n$ and the corresponding matrix has rank 1.

In this paper, we first show that the nonconvex QP (1.2) is NP-hard in either the real or the complex case, which further implies the NP-hardness of the general problem (1.1). Then, we consider a semidefinite programming (SDP) *relaxation* of (1.1) and a convex QP *restriction* of (1.2)

and study their worst-case performance. In particular, let $\upsilon_{\rm sdp}$, $\upsilon_{\rm cqp}$ and $\upsilon_{\rm qp}$ denote the optimal values of the SDP relaxation, the convex QP restriction, and the original QP (1.1), respectively. We establish a performance ratio of $\upsilon_{\rm qp}/\upsilon_{\rm sdp} = O(m^2)$ for the SDP relaxation in the real case, and we give an example showing that this bound is tight up to a constant factor. Similarly, we establish a performance ratio of $\upsilon_{\rm qp}/\upsilon_{\rm sdp} = O(m)$ in the complex case, and we give an example showing the tightness of this bound. We further show that, in the case when the phase spread of the entries of $h_1, ..., h_M$ is bounded away from $\pi/2$, the performance ratios $\upsilon_{\rm qp}/\upsilon_{\rm sdp}$ and $\upsilon_{\rm cqp}/\upsilon_{\rm qp}$ for the SDP relaxation and the convex QP restriction, respectively, are independent of m and n.

In recent years, there have been extensive studies of the performance of SDP relaxations for nonconvex QP. However, to our knowledge, this is the first performance analysis of SDP relaxation for QP with concave quadratic constraints. Our proof techniques also extend to a maximization version of the QP (1.1) with convex homogeneous quadratic constraints. In particular, we give a simple proof of a result analogous to one of Nemirovski, Roos and Terlaky [14] (also see [13, Theorem 4.7]) for the real case, namely, the SDP relaxation for this nonconvex QP has a performance ratio of $O(1/\ln(m))$.

2. NP-hardness. In this section, we show that the nonconvex QP (1.1) is NP-hard in general. First, we notice that, by a linear transformation if necessary, the following problem

$$\begin{array}{ll} \text{minimize} & z^H Q z \\ \text{subject to} & |z_\ell| \geq 1, \quad \ell = 1, ..., n, \\ & z \in \mathbb{F}^{\rm n}, \end{array} \tag{2.1}$$

is a special case of (1.1), where $Q \in \mathbb{F}^{\rm n\times n}$ is a Hermitian positive definite matrix (i.e., $Q \succ 0$), and z_ℓ denotes the ℓth component of z. Hence, it suffices to establish the NP-hardness of (2.1). To this end, we consider a reduction from the NP-complete partition problem: Given positive integers $a_1,\ a_2,\ ...,\ a_N$, decide whether there exists a subset $\mathcal{I}$ of $\{1, ..., N\}$ satisfying

$$\sum_{\ell\in\mathcal{I}} a_\ell = \frac{1}{2}\sum_{\ell=1}^{N} a_\ell. \tag{2.2}$$

Our reductions differ for the real and complex cases. As will be seen, the NP-hardness proof in the complex case[1] is more intricate than in the real case.

2.1. The Real Case. We consider the real case of $\mathbb{F} = \mathbb{R}$. Let $n := N$ and

$$\begin{aligned} a &:= (a_1, \ldots, a_N)^T, \\ Q &:= aa^T + I_n \ \succ \ 0, \end{aligned}$$

where I_n denotes the $n \times n$ identity matrix.

[1] This NP-hardness proof was first presented in an appendix of [20] and is included here for completeness; also see [26, Proposition 3.5] for a related proof.

We show that a subset $\mathcal{I}$ satisfying (2.2) exists if and only if the optimization problem (2.1) has a minimum value of n. Since

$$z^T Q z = |a^T z|^2 + \sum_{\ell=1}^{n} |z_\ell|^2 \geq n \quad \text{whenever } |z_\ell| \geq 1 \ \forall\, \ell,\ z \in \mathbb{R}^n,$$

we see that (2.1) has a minimum value of n if and only if there exists a $z \in \mathbb{R}^n$ satisfying

$$a^T z = 0, \quad |z_\ell| = 1 \ \ \forall\, \ell.$$

The above condition is equivalent to the existence of a subset $\mathcal{I}$ satisfying (2.2), with the correspondence $\mathcal{I} = \{\ell \mid z_\ell = 1\}$. This completes the proof.

2.2. The Complex Case. We consider the complex case of $\mathbb{F} = \mathbb{C}$. Let $n := 2N + 1$ and

$$\begin{aligned} a &:= (a_1, \ldots, a_N)^T, \\ A &:= \begin{pmatrix} I_N & I_N & -e_N \\ a^T & 0_N^T & -\frac{1}{2} a^T e_N \end{pmatrix}, \\ Q &:= A^T A + I_n \ \succ \ 0, \end{aligned}$$

where e_N denotes the N-dimensional vector of ones, 0_N denotes the N-dimensional vector of zeros, and I_n and I_N are identity matrices of sizes $n \times n$ and $N \times N$, respectively.

We show that a subset $\mathcal{I}$ satisfying (2.2) exists if and only if the optimization problem (2.1) has a minimum value of n. Since

$$z^H Q z = \|Az\|^2 + \sum_{\ell=1}^{n} |z_\ell|^2 \geq n \quad \text{whenever } |z_\ell| \geq 1 \ \forall\, \ell,\ z \in \mathbb{C}^n,$$

we see that (2.1) has a minimum value of n if and only if there exists a $z \in \mathbb{C}^n$ satisfying

$$Az = 0, \quad |z_\ell| = 1 \ \ \forall\, \ell.$$

Expanding $Az = 0$ gives the following set of linear equations:

$$0 = z_\ell + z_{N+\ell} - z_n, \quad \ell = 1, ..., N, \tag{2.3}$$

$$0 = \sum_{\ell=1}^{N} a_\ell z_\ell - \frac{1}{2} \left(\sum_{\ell=1}^{N} a_\ell \right) z_n. \tag{2.4}$$

For $\ell = 1, ..., 2N$, since $|z_\ell| = |z_n| = 1$ so that $z_\ell / z_n = e^{\mathbf{i}\theta_\ell}$ for some $\theta_\ell \in [0, 2\pi)$, we can rewrite (2.3) as

$$\begin{array}{rcl} \cos\theta_\ell + \cos\theta_{N+\ell} & = & 1, \\ \sin\theta_\ell + \sin\theta_{N+\ell} & = & 0, \end{array} \quad \ell = 1, ..., N.$$

These equations imply that $\theta_\ell \in \{-\pi/3, \pi/3\}$ for all $\ell \neq n$. In fact, these equations further imply that $\cos\theta_\ell = \cos\theta_{N+\ell} = 1/2$ for $\ell = 1, ..., N$, so that

$$\mathrm{Re} \left(\sum_{\ell=1}^{N} a_\ell \frac{z_\ell}{z_n} - \frac{1}{2} \left(\sum_{\ell=1}^{N} a_\ell \right) \right) = 0.$$

Therefore, (2.4) is satisfied if and only if

$$\mathrm{Im}\left(\sum_{\ell=1}^{N} a_\ell \frac{z_\ell}{z_n} - \frac{1}{2}\left(\sum_{\ell=1}^{N} a_\ell\right)\right) = \mathrm{Im}\left(\sum_{\ell=1}^{N} a_\ell \frac{z_\ell}{z_n}\right) = 0,$$

which is further equivalent to the existence of a subset $\mathcal{I}$ satisfying (2.2), with the correspondence $\mathcal{I} = \{\ell \mid \theta_\ell = \pi/3\}$. This completes the proof.

3. Performance analysis of SDP relaxation. In this section, we study the performance of an SDP relaxation of (1.2). Let

$$H_i := \sum_{\ell \in \mathcal{I}_i} h_\ell h_\ell^H, \quad i = 1, ..., m.$$

The well-known SDP relaxation of (1.1) [11, 19] is

$$\begin{aligned} \upsilon_{\text{sdp}} \;:=\; & \min \quad \mathrm{Tr}(Z) \\ & \text{s.t.} \quad \mathrm{Tr}(H_i Z) \geq 1, \quad i = 1, ..., m, \\ & \quad\; Z \succeq 0, \quad Z \in \mathbb{F}^{\mathrm{n}\times\mathrm{n}} \text{ is Hermitian.} \end{aligned} \tag{3.1}$$

An optimal solution of the SDP relaxation (3.1) can be computed efficiently using, say, interior-point methods; see [18] and references therein.

Clearly $\upsilon_{\text{sdp}} \leq \upsilon_{\text{qp}}$. We are interested in upper bounds for the relaxation performance of the form

$$\upsilon_{\text{qp}} \leq C \upsilon_{\text{sdp}},$$

where $C \geq 1$. Since we assume $H_i \neq 0$ for all i, it is easily checked that (3.1) has an optimal solution, which we denote by Z^*.

3.1. General steering vectors: the real case. We consider the real case of $\mathbb{F} = \mathbb{R}$. Upon obtaining an optimal solution Z^* of (3.1), we construct a feasible solution of (1.1) using the following randomization procedure:

1. Generate a random vector $\xi \in \mathbb{R}^{\mathrm{n}}$ from the real-valued normal distribution $N(0, Z^*)$.
2. Let $z^*(\xi) = \xi / \min\limits_{1\leq i\leq m} \sqrt{\xi^T H_i \xi}$.

We will use $z^*(\xi)$ to analyze the performance of the SDP relaxation. Similar procedures have been used for related problems [1, 3, 4, 5, 14]. First, we need to develop two lemmas. The first lemma estimates the left-tail of the distribution of a convex quadratic form of a Gaussian random vector.

LEMMA 3.1. *Let $H \in \mathbb{R}^{\mathrm{n}\times\mathrm{n}}$, $Z \in \mathbb{R}^{\mathrm{n}\times\mathrm{n}}$ be two symmetric positive semidefinite matrices (i.e., $H \succeq 0$, $Z \succeq 0$). Suppose $\xi \in \mathbb{R}^{\mathrm{n}}$ is a random vector generated from the real-valued normal distribution $N(0, Z)$. Then, for any $\gamma > 0$,*

$$\mathrm{Prob}\left(\xi^T H \xi < \gamma E(\xi^T H \xi)\right) \leq \max\left\{\sqrt{\gamma}, \frac{2(\bar{r}-1)\gamma}{\pi - 2}\right\}, \tag{3.2}$$

where $\bar{r} := \min\{\mathrm{rank}(H), \mathrm{rank}(Z)\}$.

Proof. Since the covariance matrix $Z \succeq 0$ has rank $r := \mathrm{rank}(Z)$, we can write $Z = UU^T$, for some $U \in \mathbb{R}^{n\times r}$ satisfying $U^T Z U = I_r$. Let $\bar{\xi} := Q^T U^T \xi \in \mathbb{R}^r$, where $Q \in \mathbb{R}^{r\times r}$ is an orthogonal matrix corresponding to the eigen-decomposition of the matrix

$$U^T H U = Q \Lambda Q^T,$$

for some diagonal matrix $\Lambda = \mathrm{diag}\{\lambda_1, \lambda_2, ..., \lambda_r\}$, with $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_r \geq 0$. Since $U^T H U$ has rank at most $\bar{r}$, we have $\lambda_i = 0$ for all $i > \bar{r}$. It is readily checked that $\bar{\xi}$ has the normal distribution $N(0, I_r)$. Moreover, ξ is statistically identical to $UQ\bar{\xi}$, so that $\xi^T H \xi$ is statistically identical to

$$\bar{\xi}^T Q^T U^T H U Q \bar{\xi} = \bar{\xi}^T \Lambda \bar{\xi} = \sum_{i=1}^{\bar{r}} \lambda_i |\bar{\xi}_i|^2.$$

Then, we have

$$\begin{aligned}\mathrm{Prob}\left(\xi^T H \xi < \gamma E(\xi^T H \xi)\right) &= \mathrm{Prob}\left(\sum_{i=1}^{\bar{r}} \lambda_i |\bar{\xi}_i|^2 < \gamma E\left(\sum_{i=1}^{\bar{r}} \lambda_i |\bar{\xi}_i|^2\right)\right) \\ &= \mathrm{Prob}\left(\sum_{i=1}^{\bar{r}} \lambda_i |\bar{\xi}_i|^2 < \gamma \sum_{i=1}^{\bar{r}} \lambda_i\right).\end{aligned}$$

If $\lambda_1 = 0$, then this probability is zero, which proves (3.2). Thus, we will assume that $\lambda_1 > 0$. Let $\bar{\lambda}_i := \lambda_i/(\lambda_1 + \cdots + \lambda_{\bar{r}})$, for $i = 1, ..., \bar{r}$. Clearly, we have

$$\bar{\lambda}_1 + \cdots + \bar{\lambda}_{\bar{r}} = 1, \quad \bar{\lambda}_1 \geq \bar{\lambda}_2 \geq \ldots \geq \bar{\lambda}_{\bar{r}} \geq 0.$$

We consider two cases. First, suppose $\bar{\lambda}_1 \geq \alpha$, where $0 < \alpha < 1$. Then, we can bound the above probability as follows:

$$\begin{aligned}\mathrm{Prob}\left(\xi^T H \xi < \gamma E(\xi^T H \xi)\right) &= \mathrm{Prob}\left(\sum_{i=1}^{\bar{r}} \bar{\lambda}_i |\bar{\xi}_i|^2 < \gamma\right) \\ &\leq \mathrm{Prob}\left(\bar{\lambda}_1 |\bar{\xi}_1|^2 < \gamma\right) \\ &\leq \mathrm{Prob}\left(|\bar{\xi}_1|^2 < \gamma/\alpha\right) \\ &\leq \sqrt{\frac{2\gamma}{\pi\alpha}},\end{aligned} \tag{3.3}$$

where the last step is due to the fact that $\bar{\xi}_1$ is a real-valued zero mean Gaussian random variable with unit variance.

In the second case, we have $\bar{\lambda}_1 < \alpha$, so that

$$\bar{\lambda}_2 + \cdots + \bar{\lambda}_{\bar{r}} = 1 - \bar{\lambda}_1 > 1 - \alpha.$$

This further implies $(\bar{r} - 1)\bar{\lambda}_2 \geq \bar{\lambda}_2 + \cdots + \bar{\lambda}_{\bar{r}} > 1 - \alpha$. Hence

$$\bar{\lambda}_1 \geq \bar{\lambda}_2 > \frac{1-\alpha}{\bar{r}-1}.$$

Using this bound, we obtain the following probability estimate:

$$\begin{aligned}\text{Prob}\left(\xi^T H\xi < \gamma E(\xi^T H\xi)\right) &= \text{Prob}\left(\sum_{i=1}^{\bar{r}} \bar{\lambda}_i |\bar{\xi}_i|^2 < \gamma\right)\\ &\le \text{Prob}\left(\bar{\lambda}_1|\bar{\xi}_1|^2 < \gamma,\ \bar{\lambda}_2|\bar{\xi}_2|^2 < \gamma\right)\\ &= \text{Prob}\left(\bar{\lambda}_1|\bar{\xi}_1|^2 < \gamma\right)\cdot \text{Prob}\left(\bar{\lambda}_2|\bar{\xi}_2|^2 < \gamma\right)\\ &\le \sqrt{\frac{2\gamma}{\pi\bar{\lambda}_1}}\cdot\sqrt{\frac{2\gamma}{\pi\bar{\lambda}_2}}\\ &\le \frac{2(\bar{r}-1)\gamma}{\pi(1-\alpha)}.\end{aligned} \tag{3.4}$$

Combining the estimates for the above two cases and setting $\alpha = 2/\pi$, we immediately obtain the desired bound (3.2). ∎

LEMMA 3.2. *Let $\mathbb{F} = \mathbb{R}$. Let $Z^* \succeq 0$ be a feasible solution of (3.1) and let $z^*(\xi)$ be generated by the randomization procedure described earlier. Then, with probability 1, $z^*(\xi)$ is well defined and feasible for (1.1). Moreover, for every $\gamma > 0$ and $\mu > 0$,*

$$\text{Prob}\left(\min_{1\le i\le m} \xi^T H_i\xi \ge \gamma,\ \|\xi\|^2 \le \mu\text{Tr}(Z^*)\right) \ge 1 - m\cdot\max\left\{\sqrt{\gamma}, \frac{2(r-1)\gamma}{\pi-2}\right\} - \frac{1}{\mu}, \tag{3.5}$$

where $r := \text{rank}(Z^*)$.

Proof. Since $Z^* \succeq 0$ is feasible for (3.1), it follows that $\text{Tr}(H_i Z^*) \ge 1$ for all $i = 1, ..., m$. Since $E(\xi^T H_i\xi) = \text{Tr}(H_i Z^*) \ge 1$ and the density of $\xi^T H_i\xi$ is absolutely continuous, the probability of $\xi^T H_i\xi = 0$ is zero, implying that $z^*(\xi)$ is well defined with probability 1. The feasibility of $z^*(\xi)$ is easily verified.

To prove (3.5), we first note that $E(\xi\xi^T) = Z^*$. Thus, for any $\gamma > 0$ and $\mu > 0$,

$$\begin{aligned}&\text{Prob}\left(\min_{1\le i\le m} \xi^T H_i\xi \ge \gamma,\ \|\xi\|^2 \le \mu\text{Tr}(Z^*)\right)\\ &= \text{Prob}\left(\xi^T H_i\xi \ge \gamma\ \forall\ i = 1, ..., m \ \text{ and } \ \|\xi\|^2 \le \mu\text{Tr}(Z^*)\right)\\ &\ge \text{Prob}\left(\xi^T H_i\xi \ge \gamma\text{Tr}(H_i Z^*)\ \forall\ i = 1, ..., m \ \text{ and } \ \|\xi\|^2 \le \mu\text{Tr}(Z^*)\right)\\ &= \text{Prob}\left(\xi^T H_i\xi \ge \gamma E(\xi^T H_i\xi)\ \forall\ i = 1, ..., m \ \text{ and } \ \|\xi\|^2 \le \mu E(\|\xi\|^2)\right)\\ &= 1 - \text{Prob}\left(\xi^T H_i\xi < \gamma E(\xi^T H_i\xi) \text{ for some } i \text{ or } \|\xi\|^2 > \mu E(\|\xi\|^2)\right)\\ &\ge 1 - \sum_{i=1}^{m}\text{Prob}\left(\xi^T H_i\xi < \gamma E(\xi^T H_i\xi)\right) - \text{Prob}\left(\|\xi\|^2 > \mu E(\|\xi\|^2)\right)\\ &> 1 - m\cdot\max\left\{\sqrt{\gamma}, \frac{2(r-1)\gamma}{\pi-2}\right\} - \frac{1}{\mu},\end{aligned}$$

where the last step uses Lemma 3.1 as well as Markov's inequality:

$$\text{Prob}\left(\|\xi\|^2 > \mu E(\|\xi\|^2)\right) \le \frac{1}{\mu}.$$

This completes the proof. ■

We now use Lemma 3.2 to bound the performance of the SDP relaxation.

THEOREM 3.3. *Let* $\mathbb{F} = \mathbb{R}$. *For the QP* (1.1) *and its SDP relaxation* (3.1), *we have* $v_{\rm qp} = v_{\rm sdp}$ *if* $m \leq 2$, *and otherwise*

$$v_{\rm qp} \leq \frac{27m^2}{\pi} v_{\rm sdp}.$$

Proof. By applying a suitable rank reduction procedure if necessary, we can assume that the rank r of the optimal SDP solution Z^* satisfies $r(r+1)/2 \leq m$; see e.g. [17]. Thus $r < \sqrt{2m}$. If $m \leq 2$, then $r = 1$, implying that $Z^* = z^*(z^*)^T$ for some $z^* \in \mathbb{R}^{\rm n}$ and it is readily seen that z^* is an optimal solution of (1.1), so that $v_{\rm qp} = v_{\rm sdp}$. Otherwise, we apply the randomization procedure to Z^*. We also choose

$$\mu = 3, \quad \gamma = \frac{\pi}{4m^2}\left(1 - \frac{1}{\mu}\right)^2 = \frac{\pi}{9m^2}.$$

Then, it is easily verified using $r < \sqrt{2m}$ that

$$\sqrt{\gamma} \geq \frac{2(r-1)\gamma}{\pi - 2} \quad \forall\ m = 1, 2, ...$$

Plugging these choices of γ and μ into (3.5), we see that there is a positive probability (independent of problem size) of at least

$$1 - m\sqrt{\gamma} - \frac{1}{\mu} = 1 - \frac{\sqrt{\pi}}{3} - \frac{1}{3} = 0.0758...$$

that ξ generated by the randomization procedure satisfies

$$\min_{1 \leq i \leq m} \xi^T H_i \xi \geq \frac{\pi}{9m^2} \quad \text{and} \quad \|\xi\|^2 \leq 3\,{\rm Tr}(Z^*).$$

Let ξ be any vector satisfying these two conditions.[2] Then, $z^*(\xi)$ is feasible for (1.1), so that

$$v_{\rm qp} \leq \|z^*(\xi)\|^2 = \frac{\|\xi\|^2}{\min_i \xi^T H_i \xi} \leq \frac{3\,{\rm Tr}(Z^*)}{(\pi/9m^2)} = \frac{27m^2}{\pi}\, v_{\rm sdp},$$

where the last equality uses ${\rm Tr}(Z^*) = v_{\rm sdp}$. ■

In the above proof, other choices of μ can also be used, but the resulting bound seems not as sharp. Theorem 3.3 suggests that the worst-case performance of the SDP relaxation deteriorates

[2]The probability that no such ξ is generated after N independent trials is at most $(1 - 0.0758..)^N$, which for $N = 100$ equals 0.000375.. Thus, such ξ requires relatively few trials to generate.

quadratically with the number of quadratic constraints. Below we give an example demonstrating that this bound is in fact tight up to a constant factor.

EXAMPLE 1: For any $m \geq 2$ and $n \geq 2$, consider a special instance of (1.2), corresponding to (1.1) with $|\mathcal{I}_i| = 1$ (i.e., each H_i has rank 1), whereby

$$h_\ell = \left(\cos\left(\frac{\ell\pi}{m}\right), \sin\left(\frac{\ell\pi}{m}\right), 0, \ldots, 0\right)^T, \qquad \ell = 1, ..., m.$$

Let $z^* = (z_1^*, \ldots, z_n^*)^T \in \mathbb{R}^n$ be an optimal solution of (1.2) corresponding to the above choice of steering vectors h_ℓ. We can write

$$(z_1^*, z_2^*) = \rho(\cos\theta, \sin\theta), \quad \text{for some } \theta \in [0, 2\pi).$$

Since $\{\ell\pi/m,\ \ell = 1, ..., m\}$ is uniformly spaced on $[0, \pi)$, there must exist an integer ℓ such that

$$\text{either} \quad \left|\theta - \frac{\ell\pi}{m} - \frac{\pi}{2}\right| \leq \frac{\pi}{2m} \quad \text{or} \quad \left|\theta - \frac{\ell\pi}{m} + \frac{\pi}{2}\right| \leq \frac{\pi}{2m}.$$

For simplicity, we assume the first case. (The second case can be treated similarly.) Since the last $(n-2)$ entries of h_ℓ are zero, it is readily checked that

$$|h_\ell^T z^*| = \rho\left|\cos\left(\theta - \frac{\ell\pi}{m}\right)\right| = \rho\left|\sin\left(\theta - \frac{\ell\pi}{m} - \frac{\pi}{2}\right)\right| \leq \rho\left|\sin\left(\frac{\pi}{2m}\right)\right| \leq \frac{\rho\pi}{2m}.$$

Since z^* satisfies the constraint $|h_\ell^T z^*| \geq 1$, it follows that

$$\|z^*\| \geq \rho \geq \frac{2m|h_\ell^T z^*|}{\pi} \geq \frac{2m}{\pi},$$

implying

$$\upsilon_{\text{qp}} = \|z^*\|^2 \geq \frac{4m^2}{\pi^2}.$$

On the other hand, the positive semidefinite matrix

$$Z^* = \text{diag}\{1, 1, 0, \ldots, 0\}$$

is feasible for the SDP relaxation (3.1), and it has an objective value of $\text{Tr}(Z^*) = 2$. Thus, for this instance, we have

$$\upsilon_{\text{qp}} \geq \frac{2m^2}{\pi^2}\,\upsilon_{\text{sdp}}.$$

The preceding example and Theorem 3.3 show that the SDP relaxation (3.1) can be weak if the number of quadratic constraints is large, especially when the steering vectors h_ℓ are in a certain sense "uniformly distributed" in space.

3.2. General steering vectors: the complex case. We consider the complex case of $\mathbb{F} = \mathbb{C}$. We will show that the performance ratio of the SDP relaxation (3.1) improves to $O(m)$ in the complex case (as opposed to $O(m^2)$ in the real case). Similar to the real case, upon obtaining an optimal solution Z^* of (3.1), we construct a feasible solution of (1.1) using the following randomization procedure:

1. Generate a random vector $\xi \in \mathbb{C}^n$ from the *complex-valued* normal distribution $N_c(0, Z^*)$ [2, 26].
2. Let $z^*(\xi) = \xi / \min\limits_{1\le i\le m} \sqrt{\xi^H H_i \xi}$.

Most of the ensuing performance analysis is similar to that of the real case. In particular, we will also need the following two lemmas analogous to Lemmas 3.1 and 3.2.

LEMMA 3.4. *Let $H \in \mathbb{C}^{n\times n}$, $Z \in \mathbb{C}^{n\times n}$ be two Hermitian positive semidefinite matrices (i.e., $H \succeq 0$, $Z \succeq 0$). Suppose $\xi \in \mathbb{C}^n$ is a random vector generated from the complex-valued normal distribution $N_c(0, Z)$. Then, for any $\gamma > 0$,*

$$\text{Prob}\left(\xi^H H\xi < \gamma E(\xi^H H\xi)\right) \le \max\left\{\frac{4}{3}\gamma, 16(\bar{r}-1)^2\gamma^2\right\}, \tag{3.6}$$

where $\bar{r} := \min\{\text{rank}(H), \text{rank}(Z)\}$.

Proof. We follow the same notations and proof as for Lemma 3.1, except for two blanket changes:

$$\begin{array}{rcl} \text{matrix transpose} & \rightarrow & \text{Hermitian transpose,} \\ \text{orthogonal matrix} & \rightarrow & \text{unitary matrix.} \end{array}$$

Also, $\bar{\xi}$ has the complex-valued normal distribution $N_c(0, I_r)$. With these changes, we consider the same two cases: $\bar{\lambda}_1 \ge \alpha$ and $\bar{\lambda}_1 < \alpha$, where $0 < \alpha < 1$. In the first case, we have similar to (3.3) that

$$\text{Prob}\left(\xi^H H\xi < \gamma E(\xi^H H\xi)\right) \le \text{Prob}\left(|\bar{\xi}_1|^2 < \gamma/\alpha\right). \tag{3.7}$$

Recall that the density function of a complex-valued circular normal random variable $u \sim N_c(0, \sigma^2)$, where σ is the standard deviation, is

$$\frac{1}{\pi\sigma^2} e^{-\frac{|u|^2}{\sigma^2}} \quad \forall\ u \in \mathbb{C}.$$

In polar coordinates, the density function can be written as

$$f(\rho, \theta) = \frac{\rho}{\pi\sigma^2} e^{-\frac{\rho^2}{\sigma^2}} \quad \forall\ \rho \in [0, +\infty),\ \theta \in [0, 2\pi).$$

In fact, a complex-valued normal distribution can be viewed as a joint distribution of its modulus and its argument, with the following particular properties: (1) the modulus and argument are independently distributed; (2) the argument is uniformly distributed over $[0, 2\pi)$; (3) the modulus follows a Weibull distribution with density

$$f(\rho) = \begin{cases} \frac{2\rho}{\sigma^2} e^{-\frac{\rho^2}{\sigma^2}}, & \text{if } \rho \ge 0; \\ 0, & \text{if } \rho < 0, \end{cases}$$

and distribution function

$$\text{Prob}\{|u| \le t\} = 1 - e^{-\frac{t^2}{\sigma^2}}. \tag{3.8}$$

Since $\bar{\xi}_1 \sim N_c(0,1)$, substituting this into (3.7) yields

$$\text{Prob}\left(\xi^H H\xi < \gamma E(\xi^H H\xi)\right) \le \text{Prob}\left(|\bar{\xi}_1|^2 < \gamma/\alpha\right) \le 1 - e^{-\gamma/\alpha} \le \gamma/\alpha,$$

where the last inequality uses the convexity of the exponential function.

In the second case of $\bar{\lambda}_1 < \alpha$, we have similar to (3.4) that

$$\begin{aligned}
\text{Prob}\left(\xi^H H\xi < \gamma E(\xi^H H\xi)\right) &\le \text{Prob}\left(\bar{\lambda}_1|\bar{\xi}_1|^2 < \gamma\right) \cdot \text{Prob}\left(\bar{\lambda}_2|\bar{\xi}_2|^2 < \gamma\right) \\
&= (1 - e^{-\gamma/\bar{\lambda}_1})(1 - e^{-\gamma/\bar{\lambda}_2}) \\
&\le \frac{\gamma^2}{\bar{\lambda}_1\bar{\lambda}_2} \\
&\le \frac{(\bar{r}-1)^2\gamma^2}{(1-\alpha)^2},
\end{aligned}$$

where last step uses the fact that $\bar{\lambda}_1 \ge \bar{\lambda}_2 \ge (1-\alpha)/(\bar{r}-1)$. Combining the estimates for the above two cases and setting $\alpha = 3/4$, we immediately obtain the desired bound (3.6). ∎

LEMMA 3.5. *Let $\mathbb{F} = \mathbb{C}$. Let $Z^* \succeq 0$ be a feasible solution of* (3.1) *and let $z^*(\xi)$ be generated by the randomization procedure described earlier. Then, with probability 1, $z^*(\xi)$ is well defined and feasible for* (1.1). *Moreover, for every $\gamma > 0$ and $\mu > 0$,*

$$\text{Prob}\left(\min_{1\le i\le m} \xi^H H_i \xi \ge \gamma,\ \|\xi\|^2 \le \mu \text{Tr}(Z^*)\right) \ge 1 - m \cdot \max\left\{\frac{4}{3}\gamma, 16(r-1)^2\gamma^2\right\} - \frac{1}{\mu},$$

where $r := \text{rank}(Z^*)$.

Proof. The proof is mostly the same as that for the real case (see Lemma 3.2). In particular, for any $\gamma > 0$ and $\mu > 0$, we still have

$$\begin{aligned}
&\text{Prob}\left(\min_{1\le i\le m} \xi^H H_i \xi \ge \gamma,\ \|\xi\|^2 \le \mu \text{Tr}(Z^*)\right) \\
&\quad \ge 1 - \sum_{i=1}^{m} \text{Prob}\left(\xi^H H_i\xi < \gamma E(\xi^H H_i \xi)\right) - \text{Prob}\left(\|\xi\|^2 > \mu E(\|\xi\|^2)\right).
\end{aligned}$$

Therefore, we can invoke Lemma 3.4 to obtain

$$\begin{aligned}
&\text{Prob}\left(\min_{1\le i\le m} \xi^H H_i \xi \ge \gamma,\ \|\xi\|^2 \le \mu \text{Tr}(Z^*)\right) \\
&\quad \ge 1 - m \cdot \max\left\{\frac{4}{3}\gamma, 16(r-1)^2\gamma^2\right\} - \text{Prob}\left(\|\xi\|^2 > \mu E(\|\xi\|^2)\right) \\
&\quad \ge 1 - m \cdot \max\left\{\frac{4}{3}\gamma, 16(r-1)^2\gamma^2\right\} - \frac{1}{\mu},
\end{aligned}$$

which completes the proof. ■

THEOREM 3.6. *Let* $\mathbb{F} = \mathbb{C}$. *For the QP* (1.1) *and its SDP relaxation* (3.1), *we have* $v_{\rm sdp} = v_{\rm qp}$ *if* $m \le 3$ *and otherwise*

$$v_{\rm qp} \le 8m \cdot v_{\rm sdp}.$$

Proof. By applying a suitable rank reduction procedure if necessary, we can assume that the rank r of the optimal SDP solution Z^* satisfies $r = 1$ if $m \le 3$ and $r \le \sqrt{m}$ if $m \ge 4$; see [9, §5]. Thus, if $m \le 3$, then $Z^* = z^*(z^*)^H$ for some $z^* \in \mathbb{C}^{\rm n}$ and it is readily seen that z^* is an optimal solution of (1.1), so that $v_{\rm sdp} = v_{\rm qp}$. Otherwise, we apply the randomization procedure to Z^*. By choosing $\mu = 2$ and $\gamma = \frac{1}{4m}$, it is easily verified using $r \le \sqrt{m}$ that

$$\frac{4}{3}\gamma \ge 16(r-1)^2\gamma^2 \quad \forall\ m = 1, 2, ...$$

Therefore, it follows from Lemma 3.5 that

$$\mathrm{Prob}\left\{\min_{1\le i\le m} \xi^H H_i \xi \ge \gamma,\ \|\xi\|^2 \le \mu \mathrm{Tr}(Z^*)\right\} \ge 1 - m\frac{4}{3}\gamma - \frac{1}{\mu} = \frac{1}{6}.$$

Then, similar to the proof of Theorem 3.3, we obtain that with probability of at least 1/6, $z^*(\xi)$ is a feasible solution of (1.1) and $v_{\rm qp} \le \|z^*(\xi)\|^2 \le 8m \cdot v_{\rm sdp}$.[3] ■

The proof of Theorem 3.6 shows that, by repeating the randomization procedure, the probability of generating a feasible solution with a performance ratio no more than $8m$ approaches 1 exponentially fast (independent of problem size). Alternatively, a de-randomization technique from theoretical computer science can perhaps convert the above randomization procedure into a polynomial-time deterministic algorithm [12]; also see [14].

Theorem 3.6 shows that the worst-case performance of SDP relaxation deteriorates *linearly* with the number of quadratic constraints. This contrasts with the *quadratic* rate of deterioration in the real case (see Theorem 3.3). Thus, the SDP relaxation can yield better performance in the complex case. This is in the same spirit as the recent results in [26] which showed that the quality of SDP relaxation improves by a constant factor for certain quadratic *maximization* problems when the space is changed from $\mathbb{R}^{\rm n}$ to $\mathbb{C}^{\rm n}$. Below we give an example demonstrating that this approximation bound is tight up to a constant factor.

EXAMPLE 2: For any $m \ge 2$ and $n \ge 2$, let $K = \lceil \sqrt{m} \rceil$ (so $K \ge 2$). Consider a special instance of (1.2), corresponding to (1.1) with $|\mathcal{I}_i| = 1$ (i.e., each H_i has rank 1), whereby

$$h_\ell = \left(\cos\frac{j\pi}{K}, \sin\frac{j\pi}{K} e^{\frac{\mathrm{i}2k\pi}{K}}, 0, \ldots, 0\right)^T \quad \text{with} \quad \ell = jK - K + k, \qquad j, k = 1, ..., K.$$

[3] The probability that no such ξ is generated after N independent trials is at most $(5/6)^N$, which for $N = 30$ equals 0.00421.. Thus, such ξ requires relatively few trials to generate.

Hence there are K^2 complex rank-1 constraints. Let $z^* = (z_1^*, \ldots, z_n^*)^T \in \mathbb{C}^n$ be an optimal solution of (1.2) corresponding to the above choice of $\lceil\sqrt{m}\rceil^2$ steering vectors h_ℓ. By a phase rotation if necessary, we can without loss of generality assume that z_1^* is real and write

$$(z_1^*, z_2^*) = \rho(\cos\theta, \sin\theta e^{\mathbf{i}\psi}), \quad \text{for some } \theta, \psi \in [0, 2\pi).$$

Since $\{2k\pi/K,\ k = 1, ..., K\}$ and $\{j\pi/K,\ j = 1, ..., K\}$ are uniformly spaced in $[0, 2\pi)$ and $[0, \pi)$ respectively, there must exist integers j and k such that

$$\left|\psi - \frac{2k\pi}{K}\right| \le \frac{\pi}{K} \quad \text{and} \quad \text{either } \left|\theta - \frac{j\pi}{K} - \frac{\pi}{2}\right| \le \frac{\pi}{2K} \text{ or } \left|\theta - \frac{j\pi}{K} + \frac{\pi}{2}\right| \le \frac{\pi}{2K}.$$

Without loss of generality, we assume

$$\left|\theta - \frac{j\pi}{K} - \frac{\pi}{2}\right| \le \frac{\pi}{2K}.$$

Since the last $(n-2)$ entries of each h_ℓ are zero, it is readily seen that, for $\ell = jK - K + k$,

$$\begin{aligned}
\left|\mathrm{Re}(h_\ell^H z^*)\right| &= \rho\left|\cos\theta\cos\frac{j\pi}{K} + \sin\theta\sin\frac{j\pi}{K}\cos\left(\psi - \frac{2k\pi}{K}\right)\right| \\
&= \rho\left|\cos\left(\theta - \frac{j\pi}{K}\right) + \sin\theta\sin\frac{j\pi}{K}\left(\cos\left(\psi - \frac{2k\pi}{K}\right) - 1\right)\right| \\
&= \rho\left|\sin\left(\theta - \frac{j\pi}{K} - \frac{\pi}{2}\right) - 2\sin\theta\sin\frac{j\pi}{K}\sin^2\left(\frac{K\psi - 2k\pi}{2K}\right)\right| \\
&\le \rho\left|\sin\frac{\pi}{2K}\right| + 2\rho\sin^2\frac{\pi}{2K} \\
&\le \frac{\rho\pi}{2K} + \frac{\rho\pi^2}{2K^2}.
\end{aligned}$$

In addition, we have

$$\begin{aligned}
\left|\mathrm{Im}(h_\ell^H z^*)\right| &= \rho\left|\sin\theta\sin\frac{j\pi}{K}\sin\left(\psi - \frac{2k\pi}{K}\right)\right| \\
&\le \rho\left|\sin\left(\psi - \frac{2k\pi}{K}\right)\right| \\
&\le \rho\left|\psi - \frac{2k\pi}{K}\right| \le \frac{\rho\pi}{K}.
\end{aligned}$$

Combining the above two bounds, we obtain

$$\left|h_\ell^H z^*\right| \le \left|\mathrm{Re}(h_\ell^H z^*)\right| + \left|\mathrm{Im}(h_\ell^H z^*)\right| \le \frac{3\rho\pi}{2K} + \frac{\rho\pi^2}{2K^2}.$$

Since z^* satisfies the constraint $|h_\ell^H z^*| \ge 1$, it follows that

$$\|z^*\| \ge \rho \ge \frac{2K^2|h_\ell^H z^*|}{\pi(3K+\pi)} \ge \frac{2K^2}{\pi(3K+\pi)},$$

implying

$$v_{\rm qp} = \|z^*\|^2 \geq \frac{4K^4}{\pi^2(3K+\pi)^2} = \frac{4\lceil\sqrt{m}\rceil^4}{\pi^2(3\lceil\sqrt{m}\rceil+\pi)^2}.$$

On the other hand, the positive semidefinite matrix

$$Z^* = {\rm diag}\{1, 1, 0, \ldots, 0\}$$

is feasible for the SDP relaxation (3.1), and it has an objective value of ${\rm Tr}(Z^*) = 2$. Thus, for this instance, we have

$$v_{\rm qp} \geq \frac{2\lceil\sqrt{m}\rceil^4}{\pi^2(3\lceil\sqrt{m}\rceil+\pi)^2}\ v_{\rm sdp} \geq \frac{2m}{\pi^2(3+\pi/2)^2}\ v_{\rm sdp}.$$

The preceding example and Theorem 3.6 show that the SDP relaxation (3.1) can be weak if the number of quadratic constraints is large, especially when the steering vectors h_ℓ are in a certain sense "uniformly distributed" in space. In the next subsection, we will tighten the approximation bound in Theorem 3.6 by considering special cases where the steering vectors are "not too spread out in space".

3.3. Specially configured steering vectors: the complex case. We consider the complex case of $\mathbb{F} = \mathbb{C}$. Let Z^* be any optimal solution of (3.1). Since Z^* is feasible for (3.1), $Z^* \neq 0$. Then

$$Z^* = \sum_{k=1}^r w_k w_k^H, \tag{3.9}$$

for some nonzero $w_k \in \mathbb{C}^{\rm n}$, where $r := {\rm rank}(Z^*) \geq 1$. By decomposing $w_k = u_k + v_k$, with $u_k \in {\rm span}\{h_1, ..., h_M\}$ and $v_k \in {\rm span}\{h_1, ..., h_M\}^\perp$, it is easily checked that $\tilde{Z} := \sum_{k=1}^r u_k u_k^H$ is feasible for (3.1) and

$$\langle I, Z^*\rangle = \sum_{k=1}^r \|u_k + v_k\|^2 = \sum_{k=1}^r (\|u_k\|^2 + \|v_k\|^2) = \langle I, \tilde{Z}\rangle + \sum_{k=1}^r \|v_k\|^2.$$

This implies $v_k = 0$ for all k, so that

$$w_k \in {\rm span}\{h_1, ..., h_M\}. \tag{3.10}$$

Below we show that the SDP relaxation (3.1) provides a *constant factor approximation* to the QP (1.1) when the phase spread of the entries of h_ℓ is bounded away from $\pi/2$.

THEOREM 3.7. *Suppose that*

$$h_\ell = \sum_{i=1}^p \beta_{i\ell} g_i \quad \forall\ \ell = 1, ..., M, \tag{3.11}$$

for some $p \geq 1$, $\beta_{i\ell} \in \mathbb{C}$ and $g_i \in \mathbb{C}^{\rm n}$ such that $\|g_i\| = 1$ and $g_i^H g_j = 0$ for all $i \neq j$. Then the following results hold.

(a) *If* $\mathrm{Re}(\beta_{i\ell}^H\beta_{j\ell}) > 0$ *whenever* $\beta_{i\ell}^H\beta_{j\ell} \neq 0$, *then* $v_{\rm qp} \leq Cv_{\rm sdp}$, *where*

$$C := \max_{i,j,\ell \mid \beta_{i\ell}^H\beta_{j\ell}\neq 0} \left(1 + \frac{|\mathrm{Im}(\beta_{i\ell}^H\beta_{j\ell})|^2}{|\mathrm{Re}(\beta_{i\ell}^H\beta_{j\ell})|^2}\right)^{1/2}. \tag{3.12}$$

(b) *If* $\beta_{i\ell} = |\beta_{i\ell}|e^{\mathbf{i}\phi_{i\ell}}$, *where*

$$\phi_{i\ell} \in [\bar{\phi}_\ell - \phi, \bar{\phi}_\ell + \phi] \quad \forall\, i,\ell, \quad \textit{for some } 0 \leq \phi < \frac{\pi}{4} \textit{ and some } \bar{\phi}_\ell \in \mathbb{R}, \tag{3.13}$$

then $\mathrm{Re}(\beta_{i\ell}^H\beta_{j\ell}) > 0$ *whenever* $\beta_{i\ell}^H\beta_{j\ell} \neq 0$, *and* C *given by* (3.12) *satisfies*

$$C \leq \frac{1}{\cos(2\phi)}. \tag{3.14}$$

Proof. (a) By (3.10), we have

$$w_k = \sum_{i=1}^p \alpha_{ki} g_i,$$

for some $\alpha_{ki} \in \mathbb{C}$. This together with (3.9) yields

$$\begin{aligned}\langle I, Z^*\rangle = \sum_{k=1}^r \|w_k\|^2 \;&=\; \sum_{k=1}^r \left\|\sum_{i=1}^p \alpha_{ki} g_i\right\|^2 \\ &= \sum_{k=1}^r \sum_{i=1}^p |\alpha_{ki}|^2 \;=\; \sum_{i=1}^p \lambda_i^2,\end{aligned}$$

where the third equality uses the orthonormal properties of $g_1, ..., g_p$, and the last equality uses $\lambda_i := \left(\sum_{k=1}^r |\alpha_{ki}|^2\right)^{1/2} = \|(\alpha_{ki})_{k=1}^r\|$.

Let

$$z^* := \sum_{i=1}^p \lambda_i g_i.$$

Then, the orthonormal properties of $g_1, ..., g_p$ yields

$$\|z^*\|^2 = \left\|\sum_{i=1}^p \lambda_i g_i\right\|^2 = \sum_{i=1}^p \lambda_i^2 = \langle I, Z^*\rangle = v_{\rm sdp}. \tag{3.15}$$

Moreover, for each $\ell \in \{1, ..., M\}$, we obtain from (3.9) that

$$\begin{aligned}\langle h_\ell h_\ell^H, Z^*\rangle &= \sum_{k=1}^r \langle h_\ell h_\ell^H, w_k w_k^H\rangle = \sum_{k=1}^r |h_\ell^H w_k|^2 \\ &= \sum_{k=1}^r \left|\sum_{i=1}^p \alpha_{ki} h_\ell^H g_i\right|^2 = \sum_{k=1}^r \left|\sum_{i=1}^p \alpha_{ki}\beta_{i\ell}\right|^2\end{aligned}$$

$$
\begin{aligned}
&= \mathrm{Re}\left(\sum_{k=1}^{r}\sum_{i=1}^{p}\sum_{j=1}^{p}\alpha_{ki}^H\alpha_{kj}\beta_{i\ell}^H\beta_{j\ell}\right) = \mathrm{Re}\left(\sum_{i=1}^{p}\sum_{j=1}^{p}\beta_{i\ell}^H\beta_{j\ell}\sum_{k=1}^{r}\alpha_{ki}^H\alpha_{kj}\right)\\
&= \sum_{i=1}^{p}\sum_{j=1}^{p}\mathrm{Re}\left(\beta_{i\ell}^H\beta_{j\ell}\sum_{k=1}^{r}\alpha_{ki}^H\alpha_{kj}\right)\\
&\le \sum_{i=1}^{p}\sum_{j=1}^{p}\left|\beta_{i\ell}^H\beta_{j\ell}\right|\left|\sum_{k=1}^{r}\alpha_{ki}^H\alpha_{kj}\right| \le \sum_{i=1}^{p}\sum_{j=1}^{p}\left|\beta_{i\ell}^H\beta_{j\ell}\right|\|(\alpha_{ki})_{k=1}^r\|\|(\alpha_{kj})_{k=1}^r\|\\
&= \sum_{i=1}^{p}\sum_{j=1}^{p}\left|\beta_{i\ell}^H\beta_{j\ell}\right|\lambda_i\lambda_j,
\end{aligned}
$$

where the fourth equality uses (3.11) and the orthonormal properties of $g_1, ..., g_p$; the last inequality is due to the Cauchy-Schwarz inequality. Then, it follows that

$$
\begin{aligned}
\langle h_\ell h_\ell^H, Z^*\rangle &\le \sum_{i=1}^{p}\sum_{j=1}^{p}\left(|\mathrm{Re}(\beta_{i\ell}^H\beta_{j\ell})|^2 + |\mathrm{Im}(\beta_{i\ell}^H\beta_{j\ell})|^2\right)^{1/2}\lambda_i\lambda_j\\
&= \sum_{i=1}^{p}\sum_{j=1}^{p}\left|\mathrm{Re}(\beta_{i\ell}^H\beta_{j\ell})\right|\left(1+\frac{|\mathrm{Im}(\beta_{i\ell}^H\beta_{j\ell})|^2}{|\mathrm{Re}(\beta_{i\ell}^H\beta_{j\ell})|^2}\right)^{1/2}\lambda_i\lambda_j\\
&\le \sum_{i=1}^{p}\sum_{j=1}^{p}\left|\mathrm{Re}(\beta_{i\ell}^H\beta_{j\ell})\right|C\lambda_i\lambda_j\\
&= \sum_{i=1}^{p}\sum_{j=1}^{p}\mathrm{Re}(\beta_{i\ell}^H\beta_{j\ell})C\lambda_i\lambda_j,
\end{aligned}
$$

where the summation in the second step is taken over i, j with $\beta_{i\ell}^H\beta_{j\ell} \neq 0$, the third step is due to (3.12), and the last step is due to the assumption that $\mathrm{Re}(\beta_{i\ell}^H\beta_{j\ell}) > 0$ whenever $\beta_{i\ell}^H\beta_{j\ell} \neq 0$. Also, we have from (3.11) and the orthonormal properties of $g_1, ..., g_p$ that

$$
|h_\ell^H z^*|^2 = \left\|\sum_{i=1}^{p}\lambda_i h_\ell^H g_i\right\|^2 = \left\|\sum_{i=1}^{p}\lambda_i\beta_{i\ell}\right\|^2 = \sum_{i=1}^{p}\sum_{j=1}^{p}\lambda_i\lambda_j\mathrm{Re}(\beta_{i\ell}^H\beta_{j\ell}).
$$

Comparing the above two displayed equations, we see that

$$
\langle h_\ell h_\ell^H, Z^*\rangle \le C|h_\ell^H z^*|^2, \quad \ell = 1, ..., M.
$$

Since Z^* is feasible for (3.1), this shows that $\sqrt{C}z^*$ is feasible for (1.1), which further implies

$$
v_{\mathrm{qp}} \le \left\|\sqrt{C}z^*\right\|^2 = C\|z^*\|^2 = Cv_{\mathrm{sdp}}.
$$

This proves the desired result.

(b) The condition (3.13) implies that $|\phi_{i\ell} - \phi_{j\ell}| \le 2\phi < \pi/2$. In other words, the phase angle spread of the entries of each $\beta_\ell = (\beta_{1\ell}, \beta_{2\ell}, \ldots, \beta_{n\ell})^T$ is no more than 2ϕ. This further implies that

$$\cos(\phi_{i\ell} - \phi_{j\ell}) \ge \cos(2\phi) \quad \forall\, i, j, \ell. \tag{3.16}$$

We have

$$\begin{aligned}\beta_{i\ell}^H \beta_{j\ell} &= |\beta_{i\ell}| e^{-\mathbf{i}\phi_{i\ell}} |\beta_{j\ell}| e^{\mathbf{i}\phi_{j\ell}} \\ &= |\beta_{i\ell}||\beta_{j\ell}| e^{\mathbf{i}(\phi_{j\ell} - \phi_{i\ell})} \\ &= |\beta_{i\ell}||\beta_{j\ell}| (\cos(\phi_{j\ell} - \phi_{i\ell}) + \mathbf{i}\sin(\phi_{j\ell} - \phi_{i\ell})).\end{aligned}$$

Since $|\phi_{i\ell} - \phi_{j\ell}| < \pi/2$ so that $\cos(\phi_{j\ell} - \phi_{i\ell}) > 0$, we see that $\mathrm{Re}(\beta_{i\ell}^H \beta_{j\ell}) > 0$ whenever $\beta_{i\ell}^H \beta_{j\ell} \ne 0$. Then

$$\left(1 + \frac{|\mathrm{Im}(\beta_{i\ell}^H \beta_{j\ell})|^2}{|\mathrm{Re}(\beta_{i\ell}^H \beta_{j\ell})|^2}\right)^{1/2} \le \left(1 + \tan^2(\phi_{j\ell} - \phi_{i\ell})\right)^{1/2} = \frac{1}{\cos(\phi_{j\ell} - \phi_{i\ell})} \le \frac{1}{\cos(2\phi)},$$

where the last step uses (3.16). Using this in (3.12) completes the proof. ∎

In Theorem 3.7(b), we can more generally consider $\beta_{i\ell}$ of the form $\beta_{i\ell} = \omega_{i\ell} e^{\mathbf{i}\phi_{i\ell}}(1 + \mathbf{i}\theta_{i\ell})$, where $\omega_{i\ell} \ge 0$, $\alpha_{i\ell}$ satisfies (3.13), and

$$|\theta_{j\ell} - \theta_{i\ell}| \le \sigma |1 + \theta_{i\ell}\theta_{j\ell}| \quad \forall\, i, j, \ell, \quad \text{for some } \sigma \ge 0 \text{ with } \tan(2\phi)\sigma < 1. \tag{3.17}$$

Then the proof of Theorem 3.7(b) can be extended to show the following upper bound on C given by (3.12):

$$C \le \frac{1}{\cos(2\phi)} \cdot \frac{\sqrt{1+\sigma^2}}{1 - \tan(2\phi)\sigma}. \tag{3.18}$$

However, this generalization is superficial as we can also derive (3.18) from (3.14) by rewriting $\beta_{i\ell}$ as

$$\beta_{i\ell} = |\beta_{i\ell}| e^{\mathbf{i}\tilde{\phi}_{i\ell}} \quad \text{with} \quad \tilde{\phi}_{i\ell} = \phi_{i\ell} + \tan^{-1}(\theta_{i\ell}).$$

Then, applying (3.14) yields $C \ge \cos(2\tilde{\phi})$, where $\tilde{\phi} = \max_{i,j,\ell} |\tilde{\phi}_{i\ell} - \tilde{\phi}_{j\ell}|/2$. Using trigonometric identity, it can be shown that $\cos(2\tilde{\phi})$ equals the right-hand side of (3.18) with $\sigma = \max_{i,j,\ell \;|\; \theta_{i\ell}\theta_{j\ell} \ne -1} |\theta_{j\ell} - \theta_{i\ell}|/|1 + \theta_{i\ell}\theta_{j\ell}|$.

Notice that Theorem 3.7(b) implies that if $\phi = 0$, then the SDP relaxation (3.1) is tight for the quadratically constrained QP (1.1) with $\mathbb{F} = \mathbb{C}$. Such is the case when all components of h_ℓ, $\ell = 1, ..., M$, are real and nonnegative.

4. A convex QP restriction. In this subsection, we consider a convex quadratic programming *restriction* of (1.2) in the complex case of $\mathbb{F} = \mathbb{C}$ and analyze its approximation bound. Let us write h_ℓ (the channel steering vector) as

$$h_\ell = (\ldots, |h_{j\ell}| e^{\mathbf{i}\phi_{j\ell}}, \ldots)_{j=1,\ldots,n}^T.$$

For any $\bar{\phi}_j \in [0, 2\pi)$, $j = 1, ..., n$, and any $\phi \in (0, \pi/2)$, define the four corresponding index subsets:

$$\begin{aligned}
J_\ell^1 &:= \{j \mid \phi_{j\ell} \in [\bar{\phi}_j - \phi, \bar{\phi}_j + \phi]\},\\
J_\ell^2 &:= \{j \mid \phi_{j\ell} \in [\bar{\phi}_j - \phi + \pi/2, \bar{\phi}_j + \phi + \pi/2]\},\\
J_\ell^3 &:= \{j \mid \phi_{j\ell} \in [\bar{\phi}_j - \phi + \pi, \bar{\phi}_j + \phi + \pi]\},\\
J_\ell^4 &:= \{j \mid \phi_{j\ell} \in [\bar{\phi}_j - \phi + 3\pi/2, \bar{\phi}_j + \phi + 3\pi/2]\},
\end{aligned}$$

for $\ell = 1, ..., M$. The above four subsets are pairwise disjoint if and only if $\phi < \pi/4$, and are collectively exhaustive if and only if $\phi \geq \pi/4$. Choose an index subset J with the property that

$$\text{for each } \ell, \text{ at least one of } J_\ell^1,\ J_\ell^2,\ J_\ell^3,\ J_\ell^4 \text{ contains } J.$$

Of course, $J = \emptyset$ is always allowable, but we should choose J maximally since our approximation bound will depend on the ratio $n/|J|$ (see Theorem 4.1 below). Partition the constraint set index $\{1, ..., M\}$ into four subsets K^1, K^2, K^3, K^4 such that

$$J \subseteq J_\ell^k \quad \forall\ \ell \in K^k,\ k = 1, 2, 3, 4.$$

Consider the following convex QP restriction of (1.2) corresponding to K^1, K^2, K^3, K^4:

$$\begin{array}{rcll}
\upsilon_{\rm cqp} & := & \min & \|z\|^2 \\
& & \text{s.t.} & \begin{array}{rll}
\mathrm{Re}(h_\ell^H z) & \geq 1 & \forall\ \ell \in K^1,\\
-\mathrm{Im}(h_\ell^H z) & \geq 1 & \forall\ \ell \in K^2,\\
-\mathrm{Re}(h_\ell^H z) & \geq 1 & \forall\ \ell \in K^3,\\
\mathrm{Im}(h_\ell^H z) & \geq 1 & \forall\ \ell \in K^4.
\end{array}
\end{array} \tag{4.1}$$

The above problem is a restriction of (1.2) because, for any $z \in \mathbb{C}$,

$$\begin{aligned}
|z| &\geq \max\{|\mathrm{Re}(z)|, |\mathrm{Im}(z)|\}\\
&= \max\{\mathrm{Re}(z), \mathrm{Im}(z), -\mathrm{Re}(z), -\mathrm{Im}(z)\}.
\end{aligned}$$

If $J \neq \emptyset$ and $(\ldots, h_{j\ell}, \ldots)_{j\in J} \neq 0$ for $\ell = 1, ..., M$, then (4.1) is feasible, and hence has an optimal solution. Since (4.1) is a restriction of (1.2), $\upsilon_{\rm qp} \leq \upsilon_{\rm cqp}$. We have the following approximation bound.

THEOREM 4.1. *Suppose that $J \neq \emptyset$ and (4.1) is feasible. Then,*

$$\upsilon_{\rm cqp} \leq \upsilon_{\rm qp}\ \frac{N}{\cos^2\phi} \max_{k=1,\ldots,N} \left(\max_{j\in\hat{J}_k} \frac{\bar{\eta}_j}{\underline{\eta}_{\pi_k(j)}} \right)^2,$$

where $N := \lceil n/|J| \rceil$, $\bar{\eta}_j := \max_\ell |h_{j\ell}|$, $\underline{\eta}_j := \min_{\ell | h_{j\ell} \neq 0} |h_{j\ell}|$, $\hat{J}_1, ..., \hat{J}_N$ is any partition of $\{1, ..., n\}$ satisfying $|\hat{J}_k| \leq |J|$ for $k = 1, ..., N$, and π_k is any injective mapping from $\hat{J}_k$ to J.

Proof. By making the substitution

$$z_j^{\rm new} \leftarrow z_j e^{\mathbf{i}\bar{\phi}_j},$$

we can without loss of generality assume that $\bar{\phi}_j = 0$ for all j and ℓ.

Let z^* denote an optimal solution of (1.2) and write

$$z^* = (\ldots, r_j e^{\mathbf{i}\beta_j}, \ldots)_{j=1,\ldots,n}^T,$$

with $r_j \geq 0$. Then, for any ℓ, we have from $|h_{j\ell}| \leq \bar{\eta}_j$ for all j that

$$1 \leq |h_\ell^H z^*| \leq r := \sum_{j=1}^n r_j \bar{\eta}_j.$$

Also, we have

$$\upsilon_{\text{qp}} = \|z^*\|^2 = \sum_{j=1}^n r_j^2.$$

Define

$$R_k := \left(\sum_{j \in \hat{J}_k} r_j^2 \right)^{1/2}, \qquad S_k := \sum_{j \in \hat{J}_k} r_j \bar{\eta}_j.$$

Then

$$1 \leq r = \sum_{k=1}^N S_k, \qquad \upsilon_{\text{qp}} = \sum_{k=1}^N R_k^2.$$

Without loss of generality, assume that $R_1/S_1 = \min_k R_k/S_k$. Then, using the fact that

$$\min_k \frac{|x_k|}{|y_k|} \leq \sqrt{N} \frac{\|x\|_2}{\|y\|_1}$$

for any $x, y \in \mathbb{R}^{\text{N}}$ with $y \neq 0$,[4] we see from the above relations that

$$\begin{aligned} \frac{R_1}{S_1} &\leq \frac{R_1}{S_1} r \\ &\leq \sqrt{N} \frac{\sqrt{\upsilon_{\text{qp}}}}{r} r \\ &= \sqrt{N} \sqrt{\upsilon_{\text{qp}}}. \end{aligned}$$

Since $|\hat{J}_1| \leq |J|$, there is an injective mapping π from $\hat{J}_1$ to J. Let $\omega := \min_{j \in \hat{J}_1} \underline{\eta}_{\pi(j)}/\bar{\eta}_j$. Define the vector $\bar{z} \in \mathbb{C}^{\text{n}}$ by

$$\bar{z}_j := \begin{cases} r_{\pi^{-1}(j)}/(S_1 \omega \cos\phi) & \text{if } j \in \pi(\hat{J}_1); \\ 0 & \text{else.} \end{cases}$$

[4] *Proof.* Suppose the contrary, so that for some $x, y \in \mathbb{R}^{\text{N}}$ with $y \neq 0$, we have $|x_k|/|y_k| > \sqrt{N}\|x\|_2/\|y\|_1$ for all k. Then, multiplying both sides by $|y_k|$ and summing over k yields $\|x\|_1 > \sqrt{N}\|x\|_2$, contradicting properties of 1- and 2-norms.

Then,

$$\|\bar{z}\|^2 = \frac{R_1^2}{S_1^2\omega^2\cos^2\phi} \le \frac{N\upsilon_{\rm qp}}{\omega^2\cos^2\phi}.$$

Moreover, for each $\ell \in K^1$, since $\pi(\hat{J}_1) \subseteq J \subseteq J_\ell^1$, we have

$$\begin{aligned}
\mathrm{Re}\left(h_\ell^H \bar{z}\right) &= \mathrm{Re}\left(\sum_{j\in\pi(\hat{J}_1)} h_{j\ell}^H \bar{z}_j\right) \\
&= \frac{1}{S_1\omega\cos\phi}\mathrm{Re}\left(\sum_{j\in\pi(\hat{J}_1)} r_{\pi^{-1}(j)}|h_{j\ell}|e^{-\mathbf{i}\phi_{j\ell}}\right) \\
&= \frac{1}{S_1\omega\cos\phi}\sum_{j\in\pi(\hat{J}_1)} r_{\pi^{-1}(j)}|h_{j\ell}|\cos\phi_{j\ell} \\
&\ge \frac{1}{S_1\omega\cos\phi}\sum_{j\in\pi(\hat{J}_1)} r_{\pi^{-1}(j)}\underline{\eta}_j\cos\phi \\
&= \frac{1}{S_1\omega}\sum_{j\in\hat{J}_1} r_j\bar{\eta}_j\,\frac{\underline{\eta}_{\pi(j)}}{\bar{\eta}_j} \\
&\ge \frac{1}{S_1\omega}\sum_{j\in\hat{J}_1} r_j\bar{\eta}_j\cdot\min_{j\in\hat{J}_1}\frac{\underline{\eta}_{\pi(j)}}{\bar{\eta}_j} \\
&= 1,
\end{aligned}$$

where the first inequality uses $|h_{j\ell}| \ge \underline{\eta}_j$ and $\phi_{j\ell} \in [-\phi, \phi]$ for $j \in J_\ell^1$. Since $\bar{z}_j = 0$ for $j \notin J_\ell^1$, this shows that $\bar{z}$ satisfies the first set of constraints in (4.1). A similar reasoning shows that $\bar{z}$ satisfies the remaining three sets of constraints in (4.1). ■

Notice that the $\bar{z}$ constructed in the proof of Theorem 4.1 is feasible for the further restriction of (4.1) whereby $z_j = 0$ for all $j \notin J$. This further restricted problem has the same (worst-case) approximation bound specified in Theorem 4.1.

Let us compare the two approximation bounds in Theorem 3.7 and Theorem 4.1. First, the required assumptions are different. On the one hand, the bound in Theorem 3.7 does not depend on $|h_{j\ell}|$, while the bound in Theorem 4.1 does. On the other hand, Theorem 3.7 requires that the bounded angular spread

$$|\phi_{j\ell} - \phi_{i\ell}| \le 2\phi \quad \forall\, j, \ell, \tag{4.2}$$

for some $\phi < \pi/4$, while Theorem 4.1 allows $\phi < \pi/2$ and only requires the condition (4.2) for all $1 \le \ell \le M$ and $j \in J$, where J is a pre-selected index set. Thus, the bounded angular spread condition required in Theorem 3.7 corresponds exactly to $|J| = n$. Thus, the assumptions required in the two theorems do not imply one another. Second, the two performance ratios are also different. Naturally, the final performance ratio in Theorem 4.1 depends on the choice of

J through the ratio $|J|/n$, so a large J is preferred. In the event that the assumptions of both theorems are satisfied and let us assume for simplicity that $\bar{\eta}_j = \underline{\eta}_j$ for all j, then $|J| = n$ and $\phi < \pi/4$, in which case Theorem 4.1 gives a performance ratio of $1/\cos^2\phi$ while Theorem 3.7 gives $1/\cos(2\phi)$. Since $\cos(2\phi) = \cos^2\phi - \sin^2\phi \leq \cos^2\phi$, we have $1/\cos(2\phi) \geq 1/\cos^2\phi$, showing that Theorem 4.1 gives a tighter approximation bound. However, this does not mean Theorem 4.1 is stronger than Theorem 3.7 since the two theorems hold under different assumptions in general.

We can specialize Theorem 4.1 to a typical situation in transmit beamforming. Consider a uniform linear transmit antenna array consisting of n elements, and let us assume that the M receivers are in a sector area from the far field, and the propagation is line-of-sight. By reciprocity, each steering vector h_ℓ will be Vandermonde with generator $e^{-\mathbf{i}2\pi\frac{d}{\lambda}\sin\theta_\ell}$ (see, e.g., [10]), where d is the inter-antenna spacing, λ is the wavelength, and θ_ℓ is the angle of arrival of the ℓth receiving antenna. In a sector of approximately 60 degrees about the array broadside, we will have $|\theta_\ell| \leq \pi/3$. Suppose that $d/\lambda = 1/2$. Then the steering vector corresponding to the ℓth receiving antenna will have the form

$$h_\ell = (\ldots, e^{-\mathbf{i}(j-1)\pi\sin\theta_\ell}, \ldots)^T_{j=1,\ldots,n}.$$

In this case, we have that $\phi_{j\ell} = (j-1)\pi\sin\theta_\ell$ and $|h_{j\ell}| = 1$ for all j and ℓ. We can take, e.g.,

$$\bar{\phi}_j = 0, \quad \phi = \bar{j}\pi\max_\ell|\sin\theta_\ell|, \quad J = \{1, ..., \bar{j}+1\},$$

where $\bar{j} := \lfloor 1/\max_\ell|\sin\theta_\ell|\rfloor$. Thus, the assumptions of Theorem 4.1 are satisfied. Moreover, since $|\theta_\ell| \leq \pi/3$ for all ℓ, it follows that $|J| = \bar{j}+1 \geq 2$. If n is not large, say, $n \leq 8$, then Theorem 4.1 gives a performance ratio of $n/(|J|\cos^2\phi) \leq 16$.

More generally, if we can choose the partition $\hat{J}_1, ..., \hat{J}_N$ and the mapping π_k in Theorem 4.1 such that

$$(\ldots, \bar{\eta}_j, \ldots)_{j\in\hat{J}_k} = (\ldots, \underline{\eta}_{\pi_k(j)}, \ldots)_{j\in J} \quad \forall\, k,$$

then the performance ratio in Theorem 4.1 simplifies to $N/\cos^2\phi$. In particular, this holds when $|h_{j\ell}| = \eta > 0$ for all j and ℓ or when $J = \{1, ..., n\}$ (so that $N = 1$) and $|h_{j\ell}|$ is independent of ℓ for all j, and more generally, when the channel coefficients periodically repeat their magnitudes. In general, we should choose the partition $\hat{J}_1, ..., \hat{J}_N$ and the mapping π_k to make the performance ratio in Theorem 4.1 small. For example, if $J = \hat{J}_1 = \{1, 2\}$ and $\bar{\eta}_1 = 100$, $\bar{\eta}_2 = 10$, $\underline{\eta}_1 = 1$, $\underline{\eta}_2 = 10$, then $\pi_1(1) = 2, \pi_1(2) = 1$ is the better choice.

5. Homogeneous QP in Maximization Form. Let us now consider the following complex norm maximization problem with convex homogeneous quadratic constraints:

$$\begin{aligned} \upsilon_{\text{qp}} := \max \quad & \|z\|^2 \\ \text{s.t.} \quad & \sum_{\ell\in\mathcal{I}_i} |h_\ell^H z|^2 \leq 1, \quad i = 1, ..., m, \\ & z \in \mathbb{C}^{\text{n}}, \end{aligned} \tag{5.1}$$

where $h_\ell \in \mathbb{C}^{\text{n}}$.

To motivate this problem, consider the problem of designing an intercept beamformer[5] capable of suppressing signals impinging on the receiving antenna array from irrelevant or hostile emitters, e.g., jammers, whose steering vectors (spatial signatures, or "footprints") have been previously estimated, while achieving as high gain as possible for all other transmissions. The jammer suppression capability is captured in the constraints of (5.1), and $|\mathcal{I}_i| > 1$ covers the case where a jammer employs more than one transmit antennas. The maximization of the objective $\|z\|^2$ can be motivated as follows. In intercept applications, the steering vector of the emitter of interest, h, is *a priori* unknown, and is naturally modelled as random. A pertinent optimization objective is then the average beamformer output power, measured by $E[|h^H z|^2]$. Under the assumption that the entries of h are uncorrelated and have equal average power, it follows that $E[|h^H z|^2]$ is proportional to $\|z\|^2$, which is often referred to as the beamformer's *white noise gain*.

Similar to (1.1), we let

$$H_i := \sum_{\ell \in \mathcal{I}_i}^{m} h_\ell h_\ell^H$$

and consider the natural SDP relaxation of (5.1):

$$\begin{array}{rcll} \upsilon_{\rm sdp} & := & \max & \mathrm{Tr}(Z) \\ & & \text{s.t.} & \mathrm{Tr}(H_i Z) \le 1, \quad i = 1, ..., m, \\ & & & Z \succeq 0, \quad Z \text{ is complex and Hermitian.} \end{array} \tag{5.2}$$

We are interested in lower bounds for the relaxation performance of the form

$$\upsilon_{\rm qp} \ge C\ \upsilon_{\rm sdp},$$

where $0 < C \le 1$. It is easily checked that (5.2) has an optimal solution.

Let Z^* be an optimal solution of (5.2). We will analyze the performance of the SDP relaxation using the following randomization procedure:

1. Generate a random vector $\xi \in \mathbb{C}^{\rm n}$ from the *complex-valued* normal distribution $N_c(0, Z^*)$.
2. Let $z^*(\xi) = \xi / \max\limits_{1 \le i \le m} \sqrt{\xi^H H_i \xi}$.

First, we need the following lemma analogous to Lemmas 3.1 and 3.4.

LEMMA 5.1. *Let $H \in \mathbb{C}^{\rm n \times n}$, $Z \in \mathbb{C}^{\rm n \times n}$ be two Hermitian positive semidefinite matrices (i.e., $H \succeq 0$, $Z \succeq 0$). Suppose $\xi \in \mathbb{C}^{\rm n}$ is a random vector generated from the complex-valued normal distribution $N_c(0, Z)$. Then, for any $\gamma > 0$,*

$$\mathrm{Prob}\left(\xi^H H \xi > \gamma E(\xi^H H \xi)\right) \le \bar{r}\ e^{-\gamma}, \tag{5.3}$$

where $\bar{r} := \min\{\mathrm{rank}(H), \mathrm{rank}(Z)\}$.

[5]Note that here we are talking about a receive beamformer, as opposed to our earlier motivating discussion of transmit beamformer design.

Proof. If $H = 0$, then (5.3) is trivially true. Suppose $H \neq 0$. Then, as in the proof of Lemma 3.1, we have

$$\text{Prob}\left(\xi^H H\xi > \gamma E(\xi^H H\xi)\right) = \text{Prob}\left(\sum_{i=1}^{\bar{r}} \bar{\lambda}_i |\bar{\xi}_i|^2 > \gamma\right),$$

where $\bar{\lambda}_1 \geq \bar{\lambda}_2 \geq \ldots \geq \bar{\lambda}_{\bar{r}} \geq 0$ satisfy $\bar{\lambda}_1 + \cdots + \bar{\lambda}_{\bar{r}} = 1$ and each $\bar{\xi}_i \in \mathbb{C}$ has the complex-valued normal distribution $N_c(0,1)$. Then

$$\begin{aligned}\text{Prob}\left(\xi^H H\xi > \gamma E(\xi^H H\xi)\right) &\leq \text{Prob}\left(|\bar{\xi}_1|^2 > \gamma \text{ or } |\bar{\xi}_2|^2 > \gamma \text{ or } \cdots \text{ or } |\bar{\xi}_{\bar{r}}|^2 > \gamma\right)\\ &\leq \sum_{i=1}^{\bar{r}} \text{Prob}\left(|\bar{\xi}_i|^2 > \gamma\right)\\ &= \bar{r}\ e^{-\gamma},\end{aligned}$$

where the last step uses (3.8). ∎

THEOREM 5.2. *For the complex QP* (5.1) *and its SDP relaxation* (5.2)*, we have* $v_{\text{sdp}} = v_{\text{qp}}$ *if* $m \leq 3$ *and otherwise*

$$v_{\text{qp}} \geq \frac{1}{4\ln(100K)} v_{\text{sdp}},$$

where $K := \sum_{i=1}^m \min\{\text{rank}(H_i), \sqrt{m}\}$.

Proof. By applying a suitable rank reduction procedure if necessary, we can assume that the rank r of the optimal SDP solution Z^* satisfies $r = 1$ if $m \leq 3$ and $r \leq \sqrt{m}$ if $m \geq 4$; see [9, §5]. Thus, if $m \leq 3$, then $Z^* = z^*(z^*)^H$ for some $z^* \in \mathbb{C}^n$ and it is readily seen that z^* is an optimal solution of (5.1), so that $v_{\text{sdp}} = v_{\text{qp}}$. Otherwise, we apply the randomization procedure to Z^*. By using Lemma 5.1, we have, for any $\gamma > 0$ and $\mu > 0$,

$$\begin{aligned}&\text{Prob}\left(\max_{1\leq i\leq m} \xi^H H_i \xi \leq \gamma,\ \|\xi\|^2 \geq \mu \text{Tr}(Z^*)\right)\\ &\geq 1 - \sum_{i=1}^m \text{Prob}\left(\xi^H H_i \xi > \gamma E(\xi^H H_i \xi)\right) - \text{Prob}\left(\|\xi\|^2 < \mu \text{Tr}(Z^*)\right)\\ &\geq 1 - Ke^{-\gamma} - \text{Prob}\left(\|\xi\|^2 < \mu \text{Tr}(Z^*)\right), \qquad (5.4)\end{aligned}$$

where the last step uses $r \leq \sqrt{m}$.

Let

$$\eta_j := \begin{cases} |\xi_j|^2 / Z^*_{jj}, & \text{if } Z^*_{jj} > 0;\\ 0, & \text{if } Z^*_{jj} = 0, \end{cases} \qquad j = 1, ..., n.$$

For simplicity, let us assume that $Z^*_{jj} > 0$ for all $j = 1, ..., n$. Since $\xi_j \sim N_c(0, Z^*_{jj})$, as we discussed in Subsection 3.2, $|\xi_j|$ follows a Weibull distribution with variance Z^*_{jj} (see (3.8)), and therefore

$$\text{Prob}\left(\eta_j \leq t\right) = 1 - e^{-t} \quad \forall\, t \in [0, \infty).$$

Hence,

$$E(\eta_j) = \int_0^\infty te^{-t}dt = 1, \quad E(\eta_j^2) = \int_0^\infty t^2e^{-t}dt = 2, \quad \mathrm{Var}(\eta_j) = 1.$$

Moreover,

$$E(|\eta_j - E(\eta_j)|) = \int_0^1 (1-t)e^{-t}dt + \int_1^\infty (t-1)e^{-t}dt = \frac{2}{e}.$$

Let us denote $\lambda_j = Z^*_{jj}/\mathrm{Tr}(Z^*)$, $j = 1, ..., n$, and $\eta := \sum_{j=1}^n \lambda_j\eta_j$. We have $E(\eta) = 1$ and

$$E(|\eta - E(\eta)|) = E\left(\left|\sum_{j=1}^n \lambda_j(\eta_j - E(\eta_j))\right|\right) \le \sum_{j=1}^n \lambda_j E(|\eta_j - E(\eta_j)|) = \frac{2}{e}.$$

Since, by Markov's inequality,

$$\mathrm{Prob}\,(|\eta - E(\eta)| > \alpha) \le \frac{E(|\eta - E(\eta)|)}{\alpha} \le \frac{2}{\alpha e}, \quad \forall\ \alpha > 0,$$

we have

$$\begin{aligned}\mathrm{Prob}\left(\|\xi\|^2 < \mu\mathrm{Tr}(Z^*)\right) &= \mathrm{Prob}\,(\eta < \mu)\\ &\le \mathrm{Prob}\,(|\eta - E(\eta)| > 1 - \mu)\\ &\le \frac{2}{e(1-\mu)}, \quad \text{for all } \mu \in (0,1).\end{aligned}$$

Substituting the above inequality into (5.4), we obtain

$$\mathrm{Prob}\left(\max_{1\le i\le m} \xi^H H_i\xi \le \gamma,\ \|\xi\|^2 \ge \mu\mathrm{Tr}(Z^*)\right) > 1 - Ke^{-\gamma} - \frac{2}{e(1-\mu)}, \quad \forall\ \mu \in (0,1).$$

Setting $\mu = 1/4$ and $\gamma = \ln(100K)$ yields a positive right-hand side of 0.00898.., which then proves the desired bound. ■

The above proof technique also applies to the real case, i.e., $h_\ell \in \mathbb{R}^n$ and $z \in \mathbb{R}^n$. The main difference is that $\xi \sim N(0, Z^*)$, so that $|\bar{\xi}_i|^2$ in the proof of Lemma 5.1 and η_j in the proof of Theorem 5.2 both follow a χ^2 distribution with one degree of freedom. Then

$$\mathrm{Prob}\left(|\bar{\xi}_i|^2 > \gamma\right) = \int_{\sqrt{\gamma}}^\infty \frac{e^{-t^2/2}}{\sqrt{2\pi}}dt \le \int_{\sqrt{\gamma}}^\infty \frac{e^{-\gamma t/2}}{\sqrt{2\pi}}dt = \sqrt{\frac{2}{\pi\gamma}}e^{-\gamma/2}, \quad \forall\ \gamma > 0,$$

$E(\eta_j) = 1$, and

$$\begin{aligned}E|\eta_j - E(\eta_j)| &= \int_0^\infty \frac{e^{-t/2}}{\sqrt{2\pi t}}|t-1|dt\\ &= \frac{1}{\sqrt{2\pi}}\int_0^1 \frac{e^{-t/2}}{\sqrt{t}}dt - \frac{1}{\sqrt{2\pi}}\int_0^1 \sqrt{t}e^{-t/2}dt\\ &\quad + \frac{1}{\sqrt{2\pi}}\int_1^\infty \sqrt{t}e^{-t/2}dt - \frac{1}{\sqrt{2\pi}}\int_1^\infty \frac{e^{-t/2}}{\sqrt{t}}dt\\ &= \frac{4}{\sqrt{2\pi e}} < 0.968,\end{aligned}$$

where in the last step we used integration by parts on the first and the fourth terms. This yields the analogous bound that, for any $\gamma \geq 1$ and $\mu \in (0,1)$,

$$\text{Prob}\left(\max_{1\leq i\leq m} \xi^T H_i \xi \leq \gamma,\ \|\xi\|^2 \geq \mu \text{Tr}(Z^*)\right) > 1 - K\sqrt{\frac{2}{\pi\gamma}}e^{-\gamma/2} - \frac{0.968}{1-\mu} > 1 - Ke^{-\gamma/2} - \frac{0.968}{1-\mu},$$

where $K := \sum_{i=1}^m \min\{\text{rank}(H_i), \sqrt{2m}\}$. Setting $\mu = 0.01$ and $\gamma = 2\ln(50K)$ yields a positive right-hand side of 0.0022... This in turn shows that $v_{\text{sdp}} = v_{\text{qp}}$ if $m \leq 2$ (see the proof of Theorem 3.3) and otherwise

$$v_{\text{qp}} \geq \frac{1}{200\ln(50K)} v_{\text{sdp}}.$$

We note that, in the real case, a sharper bound of

$$v_{\text{qp}} \geq \frac{1}{2\ln(2m\mu)} v_{\text{sdp}},$$

where $\mu := \min\{m, \max_i \text{rank}(H_i)\}$, was shown by Nemirovski, Roos and Terlaky [14] (also see [13, Theorem 4.7]), though the above proof seems simpler. Also, an example in [14] shows that the $O(1/\ln m)$ bound is tight (up to a constant factor) in the worst case. This example readily extends to the complex case by identifying $\mathbb{C}^n$ with $\mathbb{R}^{2n}$ and observing that $|h_\ell^H z| \geq |\text{Re}(h_\ell)^T \text{Re}(z) + \text{Im}(h_\ell)^T \text{Im}(z)|$ for any $h_\ell, z \in \mathbb{C}^n$. Thus, in the complex case, the $O(1/\ln m)$ bound is also tight (up to a constant factor).

6. Discussion. In this paper, we have analyzed the worst-case performance of SDP relaxation and convex restriction for a class of NP-hard quadratic optimization problems with homogeneous quadratic constraints. Our analysis is motivated by important emerging applications in transmit beamforming for physical layer multicasting and sensor localization in wireless sensor networks. Our generalization (1.1) of the basic problem in [20] is useful, for it shows that the same convex approximation approaches and bounds hold in the case where each multicast receiver is equipped with multiple antennas. This scenario is becoming more pertinent with the emergence of small and cheap multi-antenna mobile terminals. Furthermore, our consideration of the related homogeneous QP maximization problem has direct application to the design of jam-resilient intercept beamformers. In addition to these timely topics, more traditional signal processing design problems can be cast in the same mathematical framework; see [20] for further discussions.

While theoretical worst-case analysis is very useful, empirical analysis of the ratio $\frac{v_{\text{qp}}}{v_{\text{sdp}}}$ through simulations with randomly generated steering vectors $\{h_\ell\}$ is often equally important. In the context of transmit beamforming for multicasting [20] for the case $|\mathcal{I}_i| = 1\ \forall\ i$ (single receiving antenna per subscriber node), simulations have provided the following insights:

- For moderate values of m, n (e.g., $m = 24$, $n = 8$), and independent and identically distributed (i.i.d.) complex-valued circular Gaussian (i.i.d. Rayleigh) entries of the steering vectors $\{h_\ell\}$, the average value of $\frac{v_{\text{qp}}}{v_{\text{sdp}}}$ is under 3 – much lower than the worst-case value predicted by our analysis.

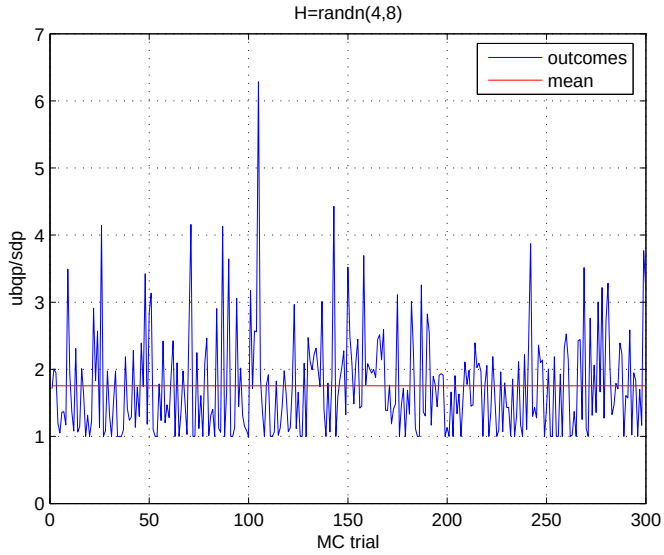


FIG. 6.1. *Upper bound on* $\frac{v_{\rm qp}}{v_{\rm sdp}}$ *for* $m = 8$, $n = 4$, 300 *realizations of real Gaussian i.i.d. steering vector entries, solution constrained to be real.*

- In all generated instances where all steering vectors have positive real and imaginary parts, the ratio $\frac{v_{\rm qp}}{v_{\rm sdp}}$ equals one (with error below 10^{-8}). This is better than what our worst-case analysis predicts for limited phase spread (see Theorem 3.7).
- In experiments with measured VDSL channel data, for which the steering vectors follow a correlated log-normal distribution, $\frac{v_{\rm qp}}{v_{\rm sdp}} = 1$ in over 50% of instances.
- Our analysis shows that the worst-case performance ratio $\frac{v_{\rm qp}}{v_{\rm sdp}}$ is smaller in the complex case than in the real case ($O(m)$ versus $O(m^2)$). Moreover, this remains true with high probability when $v_{\rm qp}$ is replaced by its upper bound

$$v_{\rm ubqp} := \min_{k=1,\ldots,N} \|z^*(\xi^k)\|^2,$$

where $\xi^1, ..., \xi^N$ are generated by N independent trials of the randomization procedure (see Subsections 3.1 and 3.2) and N is taken sufficiently large. In our simulation, we used $N = 30nm$. Figure 6.1 shows our simulation results for the real Gaussian case.[6] It plots $\frac{v_{\rm ubqp}}{v_{\rm sdp}}$ for 300 independent realizations of i.i.d. real-valued Gaussian steering vector entries, for $m = 8$, $n = 4$. Figure 6.2 plots the corresponding histogram. Figures 6.3 and 6.4 show the corresponding results for i.i.d. complex-valued circular Gaussian steering vector entries.[7] Both the mean and the maximum of the upper bound $\frac{v_{\rm ubqp}}{v_{\rm sdp}}$ are lower in the complex case. The simulations indicate that SDP approximation is better in the complex case not only in the worst case but also on average.

[6] Here the SDP solution is constrained to be real-valued, and real Gaussian randomization is used.
[7] Here the SDP solutions are complex-valued, and complex Gaussian randomization is used.

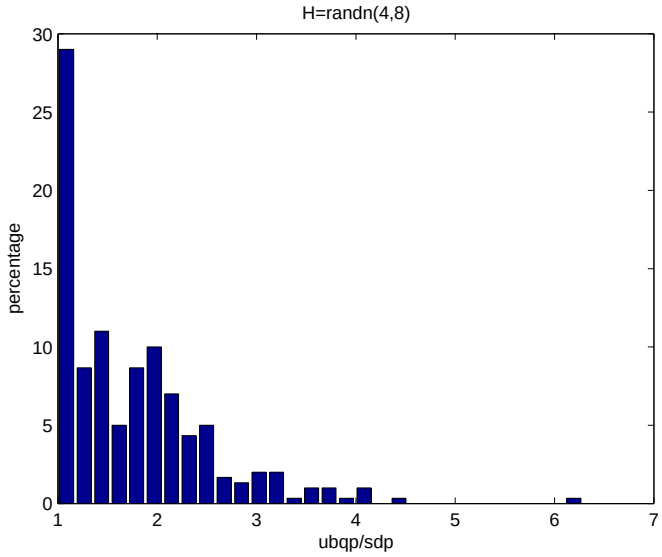


FIG. 6.2. *Histogram of the outcomes in Fig. 1.*

The above empirical (worst-case and average-case) analysis complements our theoretical worst-case analysis of the performance of SDP relaxation for the class of problems considered herein.

Finally, we remark that our worst-case analysis of SDP performance is based on the assumption that the homogeneous quadratic constraints are concave (see (1.1)). Can we extend this analysis to general homogeneous quadratic constraints? The following example in $\mathbb{R}^2$ suggests that this is not possible.

EXAMPLE 3: For any $L > 0$, consider the quadratic optimization problem with homogeneous quadratic constraints:

$$\begin{array}{ll} \min & \|z\|^2 \\ \text{s.t.} & z_2^2 \geq 1, \quad z_1^2 - Lz_1z_2 \geq 1, \quad z_1^2 + Lz_1z_2 \geq 1, \\ & z \in \mathbb{R}^2. \end{array} \tag{6.1}$$

The last two constraints imply $z_1^2 \geq L|z_1||z_2| + 1$ which, together with the first constraint $z_2^2 \geq 1$, yield $z_1^2 \geq L|z_1| + 1$ or, equivalently, $|z_1| \geq (L + \sqrt{L^2+4})/2$. So the optimal value of (6.1) is at least $1 + (L + \sqrt{L^2+4})^2/4$ (and in fact is equal to this). The natural SDP relaxation of (6.1) is

$$\begin{array}{ll} \min & Z_{11} + Z_{22} \\ \text{s.t.} & Z_{22} \geq 1, \quad Z_{11} - LZ_{12} \geq 1, \quad Z_{11} + LZ_{12} \geq 1, \\ & Z \succeq 0. \end{array}$$

Clearly, $Z = I_2$ is a feasible solution (and, in fact, an optimal solution) of this SDP, with an objective value of 2. Therefore, the SDP performance ratio for this example is at least $1/2 + (L + \sqrt{L^2+4})^2/8$, which can be arbitrarily large.

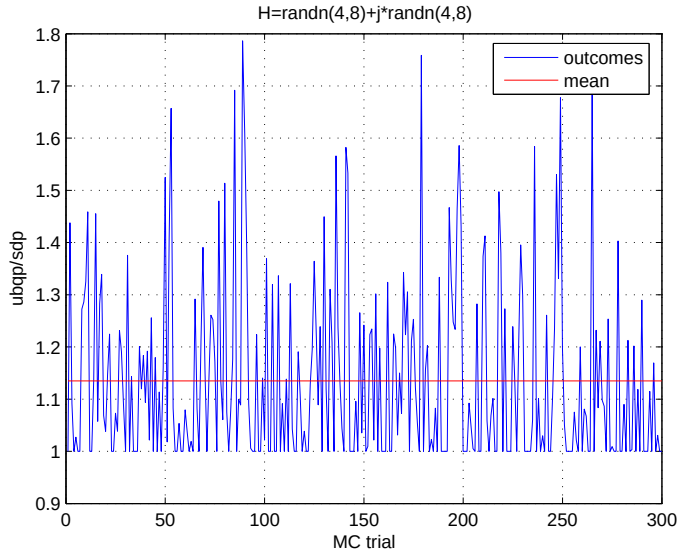


FIG. 6.3. *Upper bound on* $\frac{v_{\rm qp}}{v_{\rm sdp}}$ *for* $m = 8$, $n = 4$, 300 *realizations of complex Gaussian i.i.d. steering vector entries.*

REFERENCES

[1] N. ALON AND A. NAOR, *Approximating the Cut-Norm via Grothendiecks inequality,* Proc. 36th Annual ACM Symp. Theory Comp., Chicago, IL, USA, (2004) pp. 72–80.

[2] H.H. ANDERSEN, M. HØJBJERRE, D. SØRENSEN, AND P.S. ERIKSEN, *Linear and graphical models for the multivariate complex normal distribution,* Lecture Notes in Statistics, Vol. 101, Springer-Verlag, New York, 1995.

[3] A. BEN-TAL AND A. NEMIROVSKI, *On tractable approximations of uncertain linear matrix inequalities affected by interval uncertainty*, SIAM J. Optim., 12 (2002), pp. 811–833.

[4] A. BEN-TAL, A. NEMIROVSKI, AND C. ROOS, *Extended matrix cube theorems with applications to* μ*-theory in control*, Math. Oper. Res., 28 (2003), pp. 497–523.

[5] D. BERTSIMAS AND Y. YE, *Semidefinite relaxations, multivarite normal distribution, and order statistics*, In Handbook of Combinatorial Optimization, D.Z. Du and P.M. Pardalos, eds., Kluwer Academic Publishers, 3 (1998) pp. 1–19.

[6] B. BISWAS AND Y. YE, *Semidefinite programming for ad hoc wireless sensor network localization*, Technical report, Electrical Engineering, Stanford University, Stanford, September 2003. `http://www.stanford.edu/~yyye/`

[7] M.X. GOEMANS AND D.P. WILLIAMSON, *Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming*, J. of ACM., 42 (1995), pp. 1115–1145.

[8] ———, *Approximation algorithms for MAX-3-CUT and other problems via complex semidefinite programming*, J. Comput Syst. Sciences, 68 (2004), pp. 442–470.

[9] Y. HUANG AND S. ZHANG, *Complex matrix decomposition and quadratic programming*, Technical Report SEEM2005-02, Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, 2005. `http://www.se.cuhk.edu.hk/~zhang/#workingpaper`

[10] D.H. JOHNSON AND D.E. DUGEON, *Array Signal Processing: Concepts and Techniques,* Simon & Schuster, 1992.

[11] L. LOVÁSZ AND A. SCHRIJVER, *Cones of matrices and set-functions and* 0-1 *optimization*, SIAM J. Optim.,

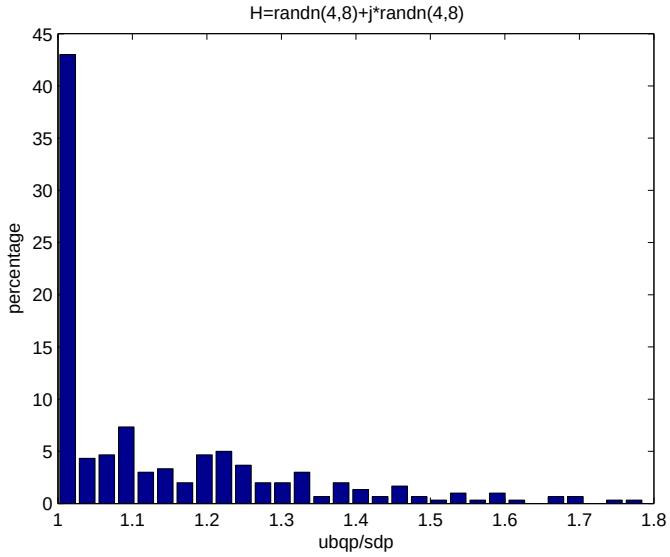


FIG. 6.4. *Histogram of the outcomes in Fig. 3.*

1 (1991), pp. 166–190.

[12] S. MAHAJAN AND H. RAMESH, *Derandomizing approximation algorithms based on semidefinite programming*, SIAM J. Comput., 28 (1999), pp. 1641–1663.

[13] A. MEGRETSKI, *Relaxations of quadratic programs in operator theory and system analysis*, Operator Theory: Adv. and Appl., 129 (2001), pp. 365–392.

[14] A. NEMIROVSKI, C. ROOS AND T. TERLAKY, *On maximization of quadratic form over intersection of ellipsoids with common center*, Math. Prog., 86 (1999), pp. 463–473.

[15] Y. NESTEROV, *Semidefinite relaxation and nonconvex quadratic optimization*, Optim. Methods and Softwares, 9 (1998), pp. 141–160.

[16] Y. NESTEROV, H. WOLKOWICZ, AND Y. YE, *Semidefinite programming relaxations of nonconvex quadratic optimization*, in Handbook of Semidefinite Programming, H. Wolkowicz, R. Saigal, and L. Vandenberghe, eds., Kluwer, Boston, (2000), pp. 360–419.

[17] G. PATAKI, *On the rank of extreme matrices in semidefinite programs and the multiplicity of optimal eigenvalues*, Math. Oper. Res., 23 (1998), pp. 339-358.

[18] G. PATAKI, editor, *Computational semidefinite and second order cone programming: the state of the art*, Math. Prog., 95 (2003).

[19] N.Z. SHOR, *Quadratic optimization problems*, Soviet J. Comput. Systems Sci., 25 (1987), pp. 1–11.

[20] N.D. SIDIROPOULOS, T.N. DAVIDSON AND Z.-Q. LUO, *Transmit beamforming for physical layer multicasting*, IEEE Trans. Signal Processing, 54 (2006), pp. 2239–2251.

[21] A. SO, J. ZHANG, AND Y. YE, *On approximating complex quadratic optimization problems via semidefinite programming relaxations*, Proc. 11th Conf. Integer Prog. and Combinatorial Optim., Lecture Notes in Computer Science, Vol. 3509, M. Junger and V. Kaibel, eds., Springer-Verlag, Berlin, 2005, pp. 125–135.

[22] P. TSENG, *Further results on approximating nonconvex quadratic optimization by semidefinite programming relaxation*, SIAM J. Optim., 14 (2003), pp. 268–283.

[23] Y. YE, *Approximating quadratic programming with bound and quadratic constraints*, Math. Prog., 84 (1999), pp. 219–226.

[24] ———, *Approximating global quadratic optimization with convex quadratic constraints*, J. Global Optim., 15 (1999), pp. 1–17.

[25] S. Zhang, *Quadratic maximization and semidefinite relaxation*, Math. Prog., 87 (2000), pp. 453–465.
[26] S. Zhang and Y. Huang, *Complex quadratic optimization and semidefinite programming*, SIAM J. Optim., 16 (2006), pp. 871–890.

Transmit Beamforming for Physical-Layer Multicasting

Nicholas D. Sidiropoulos, *Senior Member, IEEE*, Timothy N. Davidson, *Member, IEEE*, and Zhi-Quan (Tom) Luo, *Senior Member, IEEE*

***Abstract*—This paper considers the problem of downlink transmit beamforming for wireless transmission and downstream precoding for digital subscriber wireline transmission, in the context of common information broadcasting or multicasting applications wherein channel state information (CSI) is available at the transmitter. Unlike the usual "blind" isotropic broadcasting scenario, the availability of CSI allows transmit optimization. A minimum transmission power criterion is adopted, subject to prescribed minimum received signal-to-noise ratios (SNRs) at each of the intended receivers. A related max–min SNR "fair" problem formulation is also considered subject to a transmitted power constraint. It is proven that both problems are NP-hard; however, suitable reformulation allows the successful application of semidefinite relaxation (SDR) techniques. SDR yields an approximate solution plus a bound on the optimum value of the associated cost/reward. SDR is motivated from a Lagrangian duality perspective, and its performance is assessed via pertinent simulations for the case of Rayleigh fading wireless channels. We find that SDR typically yields solutions that are within 3–4 dB of the optimum, which is often good enough in practice. In several scenarios, SDR generates exact solutions that meet the associated bound on the optimum value. This is illustrated using measured very-high-bit-rate Digital Subscriber line (VDSL) channel data, and far-field beamforming for a uniform linear transmit antenna array.**

***Index Terms*—Broadcasting, convex optimization, downlink beamforming, minimization of total radiation power, multicasting, semidefinite programming, semidefinite relaxation (SDR), very-high-bit-rate Digital Subscriber line (VDSL) precoding.**

I. Introduction

CONSIDER a transmitter that utilizes an antenna array to broadcast information to multiple radio receivers within a certain service area. The traditional approach to broadcasting is to radiate transmission power isotropically, or with a fixed directional pattern. However, future digital video/audio/data broadcasting and multicasting applications are likely to be based on subscription to services; hence, it is plausible to assume that the transmitter can acquire channel state information (CSI) for all its intended receivers.

The goal of this paper is to develop efficient algorithms for the design of broadcasting schemes that exploit this channel information in order to provide better performance than the traditional approaches.

Our design approach is based on providing Quality of Service (QoS) assurance to each of the receivers. Since the received signal-to-noise ratio (SNR) determines the maximum achievable data rate and (essentially) determines the probability of error, it is an effective measure of the QoS. We consider two basic design problems. The first seeks to minimize the total transmission power (and thus leakage to neighboring cochannel transmissions), subject to meeting (potentially different) constraints on the received SNR for each individual intended receiver. The second is a "fair" design problem in which we attempt to maximize the smallest receiver SNR over the intended receivers, subject to a bound on the transmitted power. We will show that both these problems are NP-hard, but we will also show that designs that are close to being optimal can be efficiently obtained by employing semidefinite relaxation (SDR) techniques.

Our designs are initially developed for a wireless broadcast scenario in which each user employs a single receive antenna and the channel is modeled as being flat in frequency and quasi-static in time. However, the designs are also appropriate on a per-tone basis for orthogonal-frequency-division multiplexing (OFDM) and related multicarrier systems, and, as we will show, they can be generalized in a straightforward manner to single-carrier systems transmitting over frequency-selective channels. In addition to wireless systems, applications of the proposed methodology also appear in downstream multicast transmission for multicarrier and single-carrier digital subscriber line (DSL) systems. In this context, (linear) *precoding* of multiple DSL loops in the same binder that wish to subscribe to a common service (e.g., news feed, video-conference, or movie multicast) can be employed to improve QoS and/or reduce far-end crosstalk (FEXT) interference to other loops in the binder. In scenarios in which the customer-premise equipment (CPE) receivers are not physically co-located (as in residential service) or cannot be coordinated (legacy CPE), multiuser decoding of the downstream transmission is not feasible, while transmit precoding is viable. The most important difference between DSL and the wireless

Manuscript received November 1, 2004; revised May 11, 2005. The work of N. D. Sidiropoulos was supported in part by the U.S. ARO under ERO Contract N62558-03-C-0012, the E.U. under FP6 U-BROAD STREP # 506790, and the GSRT. The work of T. N. Davidson was supported in part by the Natural Sciences and Engineering Research Council of Canada and the Canada Research Chairs program. The work of Z.-Q. Luo was supported in part by the U.S. National Science Foundation, Grant no. DMS-0312416. An earlier version of part of this work appears in *Proceedings of the IEEE SAM 2004* workshop, vol. 1, pp. 489–493, Sitges, Barcelona, Spain, July 18–21, 2004. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Daniel Fuhrman.

N. D. Sidiropouplos is with the Department of Electronic and Computer Engineering, Technical University of Crete, 73100 Chania-Crete, Greece (e-mail: nikos@telecom.tuc.gr).

T. N. Davidson is with the Department of Electrical and Computer Engineering, McMaster University, Hamilton, ON L8S 4K1, Canada (e-mail: davidson@mcmaster.ca).

Z.-Q. Luo is with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455 USA (e-mail: luozq@ece.umn.edu).

Digital Object Identifier 10.1109/TSP.2006.872578

multicast scenario is that DSL channels are diagonally dominated. Still, exploiting the crosstalk coupling to reduce FEXT levels to other loops in the binder can provide significant performance gains, especially if (cooperative or competitive) power control is implemented.

It is interesting to note that, as of today, Internet multicasting (using IP's Multicast Backbone—MBone) is performed at the *network layer*, e.g., via packet-level flooding or spanning-tree access of the participant nodes and any intermediate nodes needed to access the participants. To complement that approach, what we advocate herein can be interpreted as judicious *physical layer multicasting*, that is, enabled by i) the availability of multiple transmitting elements; ii) exploiting opportunities for joint beamforming/precoding; and iii) the availability of CSI at the transmitting node or one of its proxies. This is a cross-layer optimization approach that exploits information available at the physical layer to reduce relay retransmissions at the network layer, thus providing congestion relief and QoS guarantees.

Notation: We use lowercase boldface letters to denote column vectors and uppercase bold letters to denote matrices. $(\cdot)^T$ denotes transpose, while $(\cdot)^H$ denotes Hermitian (conjugate) transpose. $\mathrm{Re}\{\cdot\}$ extracts the real part of its argument, and $\mathrm{Im}\{\cdot\}$ the imaginary part.

II. Data Model and Problem Statement

Consider a wireless scenario incorporating a single transmitter with N antenna elements and M receivers each with a single antenna. Let $\mathbf{h}_i$ denote the $N \times 1$ complex vector that models the propagation loss and phase shift of the frequency-flat quasi-static channel from each transmit antenna to the receive antenna of user $i \in \{1, \ldots, M\}$, and let $\mathbf{w}^H$ denote the beamforming weight vector applied to the N transmitting antenna elements. If the signal to be transmitted is zero-mean and white with unit variance, and if the noise[1] at receiver i is zero-mean and white with variance σ_i^2, then the receiver SNR for the ith user is $|\mathbf{w}^H\mathbf{h}_i|^2/\sigma_i^2$. Let $\rho_{\min,i}$ be the prescribed minimum SNR for the ith user and define the normalized channel vectors $\tilde{\mathbf{h}}_i := \mathbf{h}_i/\sqrt{\rho_{\min,i}\sigma_i^2}$. Then $|\mathbf{w}^H\mathbf{h}_i|^2/\sigma_i^2 \geq \rho_{\min,i} \Leftrightarrow |\mathbf{w}^H\tilde{\mathbf{h}}_i|^2 \geq 1$. Therefore, the design of the beamformer that minimizes the transmitted power, subject to (possibly different) constraints on the received SNR of each user, can be written as

$$\mathcal{Q}: \quad \min_{\mathbf{w}\in\mathbb{C}^N} \|\mathbf{w}\|_2^2 \quad \text{subject to: } |\mathbf{w}^H\tilde{\mathbf{h}}_i|^2 \geq 1, \quad i \in \{1, \cdots, M\}.$$

We will denote an instance of problem $\mathcal{Q}$ as $\mathcal{Q}(\{\tilde{\mathbf{h}}_i\}_{i=1}^M)$, keeping in mind that $\tilde{\mathbf{h}}_i = \mathbf{h}_i/\sqrt{\rho_{\min,i}\sigma_i^2}$.

Remark 1: One could think of imposing the stricter constraints $\mathbf{w}^H\tilde{\mathbf{h}}_i = 1, \ \forall i$ in order to avoid the need for single-tap equalization at the receivers. However, we are interested in the practically important case of $M > N$, wherein the stricter constraints generically yield an overdetermined system of equations, and thus an infeasible problem. On the other hand, it is easy to see that problem $\mathcal{Q}$ is always feasible, provided of course that none of the channel vectors is identically zero.

Problem $\mathcal{Q}$ is formulated under the assumption that the design center (usually the transmitter) has knowledge of the channel vector $\mathbf{h}_i$ (and the noise variance σ_i^2) for each user. This can be accomplished in a straightforward manner in fixed wireless systems and time-division-duplex (TDD) systems. In other systems, it can be accomplished through the use of beacon signals, periodically transmitted from the broadcasting station (and typically embedded in the transmission). The receiving radios can then feed back their CSI through a feedback channel. For the purposes of this paper, we will assume that the design center has perfect knowledge of the channel vectors, but extensions to cases of imperfect knowledge are under development.

Problem $\mathcal{Q}$ is a quadratically constrained quadratic programming (QCQP) problem, but unfortunately the constraints are not convex.[2] Nonconvexity, per se, does not mean that the problem is difficult to solve; however, we have the following claim, whose proof can be found in Appendix I.

Claim 1: The QoS problem $\mathcal{Q}$ is NP-hard.

The implication of Claim 1 is that if an algorithm could solve an arbitrary instance of problem $\mathcal{Q}$ in polynomial time, it would then be possible to solve a whole class of computationally very difficult problems in polynomial time [4]. The current scientific consensus indicates that this is unlikely.

A. Review of Pertinent Prior Art

The above problem is reminiscent of some closely related problems. For $M = 1$, the optimum $\mathbf{w}$ is a matched filter. When the scaled channel vectors $\tilde{\mathbf{h}}_i$ span a ball or ellipsoid about a "nominal" channel vector,[3] the problem can be transformed *exactly* into a second-order cone program, and hence can be efficiently solved [13]. Unfortunately, this transformation cannot be employed in the case of finitely many channel vectors (intended receivers).

Another closely related work is that in [1] (and references therein), which considers the problem of multiuser transmit beamforming for the cellular downlink. The key difference between [1] and our formulation is that the authors of [1] consider the transmission of independent information to each of the downlink users, whereas we focus on (common information) multicast. The mathematical problems are not equivalent. A fundamental difference is that our problem is NP-hard, whereas the formulation in [1] can be efficiently solved. To further appreciate the difference intuitively, we point out that in the generic case of our formulation most of the SNR constraints will be inactive at the optimum (i.e., most of the constraints will be oversatisfied). Consider, for example, the case of two closely located receivers with different SNR requirements: one of the two associated constraints will be oversatisfied at the optimum. On the other hand, it is proven in [1] that in the formulation of [1] the constraints are always met with equality at the optimum. The important common denominator of our work and [1] is the use of semidefinite programming tools.

[1]The noise may include unmodeled interference.

[2]This is easy to see for $N = 1$, in which case each constraint requires that the magnitude of w be greater than a constant.

[3]This implies a continuum of intended receivers.

Transmit beamforming for the dissemination of common information to multiple users has been considered in the Ph.D. dissertation of Lopez [7, ch. 5]. Lopez proposed maximizing the sum of received SNRs, which is equivalent to maximizing the average SNR over all users. This formulation leads to a principal component computational problem for the optimum beamformer, which is relatively simple to solve. The drawback is that quality of service cannot be guaranteed to all users in this way. This is important, because the weakest user link determines the common information rate. Still, the work of Lopez is the closest in spirit to ours, and for this reason we will include the maximum average SNR approach in our performance evaluations in Section VIII (see Table V).

III. Relaxation

Toward solving our problem, we first recast it as follows:

$$\min_{\mathbf{w}} \ \mathrm{trace}(\mathbf{w}\mathbf{w}^H)$$
$$\text{subject to: } \mathrm{trace}(\mathbf{w}\mathbf{w}^H\mathbf{Q}_i) \geq 1, \quad i \in \{1, \cdots, M\}$$

where we have used the fact that $\tilde{\mathbf{h}}_i^H \mathbf{w}\mathbf{w}^H \tilde{\mathbf{h}}_i = \mathrm{trace}(\tilde{\mathbf{h}}_i^H \mathbf{w}\mathbf{w}^H \tilde{\mathbf{h}}_i) = \mathrm{trace}(\mathbf{w}\mathbf{w}^H \tilde{\mathbf{h}}_i \tilde{\mathbf{h}}_i^H)$, and we have defined $\mathbf{Q}_i := \tilde{\mathbf{h}}_i \tilde{\mathbf{h}}_i^H$. Now consider the following reformulation of the problem:

$$\min_{\mathbf{X} \in \mathbb{C}^{N \times N}} \quad \mathrm{trace}(\mathbf{X})$$
$$\text{subject to: } \mathrm{trace}(\mathbf{X}\mathbf{Q}_i) \geq 1, \quad i \in \{1, \cdots, M\}$$
$$\mathbf{X} \succeq \mathbf{0}$$
$$\mathrm{rank}(\mathbf{X}) = 1$$

where now $\mathbf{X}$ is an $N \times N$ complex matrix, and the inequality $\mathbf{X} \succeq 0$ means that the matrix $\mathbf{X}$ is symmetric positive semidefinite. Note that, in the above *equivalent* formulation of our problem, the cost function is linear in $\mathbf{X}$; the trace constraints are linear inequalities in $\mathbf{X}$, and the set of symmetric positive semidefinite matrices is convex; however, the rank constraint on $\mathbf{X}$ is not convex.[4] The important observation is that the above problem is in a form suitable for semidefinite relaxation (SDR) (see, e.g., [9] and references therein); that is, dropping the rank-one constraint, one obtains the relaxed problem

$$\min_{\mathbf{X} \in \mathbb{C}^{N \times N}} \quad \mathrm{trace}(\mathbf{X})$$
$$\text{subject to: } \mathrm{trace}(\mathbf{X}\mathbf{Q}_i) \geq 1, \ i \in \{1, \cdots, M\}, \ \text{and} \ \mathbf{X} \succeq \mathbf{0}$$

which is a semidefinite programming problem (SDP), albeit not yet in standard form. In order to put it in standard form, we add M "slack" variables $s_i \in \mathbb{R}, i \in \{1, \cdots, M\}$, one for each trace constraint. In this way, we obtain the program

$\mathcal{Q}_r$:

$$\min_{\mathbf{X} \in \mathbb{C}^{N \times N}, s_i \in \mathbb{R}} \quad \mathrm{vec}(\mathbf{I}_N)^T \mathrm{vec}(\mathbf{X})$$
$$\text{s.t.: } \mathrm{vec}\left(\mathbf{Q}_i^T\right)^T \mathrm{vec}(\mathbf{X}) - s_i = 1, \ i \in \{1, \cdots, M\}$$
$$s_i \geq 0, i \in \{1, \cdots, M\}, \ \text{and} \ \mathbf{X} \succeq \mathbf{0}$$

which is now expressed in a standard form used by SDP solvers, such as SeDuMi [11]. Here, $\mathbf{I}_N$ is the identity matrix of size $N \times N$.

SDP problems can be efficiently solved using interior point methods, at a complexity cost that is at most $O((M + N^2)^{3.5})$ and is usually much less. SeDuMi [11] is a MATLAB implementation of modern interior point methods for SDP that is particularly efficient for up to moderate-sized problems, as is the case in our context. Typical run times for realistic choices of N and M are under 1/10 s, on a typical personal computer.

IV. Algorithm

Due to the relaxation, the matrix $\mathbf{X}_{\mathrm{opt}}$ obtained by solving the SDP in Problem $\mathcal{Q}_r$ will not be rank one in general. If it is, then its principal component will be the optimal solution to the original problem. If not, then $\mathrm{trace}(\mathbf{X}_{\mathrm{opt}})$ is a lower bound on the power needed to satisfy the constraints. This comes from the fact that we have removed one of the original problem's constraints. Researchers in optimization have recently developed ways of generating good solutions to the original problem, $\mathcal{Q}$, from $\mathbf{X}_{\mathrm{opt}}$, [9], [12], [15]. This process is based on *randomization*: using $\mathbf{X}_{\mathrm{opt}}$ to generate a set of candidate weight vectors, $\{\mathbf{w}_\ell\}$, from which the "best" solution will be selected. We consider three methods for generating the $\mathbf{w}_\ell$'s, which have been designed so that their computational cost is negligible compared to that of computing $\mathbf{X}_{\mathrm{opt}}$. (For consistency, the principal component is also included in the set of candidates.) In the first method (`randA`), we calculate the eigen-decomposition of $\mathbf{X}_{\mathrm{opt}} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^H$ and choose $\mathbf{w}_\ell$ such that $\mathbf{w}_\ell = \mathbf{U}\mathbf{\Sigma}^{1/2}\mathbf{e}_\ell$, where the elements of $\mathbf{e}_\ell$ are independent random variables, uniformly distributed on the unit circle in the complex plane; i.e., $[\mathbf{e}_\ell]_i = e^{j\theta_{\ell,i}}$, where the $\theta_{\ell,i}$ are independent and uniformly distributed on $[0, 2\pi)$. This ensures that $\mathbf{w}_\ell^H \mathbf{w}_\ell = \mathrm{trace}(\mathbf{X}_{\mathrm{opt}})$, irrespective of the particular realization of $\mathbf{e}_\ell$. In the second method (`randB`), inspired by Tseng [12], we choose $\mathbf{w}_\ell$ such that $[\mathbf{w}_\ell]_i = \sqrt{[\mathbf{X}_{\mathrm{opt}}]_{ii}}\,[\mathbf{e}_\ell]_i$, which ensures that $|[\mathbf{w}_\ell]_i|^2 = [\mathbf{X}_{\mathrm{opt}}]_{ii}$. The third method (`randC`), motivated by successful applications in related QCQP problems [8], uses $\mathbf{w}_\ell = \mathbf{U}\mathbf{\Sigma}^{1/2}\mathbf{v}_\ell$, where $\mathbf{v}_\ell$ is a vector of zero-mean, unit-variance complex circularly symmetric uncorrelated Gaussian random variables. This ensures that $E[\mathbf{w}_\ell \mathbf{w}_\ell^H] = \mathbf{X}_{\mathrm{opt}}$ [8].

For both `randA` and `randB`, $\|\mathbf{w}_\ell\|_2^2 = \mathrm{trace}(\mathbf{X}_{\mathrm{opt}})$, and hence when $\mathrm{rank}(\mathbf{X}_{\mathrm{opt}}) > 1$, at least one of the constraints $|\mathbf{w}_\ell^H \mathbf{h}_i|^2 \geq c_i$ will be violated.[5] However, a feasible weight vector can be found by simply scaling $\mathbf{w}_\ell$ so that all the constraints are satisfied. Under `randC`, $\|\mathbf{w}_\ell\|_2^2$ depends on the particular realization of $\mathbf{v}_\ell$, but again the resulting $\mathbf{w}_\ell$ can be scaled to the minimum length necessary to satisfy the constraints. The "best" of these randomly generated weight vectors is the one that requires the smallest scaling. For convenience, we have summarized the algorithm in Table I, which includes a simple MATLAB interface to SeDuMi [11] for the solution of the semidefinite relaxation, $\mathcal{Q}_r$. We point out that we have not yet been able to obtain theoretical *a priori* bounds on the extent

[4] The sum of two rank-one matrices has generic rank two.

[5] Recall that because of the relaxation, $\mathrm{trace}(\mathbf{X}_{\mathrm{opt}})$ is a lower bound on the energy of the optimal weight vector for the original problem.

TABLE I
BROADCAST QoS BEAMFORMING VIA SDR: ALGORITHM

• Solve the relaxed problem:

A simple MATLAB interface for SeDuMi is as follows:

```
% Input Data:
% H: N by M, columns are scaled channels h_i/sqrt(rho_{min,i} sigma_i^2)
% Output Data:
% Xopt: the solution to the SDR
vecQs = [];
for i=1:M,
  Qi = H(:,i)*H(:,i)';
  vecQs = [vecQs, vec(Qi.')];
end;
A=[-eye(M), vecQs.'];
b=ones(M,1);
c=[zeros(M,1); vec(eye(N))];
K.l=M; K.s=N; K.scomplex=1;
[x_opt,y_opt,info]=sedumi(A,b,c,K);
Xopt=mat(x_opt(M+1:end));
```

• Randomization:

Use randA, and/or randB, randC to generate the candidates, w_ell.

For each w_ell, find the most violated constraint.

Scale w_ell so that that constraint is satisfied with equality.

Pick the w_ell with the smallest norm.

of the suboptimality of solutions generated in this way, but our simulation results are quite encouraging.

V. MAX–MIN FAIR BEAMFORMING

We now consider the related problem of maximizing the minimum received SNR over all receivers, subject to a bound on the transmitted power. That is

$$\mathcal{F}: \quad \max_{\mathbf{w}\in\mathbb{C}^N} \min_i \left\{ \left| \frac{\mathbf{w}^H\mathbf{h}_i}{\sigma_i} \right|^2 \right\}_{i=1}^M \quad \text{subject to: } \|\mathbf{w}\|_2^2 \le P$$

It is easy to see that the constraint in problem $\mathcal{F}$ should be met with equality at an optimum, for otherwise $\mathbf{w}$ could be scaled up, thereby improving the objective and contradicting optimality. Thus, we can focus on the equality-constrained problem. With a scaling of the optimization variable $\mathbf{w} = \sqrt{P}\tilde{\mathbf{w}}$, the equality-constrained problem can be equivalently written as

$$P \max_{\tilde{\mathbf{w}}} \min_i \left\{ \left| \frac{\tilde{\mathbf{w}}^H\mathbf{h}_i}{\sigma_i} \right|^2 \right\}_{i=1}^M \quad \text{subject to: } \|\tilde{\mathbf{w}}\|_2^2 = 1.$$

It is clear that P is immaterial with respect to optimization; the solution scales up with $\sqrt{P}$, while the optimum value scales up with P. We will denote an instance of problem $\mathcal{F}$ as $\mathcal{F}(\{\mathbf{h}_i/\sigma_i\}_{i=1}^M, P)$. Let $\mathbf{w}_q$ be a solution to $\mathcal{Q}(\{\tilde{\mathbf{h}}_i\}_{i=1}^M)$, and P_q the associated minimum transmitted power. Consider $\mathcal{F}(\{\tilde{\mathbf{h}}_i\}_{i=1}^M, P_q)$, that is

$$\max_{\mathbf{w}} \min_i \left\{ |\mathbf{w}^H\tilde{\mathbf{h}}_i|^2 \right\}_{i=1}^M \quad \text{subject to: } \|\mathbf{w}\|_2^2 = P_q$$

and let $\mathbf{w}_f$ denote an optimal solution. Since $\mathbf{w}_q$ already attains $\left|\mathbf{w}_q^H\tilde{\mathbf{h}}_i\right|^2 \ge 1, \forall i$, it follows that $\left|\mathbf{w}_f^H\tilde{\mathbf{h}}_i\right|^2 \ge 1, \forall i$. Hence, $\mathbf{w}_f$ also satisfies the constraints of the QoS formulation, and at the same power as $\mathbf{w}_q$. It follows that $\mathbf{w}_f$ is equivalent to $\mathbf{w}_q$. This shows Claim 2.

Claim 2: $\mathcal{F}\left(\{\tilde{\mathbf{h}}_i\}_{i=1}^M, P\right)$ is equivalent to $\mathcal{Q}\left(\{\tilde{\mathbf{h}}_i\}_{i=1}^M\right)$ up to scaling. In the special case that $\rho_{\min,i} = \rho_{\min}, \forall i$, we have that $\tilde{\mathbf{h}}_i = (\mathbf{h}_i/\sigma_i)/\sqrt{\rho_{\min}}, \forall i$, and hence $\mathcal{F}\left(\{\mathbf{h}_i/\sigma_i\}_{i=1}^M, P\right)$ is equivalent to $\mathcal{Q}\left(\{\tilde{\mathbf{h}}_i\}_{i=1}^M\right)$ up to scaling.

Corollary 1: One way to solve the max–min fair problem $\mathcal{F}\left(\{\mathbf{h}_i/\sigma_i\}_{i=1}^M, P\right)$ is to solve the QoS problem $\mathcal{Q}\left(\{\mathbf{h}_i/\sigma_i\}_{i=1}^M\right)$, then scale the resulting solution to the desired power P. Conversely, scaling the solution of $\mathcal{F}\left(\{\tilde{\mathbf{h}}_i\}_{i=1}^M, 1\right)$ yields a solution to $\mathcal{Q}\left(\{\tilde{\mathbf{h}}_i\}_{i=1}^M\right)$, even in the case of unequal $\rho_{\min,i}$.

Remark 2: It is important not to lose sight of the fact that $\mathcal{F}\left(\{\mathbf{h}_i/\sigma_i\}_{i=1}^M, P\right)$ *is not equivalent up to scaling to* $\mathcal{Q}\left(\{\tilde{\mathbf{h}}_i\}_{i=1}^M\right)$ when the $\rho_{\min,i}$'s are unequal. This can be intuitively appreciated by noting that the max–min fair formulation aims to maximize the minimum received SNR, without regard to the *individual* SNR constraints. The QoS formulation, on the other hand, explicitly guarantees the prescribed minimum SNR level at each node.

From the above, and Claim 1, Claim 3 follows.

Claim 3: The max-min fair problem $\mathcal{F}$ is NP-hard.

If the QoS problem could be solved *exactly*, there would have been no need for a separate algorithm for the max–min fair problem. However, we can only solve the QoS problem approximately (cf., randomization postprocessing of the generally higher rank solution). Due to this, it is of interest to develop a customized SDR algorithm directly for the max–min fair problem. Using the fact that $\mathbf{h}_i^H\mathbf{w}\mathbf{w}^H\mathbf{h}_i = \mathrm{trace}(\mathbf{w}\mathbf{w}^H\mathbf{h}_i\mathbf{h}_i^H)$, and defining $\check{\mathbf{Q}}_i := \mathbf{h}_i\mathbf{h}_i^H/\sigma_i^2$, we recast the max–min fair problem as follows:

$$\max_{\mathbf{X}\in\mathbb{C}^{N\times N}} \min_{i=1,\cdots,M} \mathrm{trace}(\mathbf{X}\check{\mathbf{Q}}_i) \quad \text{subject to: } \mathrm{trace}(\mathbf{X}) = P, \quad \mathbf{X} \succeq \mathbf{0} \quad \mathrm{rank}(\mathbf{X}) = 1.$$

Dropping the rank constraint, we obtain the relaxation

$$\max_{\mathbf{X}\in\mathbb{C}^{N\times N}} \min_{i=1,\cdots,M} \mathrm{trace}(\mathbf{X}\check{\mathbf{Q}}_i) \quad \text{subject to: } \mathrm{trace}(\mathbf{X}) = P, \quad \mathbf{X} \succeq \mathbf{0}.$$

Introducing an additional variable, t, this relaxation can be equivalently written as

$$\max_{\mathbf{X}\in\mathbb{C}^{N\times N}, t\in\mathbb{R}} \quad t \quad \text{subject to: } \mathrm{trace}(\mathbf{X}\check{\mathbf{Q}}_i) \ge t, \quad \forall i \in \{1,\cdots,M\} \quad \mathrm{trace}(\mathbf{X}) = P, \quad \mathbf{X} \succeq \mathbf{0}.$$

TABLE II
BROADCAST MAX–MIN BEAMFORMING VIA SDR: ALGORITHM

• Solve the relaxed problem:

A suitable MATLAB interface for SeDuMi is as follows:

```
% Input Data:
% H: N by M, columns are scaled channels h_i/σ_i
% P: scalar, the total transmit power constraint
% Output Data:
% Xopt: the solution to the SDR
% t_opt: the minimum objective value of the SDR
vecQs = [];
for i=1:M,
   Qi = H(:,i)*H(:,i)';
   vecQs = [vecQs vec(Qi.')];
end;
A1=[-ones(M,1); 0];
A2=[-eye(M); zeros(1,M)];
A3=[vecQs.'; vec(eye(N)).'];
A=[A1 A2 A3];
b=[zeros(M,1); P];
c = [-1; zeros(M+N*N,1)];
K.l=M+1; K.s=N; K.scomplex=1;
[x_opt,y_opt,info]=sedumi(A,b,c,K);
Xopt=mat(x_opt(M+2:end));
t_opt = x_opt(1)
```

• Randomization:

Use randA, and/or randB, randC to generate the candidates `w_ell`.

Scale each `w_ell` to norm `P`.

Pick the one that yields the largest `min(abs(w_ell'*H))`.

Further introducing M nonnegative real slack variables, one for each inequality constraint, we convert the problem to an equivalent one involving only equality, nonnegativity, and positive-semidefinite constraints

$$\mathcal{F}_r: \quad \min_{\substack{t,s_i\in\mathbb{R}\\ \mathbf{X}\in\mathbb{C}^{N\times N}}} \; -t$$
$$\text{subject to: } -t-s_i+\text{vec}\left(\check{\mathbf{Q}}_i^T\right)^T\text{vec}(\mathbf{X})=0,\ \forall i,$$
$$\text{vec}(\mathbf{I}_N)^T\text{vec}(\mathbf{X})=P$$
$$\mathbf{X}\succeq\mathbf{0},\ s_i\geq 0,\ \forall i,\ t\geq 0.$$

This problem is formatted for direct solution via SeDuMi [11]. Table II provides a suitable MATLAB interface for solving this relaxation. Postprocessing of the solution of the relaxed problem to approximate the solution of the original max–min-fair problem can be accomplished using `randA`, `randB`, and `randC`, but the selection criterion is different (see Table II).

In closing this section, we would like to point out connections between problems $\mathcal{F}$ and $\mathcal{F}_r$ and the problem of maximizing the common mutual information of the (nondegraded) Gaussian broadcast channel in which the transmitter has N antennas and each of the M (noncooperative) receivers has a single antenna. If $\mathbf{X}$ denotes the covariance of the transmitted signal, then the maximum achievable common information rate (in the sense of Shannon) can be written as (see, e.g., [6] and references therein)

$$C:=\max_{\substack{\mathbf{X}\succeq\mathbf{0},\\ \text{trace}(\mathbf{X})\leq P}}\ \min_i\ \left\{\log\left(1+\frac{\mathbf{h}_i^H\mathbf{X}\mathbf{h}_i}{\sigma_i^2}\right)\right\}_{i=1}^{M}.$$

Alternatively, we can rewrite this max–min problem as

$$\max\ t$$
$$\text{subject to } \mathbf{X}\succeq\mathbf{0},\ \text{trace}(\mathbf{X})\leq P,\ \log\left(1+\frac{\mathbf{h}_i^H\mathbf{X}\mathbf{h}_i}{\sigma_i^2}\right)\geq t,\quad \forall i.$$

By the monotonicity of the "log" function, the above problem is further equivalent to

$$\max\tilde{t}$$
$$\text{subject to } \mathbf{X}\succeq\mathbf{0},\ \text{trace}(\mathbf{X})\leq P,\ \frac{\mathbf{h}_i^H\mathbf{X}\mathbf{h}_i}{\sigma_i^2}\geq\tilde{t},\quad \forall i$$

in the sense that they yield the same optimal transmit covariance matrix $\mathbf{X}$. The latter problem is identical to problem $\mathcal{F}_r$. In other words, the semidefinite relaxation of problem $\mathcal{F}$ actually yields a transmit covariance matrix that achieves the maximum common information rate C. In a similar manner, we can argue that the rank-one transmit covariance matrix obtained from problem $\mathcal{F}$ achieves the maximum common information rate under the restriction that beamforming is employed. However, the latter rate can be significantly lower than C for a large number of users [6]. Nonetheless, from a practical perspective, beamforming is attractive because it is simple to implement,[6] requiring only a single standard additive white Gaussian noise (AWGN) channel encoder and decoder. In contrast, achieving the maximum common information rate C in general requires higher rank transmit covariance matrix $\mathbf{X}$. In that case, a weighted sum of multiple independent signals is transmitted from each antenna, with each independent signal requiring a separate AWGN channel encoder and decoder. Hence, the beamforming strategy considered in this paper trades off a potential reduction in the maximum common information rate for implementation simplicity.

VI. CASE OF FREQUENCY-SELECTIVE MULTIPATH

Although we have focused our attention so far on frequency-flat fading channels, the situation is quite similar in the case of spatial beamforming[7] for common information transmission over frequency-selective (intersymbol interference) channels. Let $\mathbf{h}_i^{(\ell)}$ denote the ℓth $N\times 1$ vector tap of the baseband-equivalent discrete-time impulse response of the multipath channel between the transmitter antenna array and

[6]A properly weighted common temporal signal is transmitted from each antenna.

[7]It is perhaps worth emphasizing that, while *space–time precoding* would generally be preferable from a performance point of view when the channels are time dispersive, we (continue to) consider *spatial beamforming* only in this section. This is motivated from a complexity point of view. Space–time multicast precoding is an interesting topic for future research.

the (single) receive antenna of receiver i. Assume that delay spread is limited[8] to L nonzero vector channel taps. Define the channel matrix for the ith receiver as

$$\mathbf{H}_i := \left[\mathbf{h}_i^{(0)}, \cdots, \mathbf{h}_i^{(L-1)}\right].$$

Beamforming the transmit array with a fixed (time-invariant) $\mathbf{w}^H$ yields a scalar equivalent channel from the viewpoint of the ith receiver, whose scalar taps are given by

$$\left[\bar{h}_i^{(0)}, \cdots, \bar{h}_i^{(L-1)}\right]^T = \left[\mathbf{w}^H\mathbf{h}_i^{(0)}, \cdots, \mathbf{w}^H\mathbf{h}_i^{(L-1)}\right]^T$$

or, in vector form

$$\bar{\mathbf{h}}_i^T = \mathbf{w}^H\mathbf{H}_i.$$

Now, if a Viterbi equalizer is used for sequence estimation at the receiver, then the parameter that determines performance is [3]

$$\frac{\|\bar{\mathbf{h}}_i\|_2^2}{\sigma_i^2} = \frac{\mathbf{w}^H\mathbf{H}_i\mathbf{H}_i^H\mathbf{w}}{\sigma_i^2} = \mathrm{trace}(\mathbf{w}\mathbf{w}^H\bar{\mathbf{Q}}_i),$$

where now $\bar{\mathbf{Q}}_i := \mathbf{H}_i\mathbf{H}_i^H/\sigma_i^2$ and is generally of higher rank than before, but otherwise things remain conceptually the same. In particular, the relaxations $\mathcal{Q}_r$ and $\mathcal{F}_r$ and the algorithms in Tables I and II can be employed as they were in the frequency-flat case—only the definition of the input matrices changes.

VII. Insights Afforded via Duality

Let us return to our original problem $\mathcal{Q}$, as follows:

$$\min_{\mathbf{w}\in\mathbb{C}^N} \|\mathbf{w}\|_2^2$$
$$\text{subject to: } \left|\mathbf{w}^H\tilde{\mathbf{h}}_i\right|^2 \geq 1, \quad i \in \{1, \cdots, M\}.$$

We will now gain some insight into the quality of the solution generated by the semidefinite relaxation of $\mathcal{Q}$ using bounds obtained from duality. For convenience, we first convert the problem to real-valued form; this yields a $2N \times 1$ vector of real variables $\mathbf{x} := [\mathrm{Re}\{\mathbf{w}\}^T\mathrm{Im}\{\mathbf{w}\}^T]^T$, and the $\mathbf{Q}_i$'s are now $2N \times 2N$ symmetric matrices of rank 2: $\mathbf{Q}_i := \mathbf{g}_i\mathbf{g}_i^T + \bar{\mathbf{g}}_i\bar{\mathbf{g}}_i^T$, where $\mathbf{g}_i := [\mathrm{Re}\{\tilde{\mathbf{h}}_i\}^T\mathrm{Im}\{\tilde{\mathbf{h}}_i\}^T]^T$ and $\bar{\mathbf{g}}_i := [\mathrm{Im}\{\tilde{\mathbf{h}}_i\}^T - \mathrm{Re}\{\tilde{\mathbf{h}}_i\}^T]^T$. Problem $\mathcal{Q}$ can then be rewritten as

$$\mathcal{P}: \quad \min_{\mathbf{x}\in\mathbb{R}^{2N\times 2N}} \mathbf{x}^T\mathbf{x}$$
$$\text{subject to: } \mathbf{x}^T\mathbf{Q}_i\mathbf{x} \geq 1,\ i \in \{1, \cdots, M\}.$$

The Lagrangian of problem $\mathcal{P}$ is [2]

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) = \mathbf{x}^T\mathbf{x} + \sum_{i=1}^M \lambda_i(1 - \mathbf{x}^T\mathbf{Q}_i\mathbf{x})$$
$$= \mathbf{x}^T\left(\mathbf{I} - \sum_{i=1}^M \lambda_i\mathbf{Q}_i\right)\mathbf{x} + \sum_{i=1}^M \lambda_i$$

and the dual problem is

$$\max_{\boldsymbol{\lambda}\succeq\mathbf{0}} \min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda})$$

[8]or, essentially limited; the remaining taps can be treated as interference.

where $\boldsymbol{\lambda} \succeq \mathbf{0}$ denotes $\lambda_i \geq 0$. If the symmetric matrix $(\mathbf{I} - \sum_{i=1}^M \lambda_i\mathbf{Q}_i)$ has a negative eigenvalue, then it is easy to see that the quadratic term in $\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda})$ is unbounded from below (e.g., choose $\mathbf{x}$ proportional to the corresponding eigenvector). If, on the other hand, all eigenvalues are greater than or equal to zero, then the said matrix is positive semidefinite and the minimum over $\mathbf{x}$ is attained, e.g., at $\mathbf{x} = \mathbf{0}$. This yields the following equivalent of the dual problem:

$$\max_{\lambda_i\in\mathbb{R}} \sum_{i=1}^M \lambda_i$$
$$\text{subject to: } \mathbf{I} - \sum_{i=1}^M \lambda_i\mathbf{Q}_i \succeq \mathbf{0}$$
$$\lambda_i \geq 0,\ i = 1, \ldots, M$$

which is a semidefinite program.

The dual problem is interesting, because the maximum of the dual problem is a lower bound on the minimum of the original (primal) problem [2]. The dual problem is convex by virtue of its definition, however the particular dual studied above is special in the sense that optimization over $\mathbf{x}$ for a given $\boldsymbol{\lambda}$ can be carried out analytically, and the residual $\boldsymbol{\lambda}$-optimization problem is an SDP. This means that we can solve the dual problem and thus obtain the tightest bound obtainable via duality. This duality-derived bound can be compared to the SDR bound we used earlier. Let $\mathcal{D}(\cdot)$ denote the dual of a given optimization problem, and let $\mathcal{R}(\mathcal{P})$ denote the semidefinite relaxation of $\mathcal{P}$, obtained by dropping the associated rank-one constraint. Furthermore, let $\beta(\cdot)$ denote the optimal value of a given optimization problem.

Theorem 1: [14, pp. 403–404] $\mathcal{D}(\mathcal{D}(\mathcal{P})) = \mathcal{R}(\mathcal{P})$ and $\beta(\mathcal{R}(\mathcal{P})) = \beta(\mathcal{D}(\mathcal{P}))$.

More specifically, Theorem 1 states that the dual of the dual of $\mathcal{P}$ is the SDR of $\mathcal{P}$ and that the optimal objective value of the SDR of $\mathcal{P}$ is the same as the optimal objective value of the dual of $\mathcal{P}$. Hence, SDR yields the same lower bound on the optimal value of $\mathcal{P}$ as that obtained from duality, and the associated gap between this bound and the optimal value is equal to the duality gap.

Theorem 1 along with Claim 2 directly yield the following corollary for the max–min-fair problem $\mathcal{F}$.

Corollary 2: $\mathcal{D}(\mathcal{D}(\mathcal{F})) = \mathcal{R}(\mathcal{F})$ and $\beta(\mathcal{R}(\mathcal{F})) = \beta(\mathcal{D}(\mathcal{F}))$.

VIII. Simulation Results

An appropriate figure of merit for the performance of the proposed algorithm for the QoS beamforming problem $\mathcal{Q}$ would be the ratio of the minimum transmitted power achieved by the proposed algorithm and $\beta(\mathcal{Q})$, the transmitted power achieved by the (true) optimal solution. Unfortunately, problem $\mathcal{Q}$ is NP-hard, and thus $\beta(\mathcal{Q})$ can be difficult to compute. However, we can replace $\beta(\mathcal{Q})$ in the figure of merit by the lower bound obtained from the SDR; i.e., $\beta(\mathcal{Q}) \geq \beta(\mathcal{Q}_r) = \mathrm{trace}(\mathbf{X}_{\mathrm{opt}})$. If we let $\{\mathbf{w}_\ell\}$ denote the sequence of candidate weight vectors generated via randomization, and $\breve{\mathbf{w}}_\ell$ denote the minimally scaled version of $\mathbf{w}_\ell$ that satisfies the constraints of problem $\mathcal{Q}$, then a meaningful and easily computable figure of merit is $(\min_\ell \|\breve{\mathbf{w}}_\ell\|_2^2)/\mathrm{trace}(\mathbf{X}_{\mathrm{opt}})$. We will call this ratio the upper

TABLE III
MC SIMULATION RESULTS FOR QoS BEAMFORMING: MEAN AND STANDARD DEVIATION OF UPPER BOUND ON POWER BOOST. EACH ELEMENT OF $\mathbf{h}_i$ IS i.i.d. WITH A CIRCULARLY SYMMETRIC COMPLEX GAUSSIAN (RAYLEIGH) DISTRIBUTION OF VARIANCE 1. ALL THREE RANDOMIZATION TECHNIQUES (`randA`, `randB`, `randC`) ARE USED IN PARALLEL, FOR 1000 RANDOMIZATIONS EACH. $\rho_{\min,i}\sigma_i^2 = 1, \forall i$

N/M	mean	std
4/8	1.12	0.16
4/16	1.47	0.30
8/16	1.82	0.37
8/32	2.79	0.47

TABLE IV
MC SIMULATION RESULTS FOR QoS BEAMFORMING: MEAN AND STANDARD DEVIATION OF UPPER BOUND ON POWER BOOST. HERE, THE NUMBER OF POST-SDR randomizations $= 30\,NM$. REMAINING PARAMETERS ARE AS IN TABLE III

N/M	mean	std
4/8	1.12	0.16
4/16	1.44	0.29
8/16	1.76	0.34
8/32	2.49	0.38

bound on the power boost required to satisfy the constraints. If our algorithm achieves a power boost of η, then the transmitted power is guaranteed to be within a factor η of that of the optimal solution $\beta(\mathcal{Q})$ and will often be closer.

A. Rayleigh Fading Wireless Channels

We consider the standard independent and identically distributed (i.i.d.) Rayleigh fading model described in the caption of Table III. That table summarizes the results obtained using the direct QoS relaxation algorithm in Table I ($\rho_{\min,i}\sigma_i^2 = 1$, $\forall i$) with all three randomization options (`randA`, `randB`, and `randC`) employed in parallel, for a fixed number of 1000 randomization samples each. Table IV summarizes results for the same scenario, except that 30 NM randomization samples are drawn for each randomization strategy—thus the number of randomizations grows linearly in the problem size. Note that, in many cases, our solutions are within 3–4 dB from the (generally optimistic) lower bound on transmitted power provided by SDR, and thus are guaranteed to be at most 3–4 dB away from optimal; this is often good enough from an engineering perspective. Comparing the corresponding entries in Tables III and IV, it is evident that switching from 1000 to 30 NM randomizations per channel realization only yields a minor performance improvement in the cases considered.

Table V summarizes our simulation results for max–min fair beamforming, using the direct algorithm in Table II ($\sigma_i^2 = 1$, $\forall i$, $P = 1$). Table V presents Monte Carlo averages for the upper bound on the minimum SNR (the optimum attained in problem $\mathcal{F}_r$), the SDR-attained minimum SNR (after randomization), the minimum SNR attained by the maximum average SNR beamformer[9] [7, ch. 5], and the minimum SNR for the case of no beamforming. For the latter, we have used $\mathbf{w} = (1/\sqrt{N})\mathbf{1}_{N\times 1}$, which fixes transmitted power to 1. Under the i.i.d. Rayleigh fading assumption, this is *equivalent* to selecting an arbitrary transmit antenna, allocating the entire power budget to it, and shutting off all others. To see this, note that the sum channel $(1/\sqrt{N})\mathbf{1}_{N\times 1}^T\mathbf{h}_i$ viewed by any particular receiver i will still be Rayleigh, of the same variance as the elements of $\mathbf{h}_i$. For this reason, we can view the beamforming vector $\mathbf{w} = (1/\sqrt{N})\mathbf{1}_{N\times 1}$ as corresponding to no beamforming at all. All three randomization options (`randA`, `randB`, and `randC`) were employed in parallel, for 30 NM samples each. It is satisfying to note that the SDR solution attains a significant fraction of the (possibly unattainable) upper bound. Furthermore, the SDR technique provides a substantial improvement in the average minimum SNR relative to no beamforming and to maximum average SNR beamforming [7, ch. 5]. Like SDR, maximum average SNR beamforming uses full CSI at the transmitter. However, it is generally not meaningful to compare designs produced under different objectives. Accordingly, the maximum average SNR beamforming results in Table V are only meant to convey an idea of how much QoS improvement SDR can provide over computationally simpler solutions that also exploit full CSI.

TABLE V
MC SIMULATION RESULTS FOR MAX–MIN FAIR BEAMFORMING: AVERAGES FOR THE UPPER BOUND ON $\min_i \text{SNR}_i$, THE RELAXATION-ATTAINED $\min_i \text{SNR}_i$, THE $\min_i \text{SNR}_i$ ATTAINED BY MAXIMIZING AVERAGE SNR (ACROSS USERS), AND THE $\min_i \text{SNR}_i$ FOR THE CASE OF NO BEAMFORMING. THE RESULTS ARE AVERAGED OVER 1000 MONTE CARLO (MC) RUNS. FOR EACH MC RUN, THE ELEMENTS OF $\mathbf{h}_i$ ARE INDEPENDENTLY REDRAWN FROM A CIRCULARLY SYMMETRIC COMPLEX GAUSSIAN DISTRIBUTION OF VARIANCE 1. $\sigma_i^2 = 1, \forall i$, $P = 1$. ALL THREE RANDOMIZATION TECHNIQUES (`randA`, `randB`, `randC`) ARE USED IN PARALLEL, FOR 30 NM RANDOMIZATIONS EACH

N/M	upper bound	SDR	Max Avg SNR	no BMF
4/8	1.05	0.94	0.25	0.12
4/16	0.73	0.51	0.11	0.06
8/16	1.43	0.86	0.14	0.06
8/32	1.07	0.45	0.06	0.03

We observe from Tables III–V, that as N and/or M increase, the quality of the approximate solution drifts away from the respective relaxation/duality bound. This could be due to a variety of factors, or combination thereof. First, the relaxation bound may become more optimistic at higher N and/or M—remember that it is only a bound, not necessarily a tight bound. If this is true, then the apparent degradation may in fact be much milder in reality. Second, the number of randomizations required to attain a quasi-optimal solution may increase faster than linearly in the product NM. Third, the approximation quality of the

[9]This beamformer maximizes the average SNR for each channel matrix realization (Monte Carlo run), where the average is taken over the users. The resulting beamforming vector is also scaled to unit norm.

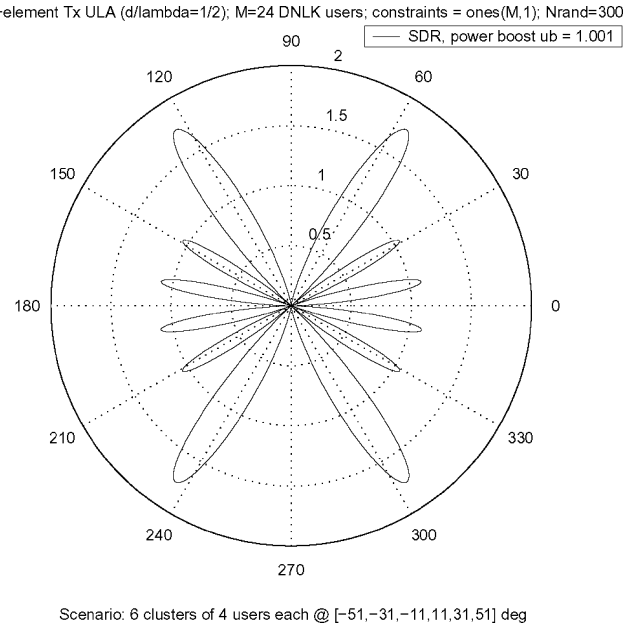


Fig. 1. Broadcast beamforming example using algorithm in Table I. Optimized beam pattern for $N = 8$-element transmit ULA ($d/\lambda = 1/2$) and $M = 24$ downlink users, in six clusters of four users each. Clusters centered at $[-51, -31, -11, 11, 31, 51]^\circ$ with extent $\pm 2^\circ$. Channel vectors are Vandermonde, of element modulus 1. $\sigma_i^2 = \sigma^2$, $\forall i$, $\rho_{\min,i} = 1/\sigma^2$, $\forall i$ (here, σ_i^2 also models propagation loss, in addition to thermal noise). Symmetric lobes appear due to the inherent ULA ambiguity. `randA`, # post-SDR randomizations = 300. In this case, the solution is guaranteed to be within 0.1% of the optimum.

method per se may degrade as the problem size grows. In a related, but distinct, problem the quality of the SDR approximation degrades logarithmically in the problem size [10].

B. Far-Field Beamforming for a Uniform Linear Transmit Antenna Array

In several scenarios, the solutions generated by the SDR technique are essentially optimal. This is illustrated in Fig. 1, which shows the optimized transmit beampattern for a particular far-field multicasting scenario using a uniform linear antenna array (ULA); the details of the simulation setup are included in the figure caption for ease of reference.

C. Measured VDSL Channels

In this section, we test the performance of our algorithms using measured VDSL channel data collected by France Telecom R&D as part of the EU-FP6 U-BROAD project # 506 790.

Gigabit VDSL technology for very short twisted copper loops (in the order of 100–500 m) is currently under development in the context of fiber to the basement (FTTB) or fiber to the curb/cabinet (FTTC) hybrid access solutions. Multiple-input multiple-output (MIMO) transmission modalities are an important component of gigabit VDSL. These so-called *vectoring* techniques rely on transmit precoding and/or multiuser detection to provide reliable communication at very high transmission rates [5]. Transmit precoding is particularly appealing when the targeted receivers are not physically co-located, or when legacy equipment is being used at the receive site. In both cases, multiuser detection is not feasible. In this context, media streaming (e.g., news-feed, pay-per-view, or video-conferencing) may involve multiple recipients in the same binder.

Let N denote the number of loops subscribing to a given multicast. With multicarrier transmission, each tone can be viewed as a flat-fading MIMO channel with N inputs and N outputs, plus noise and alien interference. The diagonal of the channel matrix consists of samples of the N direct [insertion loss (IL)] channel frequency responses, while off-diagonal elements are drawn from the corresponding FEXT channel frequency responses. Due to the noncoherent combining of the self-FEXT coupling coefficients, the useful signal power received at each output terminal is reduced, even when all inputs are fed with the same information-bearing signal. That is, the equivalent channel tap at frequency f is $h_{\rm e}(f) = h_{\rm IL}(f) + \sum_{n_{\rm FEXT}=1}^{N-1} h_{n_{\rm FEXT}}(f)$, where $h_{\rm IL}(f)$ denotes the direct (insertion loss) channel, and $h_{n_{\rm FEXT}}(f)$ denotes a generic FEXT interference channel.

Conceptually, the scenario is very similar to the wireless scenario considered earlier, but with two key differences: now $N = M$, and the channel matrix $\mathbf{H} := [\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_M]$ is diagonally dominated, because FEXT coupling is much weaker than insertion loss. The question then is whether transmit precoding can provide a meaningful benefit relative to simply ignoring FEXT altogether.

We use IL and far-end FEXT measured data for S88 cable comprising 14 quads, i.e., 28 loops. The length of the cable is 300 m. For each channel, a log-frequency sweeping scheme was used to measure the I/Q components of the frequency response from 10 kHz to 30 MHz, yielding 801 complex samples per channel. Cubic spline complex interpolation was used to convert these samples to a linear frequency scale. We consider 17 $N \times N$ channel matrices, with $N = 14$, in the frequency range 21.5 to 30 MHz. Insertion loss drops between -40 and -45 dB in this range of frequencies, while FEXT coupling is between -77 and -82 dB in the mean, with over 10-dB standard deviation and significant variation across frequency as well. For each channel matrix, we apply our max–min-fair beamforming algorithm with $\sigma_i^2 = \sigma^2$, $\forall i$, and $P/\sigma^2 = 1$. Fig. 2 shows the resulting plots of minimum received signal power, the associated relaxation/duality bound, and the minimum received signal power when no precoding is used. We observe that SDR can almost double the minimum received signal power relative to no precoding, and it often attains zero gap relative to the relaxation/duality bound. For shorter loops (e.g., 100 m), the situation is even more in favor of SDR, because then FEXT resembles near-end crosstalk (NEXT) and is relatively more pronounced.

D. Further Observations

1) Comparison of the Two Relaxations: We have shown theoretically that the two problem formulations (QoS, $\mathcal{Q}$, and max–min-fair, $\mathcal{F}$) are algorithmically equivalent, i.e., had we had an optimal algorithm that provides the exact solution to one, it could have also been used to obtain the exact solution to the other. What we have instead is two generally approximate algorithms, obtained by direct relaxation of the respective problems. The link between the two formulations can still be exploited. For example, we may obtain an approximate

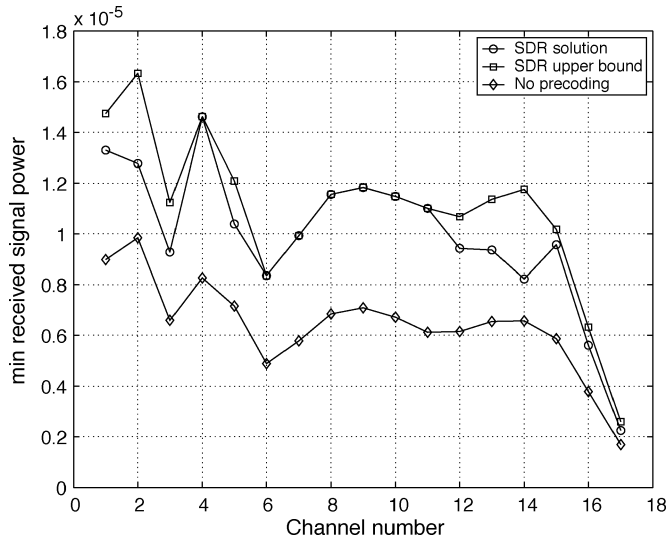


Fig. 2. Transmit precoding for VDSL multicasting.

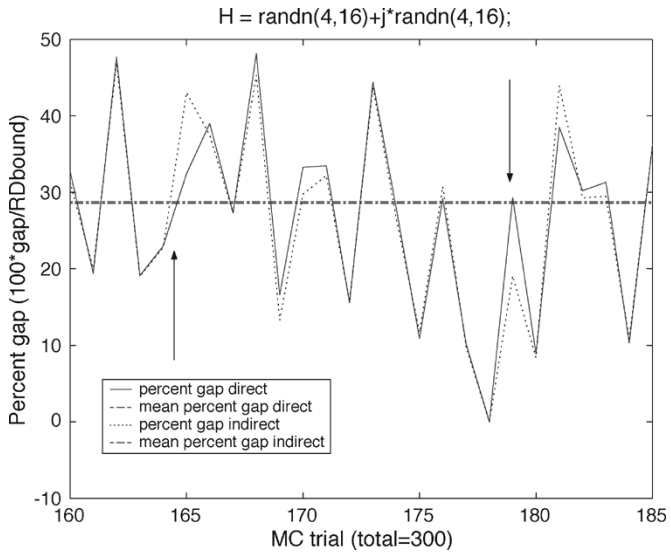


Fig. 3. Comparison of direct and indirect solutions to the max–min-fair problem.

solution to the max–min-fair problem by first running the QoS algorithm in Table I with all the $\rho_{\min,i} = 1$, then scaling the resulting solution to the desired power level P. Of course, we can also use the direct relaxation of the max–min-fair problem in Table II. Due to approximation, there is no *a priori* reason to expect that the two solutions will be identical, even in the mean.

In order to address this issue, we have compared the two strategies by means of Monte Carlo simulation. We chose $N = 4$, $M = 16$, $\sigma_i^2 = 1$, $\forall i$, and $P = 1$, and ran both algorithms for 300 i.i.d. Rayleigh fading channels. All three randomizations (`randA`, `randB`, `randC`) were employed in parallel, for 30 NM randomization samples each. For each channel, we recorded the percent gap (100 times the gap over the relaxation bound) of each algorithm. Fig. 3 shows a portion of the results, along with the mean percent gap attained by each algorithm (averaged over all 300 channels). By "direct" we refer to the algorithm in Table II, whereas by "indirect, " we refer to the algorithm in Table I with all $\rho_{\min,i} = 1$, followed by scaling.

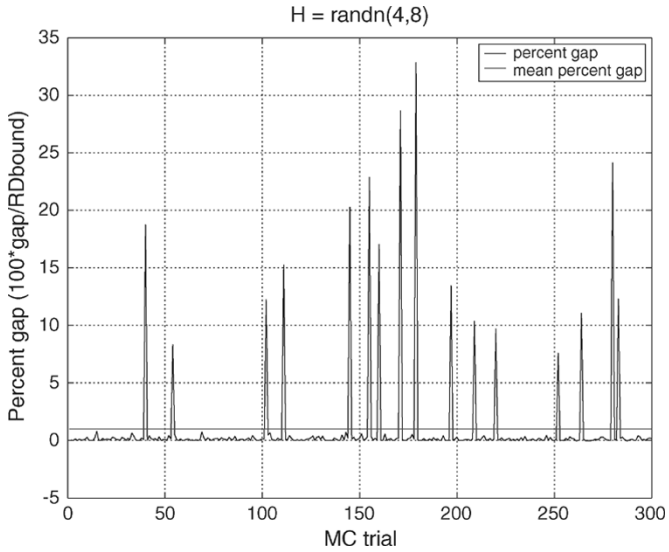


Fig. 4. Percent gap outcomes for 300 real Gaussian channel realizations.

We observe that the mean percent gaps of the two algorithms are virtually identical, and in fact most of the respective percent gaps are very close on a sample-by-sample basis. However, there are instances wherein each algorithm is significantly better than the other (over 10% difference in the gap). Two pronounced cases are highlighted by arrows in Fig. 3. We conclude that, while both approaches are equally effective on average, it pays to use both, if possible, in certain cases.

2) On the Dependence of Gap Statistics on Channel Statistics: We have seen that, for i.i.d. circular Gaussian (Rayleigh) channel matrices, the gap between our relaxation–randomization approximate solutions and the relaxation/duality bound might not be insignificant. We have also seen cases wherein the gap is very small, cf., the far-field uniform linear transmit antenna array example, and a good proportion of the VDSL channels tested earlier.

It is evident that the gap statistics depend on the channel statistics. Interestingly, the gap statistics are far more favorable for *real* (as opposed to complex circular) i.i.d. Gaussian channels. This is illustrated in Fig. 4, using the QoS algorithm in Table I for $N = 4$, $M = 8$, $\rho_{\min,i}\sigma_i^2 = 1$, $\forall i$, and 300 real i.i.d. Gaussian channels. All three randomizations (`randA`, `randB`, `randC`) are employed in parallel, for 30 NM randomization samples each. For each channel, we recorded the percent gap (100 times the gap over the relaxation bound) of the algorithm in Table I. Observe that for about 95% of the channels the percent gap is down to numerical accuracy in this case. Contrast this situation with Fig. 5, which shows the respective results for complex circular Gaussian channel matrices—the difference is remarkable.

There are other cases where we have observed that the relaxation approach operates close to zero gap. One somewhat contrived case is when the real and imaginary parts of the channel coefficients are nonnegative. This is illustrated in Fig. 6, where it is worth noting that the scaling of the y axis is 10^{-8}. In this case, the gap hovers around numerical accuracy, without exhibiting any bad runs at all for the 300 channel matrices considered.

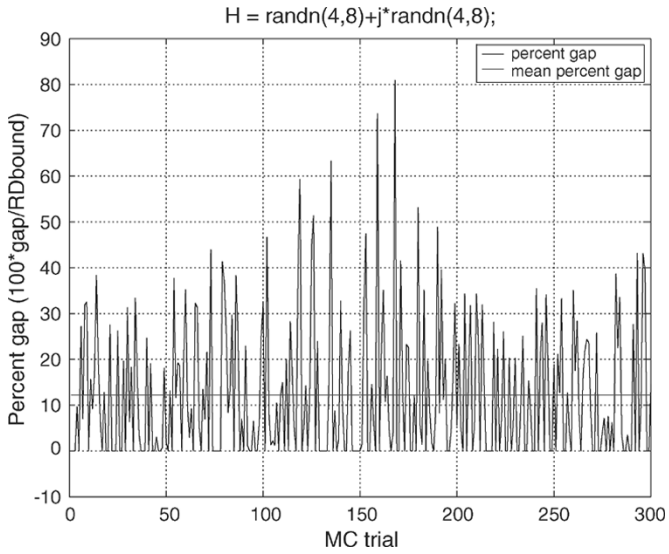


Fig. 5. Percent gap outcomes for 300 complex circular Gaussian (Rayleigh) channel realizations.

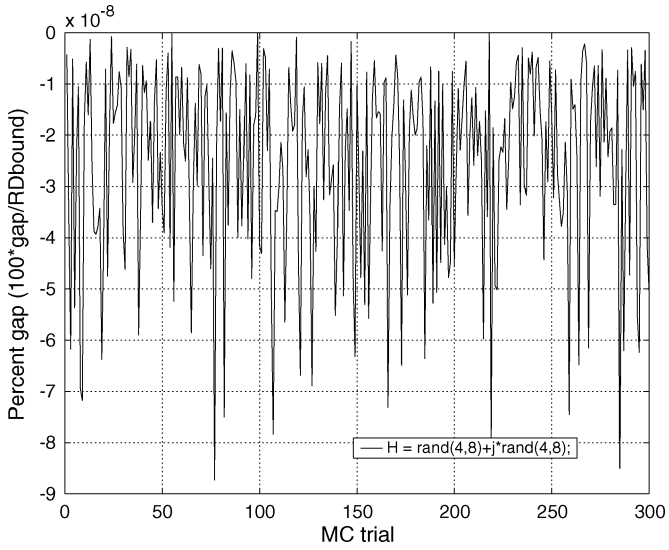


Fig. 6. Percent gap outcomes for 300 channel realizations with positive real and imaginary parts (uniformly distributed between 0 and 1). Note that the scaling of the y axis is 10^{-8}.

In conclusion, the complex circular Gaussian channel case appears to be the least favorable of the scenarios considered.

IX. Conclusion

We have taken a new look at the broadcasting/multicasting problem when channel state information is available at the transmitter. We have proposed two pertinent problem formulations: minimizing transmitted power under multiple minimum received power constraints, and maximizing the minimum received power subject to a bound on the transmitted power. We have shown that both formulations are NP-hard optimization problems; however, their solution can often be well approximated using semidefinite relaxation tools. We have explored the relationship between the two formulations and also insights afforded by Lagrangian duality theory. In view of i) our extensive numerical experiments with simulated and measured data, verifying that semidefinite relaxation consistently yields good performance, ii) proof that the basic problem is NP-hard, and thus approximation is unavoidable, and iii) corroborating motivation provided by duality theory, we conclude that the approximate solutions provided herein offer useful designs across a broad range of applications.

It would be useful to analyze the duality gap for the problem at hand, for this would yield *a priori* bounds on the degree of suboptimality introduced by relaxation, as opposed to the *a posteriori* bound that we now have by virtue of Theorem 1. Our numerical results indicate that the degree of suboptimality is often acceptable in our intended applications. In an effort to understand the apparent success of the SDR approach (e.g., in the case where the channel vectors have nonnegative real and imaginary parts), one can consider the following simple linearly constrained convex quadratic program (QP) *restriction* of the QoS problem:

$$\mathcal{Q}_s: \quad \min_{\mathbf{w}\in\mathbb{C}^N} \|\mathbf{w}\|_2^2 \quad \text{subject to: } \mathrm{Re}\left\{\tilde{\mathbf{h}}_i^H\mathbf{w}\right\} \geq 1, \quad \text{for all } i.$$

Notice that the feasible region of this problem is a subset of that of the original nonconvex (and NP-hard) QoS formulation $\mathcal{Q}$. Thus, $P^* \leq \bar{P}$, where P^* and $\bar{P}$ denote the minimum beamforming power obtained from optimal solutions of $\mathcal{Q}$ and $\mathcal{Q}_s$, respectively. We have recently shown [16] that the gap between P^* and $\bar{P}$ is never more than $1/\cos^2(\alpha/2)$, where α is the maximum phase spread across the different users measured at each transmit antenna and is assumed to be less than π. Notice that the two cases where channel vectors i) are real and nonnegative or ii) have nonnegative real and imaginary parts correspond to $\alpha = 0$ and $\alpha \leq \pi/2$. Thus, $\mathcal{Q}_s$ provides an exact solution in the first case and a factor of 2 approximation in the second case. These results indicate that problem $\mathcal{Q}$ is well approximated by $\mathcal{Q}_s$ if the phase spread α is small.

There are many other interesting extensions to the algorithms developed herein: e.g., robustness issues, and multiple cochannel multicasting groups. These are subjects of ongoing work and will be reported elsewhere. Furthermore, aside from transmit beamforming/precoding, there are also more traditional signal processing applications of the proposed methodology. One is linear filter design, in particular, the design of a linear "batch" filter that responds to certain prescribed frequencies in its input and attenuates all other frequencies. In this setting, the $\mathbf{h}_i$ vectors will be Vandermonde, with generators $e^{j\omega_i}$ and $\omega_i \in [-\pi, \pi)$. One may easily envision scenarios wherein such a problem formulation can be appropriate: radio-astronomy applications, frequency-diversity combining, and frequency-hopping communications. The context can be further generalized: design a linear filter that responds to prescribed but otherwise arbitrary signals in its input, while attenuating all else.

APPENDIX I
PROOF OF CLAIM 1

Before dealing with Claim 1 directly, we first consider the following restriction of the QoS problem $\mathcal{Q}$: the case when all $\mathbf{h}_i$ are real, and optimization is over $\mathbb{R}^N$. We will show that[10]

$$\mathcal{S}: \quad \min_{\mathbf{x}\in\mathbb{R}^N} \mathbf{x}^T\mathbf{x} \quad \text{subject to: } |\mathbf{x}^T\mathbf{h}_m| \geq 1,\ m \in \{1,\cdots,M\}$$

contains

$$\mathcal{A}: \quad \min_{y_n\in\mathbb{R}} y_1^2+\cdots+y_N^2+\left(\sum_{n=1}^N a_n y_n\right)^2 \quad \text{subject to: } y_n^2 \geq 1,\ n \in \{1,\cdots,N\}$$

as a special case and that problem $\mathcal{A}$ is at least as hard as the following problem:

Partition Problem Π*:* Given integers $a_1,\cdots,a_N$, do there exist binary variables $\{x_n\}_{n=1}^N \in \{+1,-1\}^N$, such that $\sum_{n=1}^N a_n x_n = 0$?

This is known to be NP-complete [4].

It is easy to check that the optimal value of problem $\mathcal{A}$ is equal to N if and only if the answer to problem Π is affirmative. Thus, solving problem $\mathcal{A}$ is at least as hard as solving problem Π.

To show that problem $\mathcal{S}$ contains problem $\mathcal{A}$ (i.e., an arbitrary instance of problem $\mathcal{A}$ can be posed as a special instance of problem $\mathcal{S}$), note that $y_n^2 \geq 1$ can be written as $|\mathbf{y}^T\mathbf{e}_n| \geq 1$, where $\mathbf{y} := [y_1,\cdots,y_N]^T$ and $\mathbf{e}_n$ contains one in the nth position and zeros elsewhere. Furthermore

$$y_1^2+\cdots+y_N^2+\left(\sum_{n=1}^N a_n y_n\right)^2 = \mathbf{y}^T(\mathbf{I}+\mathbf{a}\mathbf{a}^T)\mathbf{y} = \mathbf{y}^T\mathbf{Q}\mathbf{y},$$

where $\mathbf{a} := [a_1,\cdots,a_N]^T$, and $\mathbf{Q} := \mathbf{I}+\mathbf{a}\mathbf{a}^T$. The matrix $\mathbf{Q}$ is positive definite. Let $\mathbf{Q} = \mathbf{S}^T\mathbf{S}$, and $\mathbf{x} := \mathbf{S}\mathbf{y}$. Then $\mathbf{y}^T\mathbf{Q}\mathbf{y} = \mathbf{x}^T\mathbf{x}$, $\mathbf{y} = \mathbf{S}^{-1}\mathbf{x}$, and $|\mathbf{y}^T\mathbf{e}_n| \geq 1$ can be written as $|\mathbf{x}^T\mathbf{S}^{-T}\mathbf{e}_n| \geq 1$, or, with $\mathbf{h}_n := \mathbf{S}^{-T}\mathbf{e}_n$, as $|\mathbf{x}^T\mathbf{h}_n| \geq 1$. This shows that an arbitrary instance of problem $\mathcal{A}$ can be transformed to a special instance of problem $\mathcal{S}$ (with $M = N$). Thus, $\mathcal{S}$ is at least as hard as $\mathcal{A}$, which is at least as hard as the partition problem. □

Proof of Claim 1: The QoS Problem $\mathcal{Q}$ is NP-hard: Consider the problem

$$\min_{\mathbf{w}\in\mathbb{C}^N} \mathbf{w}^H\mathbf{w} \quad \text{subject to: } |\mathbf{w}^H\mathbf{h}_i| \geq 1, \quad i=1,\ldots,M. \tag{1}$$

Define the $N\times M$ matrix $\mathbf{H} = [\mathbf{h}_1,\cdots,\mathbf{h}_M]$, and the $M\times 1$ vector $\mathbf{z}$, with $\mathbf{z}^H := \mathbf{w}^H\mathbf{H}$. Consider the case that $M \geq N$, and $\mathbf{H}$ is full row-rank (N). Then $\mathbf{w}^H = \mathbf{z}^H\mathbf{H}^\dagger$, where $\mathbf{H}^\dagger = \mathbf{H}^H(\mathbf{H}\mathbf{H}^H)^{-1}$ denotes the right pseudoinverse of $\mathbf{H}$, and the problem in (1) is equivalent to

$$\min_{\mathbf{z}\in\mathbb{C}^M} \mathbf{z}^H\mathbf{Q}\mathbf{z} \quad \text{subject to: } |z_k| \geq 1, \quad k=1,\ldots,M \tag{2}$$

where $\mathbf{Q} := \mathbf{H}^\dagger(\mathbf{H}^\dagger)^H \succeq \mathbf{0}$, a positive semidefinite matrix of rank $N \leq M$; and z_k denotes the kth element of the vector $\mathbf{z}$. We will show that problem (2) is NP-hard in general. To this end, we consider a reduction from the NP-complete partition problem [4]; i.e., given $a_1 > 0, a_2 > 0,\ldots,a_P > 0$, decide whether or not a subset, say I, of $\{1,\ldots,P\}$ exists, such that

$$\sum_{k\in I} a_k = \frac{1}{2}\sum_{k=1}^P a_k. \tag{3}$$

Let $M = 2P+1$ and let the complex-valued decision vector be

$$\mathbf{z} = [z_0, z_1,\ldots,z_P,z_{P+1},\ldots,z_{2P}]^T \in \mathbb{C}^M.$$

Let us denote

$$\mathbf{a} := [a_1 \quad a_2 \quad \ldots \quad a_P]^T$$
$$\mathbf{A} := \begin{bmatrix} -\mathbf{1}_P & \mathbf{I}_P & \mathbf{I}_P \\ -\frac{1}{2}\mathbf{1}_P^T\mathbf{a} & \mathbf{a}^T & \mathbf{0}_P^T \end{bmatrix}$$
$$\mathbf{Q} := \mathbf{A}^T\mathbf{A} + \mathbf{I}_M$$

where $\mathbf{1}_P$ denotes the length-P vector of ones, and $\mathbf{0}_P$ is the length-P vector of zeros.

Next we show that a partition I satisfying (3) exists if and only if the optimization problem (2) has a minimum value of M. In other words, the existence of I is equivalent to the fact that there is $\mathbf{z}\in\mathbb{C}^M$ such that $\mathbf{z}^H\mathbf{Q}\mathbf{z} = M$ and $|z_k| \geq 1$, for all k. Since

$$\mathbf{z}^H\mathbf{Q}\mathbf{z} = \|\mathbf{A}\mathbf{z}\|_2^2 + \sum_{k=0}^{2P}|z_k|^2 \geq 2P+1 = M, \quad \text{for } |z_k|\geq 1\ \forall k,$$

it follows that

$$\mathbf{z}^H\mathbf{Q}\mathbf{z} = M, \quad |z_k| \geq 1 \quad \text{for all} \quad k$$

is equivalent to

$$\mathbf{A}\mathbf{z} = \mathbf{0}, \quad |z_k| = 1 \quad \text{for all } k.$$

The latter gives rise to a set of linear equations

$$-z_0 + z_k + z_{P+k} = 0, \quad k=1,\ldots,P \tag{4}$$
$$-\frac{1}{2}\left(\sum_{k=1}^P a_k\right)z_0 + \sum_{k=1}^P a_k z_k = 0. \tag{5}$$

[10]We henceforth use $\mathbf{h}_m$ to denote possibly scaled channel vectors, dropping the tilde for brevity.

The z_k's are all constrained to be on the unit circle; thus let $z_k/z_0 = e^{i\theta_k}$ for $k = 1, \ldots, 2P$. Using (4), we have

$$\cos\theta_k + \cos\theta_{P+k} = 1 \tag{6}$$
$$\sin\theta_k + \sin\theta_{P+k} = 0 \tag{7}$$

where $k = 1, \ldots, P$. These two equations imply that $\theta_k \in \{-\pi/3, \pi/3\}$ for all k. This, in particular, means that $\cos\theta_k = \cos\theta_{P+k} = 1/2$ for $k = 1, \ldots, P$, implying that

$$\mathrm{Re}\left\{-\frac{1}{2}\left(\sum_{k=1}^{P} a_k\right) + \sum_{k=1}^{P} \frac{a_k z_k}{z_0}\right\} = 0.$$

Therefore, (5) is satisfied if and only if

$$\mathrm{Im}\left\{-\frac{1}{2}\left(\sum_{k=1}^{P} a_k\right) + \sum_{k=1}^{P} \frac{a_k z_k}{z_0}\right\} = \mathrm{Im}\left\{\sum_{k=1}^{P} \frac{a_k z_k}{z_0}\right\} = \sum_{k=1}^{P} a_k \sin\theta_k = 0,$$

with $\theta_k \in \{-\pi/3, \pi/3\}$ for all k, and thus $\sin\theta_k \in \{(\sqrt{3}/2), -(\sqrt{3}/2)\}$, which is equivalent to the existence of a partition I of $\{a_1, \ldots, a_P\}$ such that (3) holds. In fact, we can imply take $I = \{k | \theta_k = \pi/3\}$. □

ACKNOWLEDGMENT

The authors would like to thank Dr. M. Ouzzif and Dr. R. Tarafi of France Telecom R&D, who conducted the VDSL channel measurements under the auspices of the U-BROAD project.

REFERENCES

[1] M. Bengtsson and B. Ottersten, "Optimal and suboptimal transmit beamforming," in *Handbook of Antennas in Wireless Communications*, L. C. Godara, Ed. Boca Raton, FL: CRC Press, Aug. 2001, ch. 18.

[2] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[3] G. D. Forney, "Maximum likelihood sequence estimation of digital sequences in the presence of intersymbol interference," *IEEE Trans. Inf. Theory*, vol. 18, no. 3, pp. 363–378, May 1972.

[4] M. R. Garey and D. S. Johnson, *Computers and Intractability. A Guide to the Theory of NP-Completeness*. San Francisco, CA: Freeman, 1979.

[5] G. Ginis and J. M. Cioffi, "Vectored transmission for digital subscriber line systems," *IEEE J. Sel. Areas Commun.*, vol. 20, no. 5, pp. 1085–1104, Jun. 2002.

[6] N. Jindal and A. Goldsmith, "Optimal power allocation for parallel gaussian broadcast channels with independent and common information," presented at the Int. Symp. Information Theory, Chicago, IL, Jun. 2004.

[7] M. J. Lopez, "Multiplexing, scheduling, and multicasting strategies for antenna arrays in wireless networks," Ph.D. thesis, Dept. of Elect. Eng. and Comp. Sci., MIT, Cambridge, MA, 2002.

[8] Z.-Q. Luo, "Lecture Notes for EE8950: Engineering Optimization," Univ. of Minnesota, Minneapolis, Spring 2004.

[9] W.-K. Ma, T. N. Davidson, K. M. Wong, Z.-Q. Luo, and P.-C. Ching, "Quasi-ML multiuser detection using semi-definite relaxation with application to synchronous CDMA," *IEEE Trans. Signal Process.*, vol. 50, no. 4, pp. 912–922, Apr. 2002.

[10] A. Nemirovski, C. Roos, and T. Terlaky, "On maximization of quadratic form over intersection of ellipsoids with common center," *Math. Program.*, ser. A, vol. 86, pp. 463–473, 1999.

[11] J. F. Sturm, "Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones," *Optim. Meth. Softw.*, vol. 11–12, pp. 625–653, 1999.

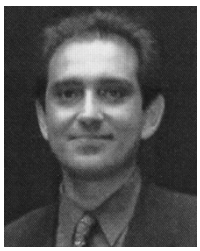
[12] P. Tseng, "Further results on approximating nonconvex quadratic optimization by semidefinite programming relaxation," *SIAM J. Optim.*, vol. 14, no. 1, pp. 268–283, Jul. 2003.

[13] S. A. Vorobyov, A. B. Gershman, and Z.-Q. Luo, "Robust adaptive beamforming using worst-case performance optimization: A solution to the signal mismatch problem," *IEEE Trans. Signal Process.*, vol. 51, no. 2, pp. 313–324, Feb. 2003.

[14] H. Wolkowicz, "Relaxations of Q2P," in *Handbook of Semidefinite Programming: Theory, Algorithms, and Applications*, H. Wolkowicz, R. Saigal, and L. Vandenberghe, Eds. Norwell, MA: Kluwer, 2000, ch. 13.4.

[15] S. Zhang, "Quadratic maximization and semidefinite relaxation," *Math. Program.*, ser. A, vol. 87, pp. 453–465, 2000.

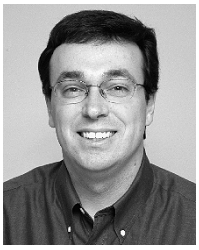
[16] Z.-Q. Luo, N. D. Sidiropoulos, P. Tseng, and S. Zhang, "Approximation bounds for quadratic optimization with homogeneous quadratic constraints," *SIAM J. Optim.*, Oct. 2005, submitted for publication.



Nicholas D. Sidiropoulos (M'92–SM'99) received the Diploma degree in electrical engineering from the Aristotelian University of Thessaloniki, Thessaloniki, Greece, in 1988 and the M.S. and Ph.D. degrees in electrical engineering from the University of Maryland, College Park (UMCP), in 1990 and 1992, respectively.

From 1988 to 1992, he was a Fulbright Fellow and a Research Assistant at the Institute for Systems Research (ISR), UMCP. From September 1992 to June 1994, he served his military service as a Lecturer in the Hellenic Air Force Academy. From October 1993 to June 1994, he also was a Member of the Technical Staff, Systems Integration Division, G-Systems, Ltd., Athens, Greece. He was a Postdoctoral Fellow from 1994 to 1995 and Research Scientist from 1996 to 1997 at ISR-UMCP, an Assistant Professor with the Department of Electrical Engineering, University of Virginia, Charlottesville, from 1997 to 1999, and an Associate Professor with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, from 2000 to 2002. He is currently a Professor with the Telecommunications Division of the Department of Electronic and Computer Engineering, Technical University of Crete, Chania, Crete, Greece, and Adjunct Professor at the University of Minnesota. His current research interests are primarily in signal processing for communications and multiway analysis. He is an active consultant for industry in the areas of frequency-hopping systems and signal processing for xDSL modems.

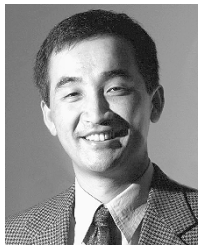
Dr. Sidiropoulos is a member of both the Signal Processing for Communications (SPCOM) and Sensor Array and Multichannel Signal Processing (SAM) Technical Committees of the IEEE Signal Processing Society and currently serves as an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING. From 2000 to 2002, he also served as Associate Editor for the IEEE SIGNAL PROCESSING LETTERS. He received the NSF/CAREER award (Signal Processing Systems Program) in June 1998 and an IEEE Signal Processing Society Best Paper Award in 2001.



Timothy N. Davidson (M'96) received the B.Eng. (Hons. I) degree in electronic engineering from the University of Western Australia (UWA), Perth, in 1991 and the D.Phil. degree in Engineering Science from the University of Oxford, U.K., in 1995.

He is currently an Associate Professor in the Department of Electrical and Computer Engineering at McMaster University, Hamilton, ON, Canada, where he holds the (Tier II) Canada Research Chair in Communication Systems. His research interests are in signal processing, communications and control, with current activity focused on signal processing for digital communication systems. He has held research positions at the Communications Research Laboratory at McMaster University, the Adaptive Signal Processing Laboratory at UWA, and the Australian Telecommunications Research Institute at Curtin University of Technology, Perth, Western Australia.

Dr. Davidson was awarded the 1991 J. A. Wood Memorial Prize [for "the most outstanding (UWA) graduand" in the pure and applied sciences] and the 1991 Rhodes Scholarship for Western Australia.



Zhi-Quan (Tom) Luo (SM'03) received the B.Sc. degree in mathematics from Peking University, China, in 1984 and the Ph.D. degree in operations research from the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, in 1989.

During the academic year of 1984 to 1985, he was with Nankai Institute of Mathematics, Tianjin, China. In 1989, he joined the Department of Electrical and Computer Engineering, McMaster University, Hamilton, ON, Canada, where he became a Professor in 1998 and has held the Canada Research Chair in Information Processing since 2001. Starting April 2003, he has been a Professor with the Department of Electrical and Computer Engineering and holds an endowed ADC research Chair in Wireless Telecommunications with the Digital Technology Center at the University of Minnesota, Minneapolis. His research interests lie in the union of large-scale optimization, information theory and coding, data communications, and signal processing.

Prof. Luo is a member of SIAM and MPS. He is also a member of the Signal Processing for Communications (SPCOM) and Signal Processing Theory and Methods (SPTM) Technical Committees of the IEEE Signal Processing Society. From 2000 to 2004, he served as an Associate Editor for the IEEE TRANSACTIONS. He is presently serving as an Associate Editor for several international journals, including the *SIAM Journal on Optimization* and *Mathematics of Operations Research*. He received an IEEE Signal Processing Society Best Paper Award in 2004.

A Hybrid Probabilistic Data Association-Sphere Decoding Detector for Multiple-Input–Multiple-Output Systems

Georgios Latsoudas and Nicholas D. Sidiropoulos, *Senior Member, IEEE*

***Abstract*—A hybrid probabilistic data association (PDA)-sphere decoding (SD) algorithm is proposed for signal detection in multiple-input–multiple-output (MIMO) systems. The key idea is to reduce the dimension of the problem solved via SD by first running a single stage of the PDA to fix symbols that can be decoded with high reliability. Simulations under a multiple antenna Rayleigh fading scenario show that this two-step algorithm attains a considerably better performance-complexity tradeoff than SD and PDA for low to moderate signal-to-noise ratio (SNR) or higher problem dimensions.**

***Index Terms*—Integer least squares, multiple-input–multiple-output (MIMO) decoding, probabilistic data association (PDA), sphere decoder.**

I. Introduction

MULTIPLE antenna systems have attracted great interest in recent years, due to the capacity improvement that they afford. Vertical Bell Laboratories Layered Space Time (V-BLAST) [3] is a widely known multiple antenna spatial multiplexing system targeting high spectral efficiencies. Unfortunately, the associated maximum-likelihood (ML) detector amounts to a constrained integer least-squares problem, whose exact solution entails exhaustive search. Thus, following the so-called *nulling and cancelling* detector [3], several computationally efficient detection algorithms have been developed for or adapted to V-BLAST.

Sphere Decoding (SD) [11], Probabilistic Data Association (PDA) [8], [10], and Semi-Definite Relaxation (SDR) [9] are three multiple-input–multiple-output (MIMO) detectors that can provide near-optimal performance at relatively low complexity in certain scenarios. Among them, SD appears to be prevalent in the recent literature. Numerous variants and improvements of SD have recently been developed, e.g., [1], [2], [12], and [13], incorporating more sophisticated schemes for increasing the associated search radius and organizing the computations in a more efficient manner, e.g., the Schnorr–Euchner (SE) SD, which uses an improved search strategy [1], [2]. A drawback of the SD family of detectors is that, for close-to-ML performance, complexity remains high in the low signal-to-noise ratio (SNR) regime or when the number of symbols to be jointly detected is large [5], [6].

The PDA is a simpler detection method, which, however, generally provides worse performance than SD. SD, PDA, SDR, and several other algorithms have recently been compared in the context of code division multiple access (CDMA) multiuser detection [4]. A corresponding comparison for the multiple antenna Rayleigh fading scenario (as in V-BLAST) has not been undertaken, to the best of our knowledge. Thorough comparisons are nontrivial, because complexity and performance of SD and SDR depend on a number of parameters. Our experience in [7] indicates that SDR is inferior to SD at high SNR.

In this letter, we propose a hybrid PDA-SD algorithm that attains a better performance-complexity tradeoff than either of its constituent components. At each stage of the decoding process, the PDA produces a set of soft decision metrics that can be used to assess how reliable associated hard decisions would be at that point. The basic idea, then, is to execute *a single stage* of the PDA algorithm and fix those symbols that can be detected with high reliability. After cancelling the effect of those symbols, a reduced-dimensionality problem is passed to SD for decoding. This reduces the complexity of SD and improves the performance of PDA. Our simulations show that the proposed algorithm enjoys an error performance close to that of SD over a wide range of SNR, at a significantly reduced computational cost.

We use the SD algorithm in Viterbo–Boutros (VB-SD) [11], with an initial radius chosen according to [5], and the SE-SD in [1] and [2], with a search radius set to infinity. Note, however, that the initial PDA stage can also be combined with other variants of SD or SDR. The key here is that *dimensionality reduction* via single-stage PDA preprocessing can provide significant computational relief at a small performance cost.

II. System Model

The aforementioned techniques are applicable to a broad range of MIMO communication systems. Herein, we focus on V-BLAST for concreteness. V-BLAST is a symbol synchronized multiple antenna system with n_T transmit and n_R receive antennas, with $n_T \leq n_R$. The input stream of bits is mapped to a particular constellation, and the resulting symbol stream is demultiplexed into n_T substreams. The transmissions are organized into bursts of L symbol periods. It is assumed that

Manuscript received July 28, 2004; revised October 28, 2004. This work was supported in part by the U.S. ARO under ERO Contract N62558-03-C-0012 and by the EU under U-BROAD STREP 506790S. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Yiteng (Arden) Huang.

The authors are with the Department of Electronic and Computer Engineering, Technical University of Crete, University Campus-Kounoupidiana, 731 00 Chania-Crete, Greece (e-mail: latsoud@telecom.tuc.gr; nikos@telecom.tuc.gr).

Digital Object Identifier 10.1109/LSP.2005.843779

the channel is frequency flat and block fading (i.e., its variation is negligible over the L symbol periods comprising a burst and random from one burst to the next). The channel is assumed to be known to the receiver but not to the transmitter. From the discrete-time baseband-equivalent viewpoint, the system can be represented as

$$\tilde{\mathbf{r}} = \sqrt{\frac{\rho}{n_T}}\tilde{\mathbf{A}}\tilde{\mathbf{s}} + \tilde{\mathbf{n}} = \tilde{\mathbf{H}}\tilde{\mathbf{s}} + \tilde{\mathbf{n}} \tag{1}$$

where $\tilde{\mathbf{r}} = [\tilde{r}_1, \tilde{r}_2, \ldots, \tilde{r}_{n_R}]^T, \tilde{\mathbf{s}} = [\tilde{s}_1, \tilde{s}_2, \ldots, \tilde{s}_{n_T}]^T$ are the receive and the transmit vector, respectively, $\tilde{\mathbf{A}}$ is a generally complex $n_R \times n_T$ channel matrix with entries $\tilde{a}_{ij}$, and $\tilde{\mathbf{n}}$ is a white Gaussian circularly symmetric $n_R \times 1$ noise vector with covariance matrix $2\sigma^2\mathbf{I}$. The normalized amplitude $\sqrt{(\rho/n_T)}$ ensures that the SNR is constant for a given noise variance, irrespective of n_T. Assuming rich scattering, the elements of $\tilde{\mathbf{A}}$ are modeled as independent and identically distributed (i.i.d.) circularly symmetric Gaussian variables with zero mean and unit variance of the real and imaginary parts. For simplicity, we assume that the transmitted symbols are taken from a 4-QAM constellation, but the ideas generalize to higher order constellations. In order to transform the above model to a real-valued one, define

$$\mathbf{s} := [\Re(\tilde{\mathbf{s}}^T) \quad \Im(\tilde{\mathbf{s}}^T)]^T \tag{2}$$

$$\mathbf{r} := [\Re\{\tilde{\mathbf{r}}^T\} \quad \Im\{\tilde{\mathbf{r}}^T\}]^T \tag{3}$$

$$\mathbf{A} := \begin{bmatrix} \Re\{\tilde{\mathbf{A}}\} & -\Im\{\tilde{\mathbf{A}}\} \\ \Im\{\tilde{\mathbf{A}}\} & \Re\{\tilde{\mathbf{A}}\} \end{bmatrix} \tag{4}$$

$$\mathbf{n} := [\Re\{\tilde{\mathbf{n}}^T\} \quad \Im\{\tilde{\mathbf{n}}^T\}]^T \tag{5}$$

where $\Re, \Im$ denote the real and the imaginary part, respectively. Using the above vectors and matrices, we obtain the real-valued vector equation

$$\mathbf{r} = \sqrt{\frac{\rho}{n_T}}\mathbf{A}\mathbf{s} + \mathbf{n} = \mathbf{H}\mathbf{s} + \mathbf{n}. \tag{6}$$

III. Hybrid Algorithm

The hybrid algorithm consists of the following steps. As in [10], we premultiply (6) with $\mathbf{H}^T$, which yields

$$\mathbf{z} = \mathbf{H}^T\mathbf{r} = \mathbf{G}\mathbf{s} + \mathbf{v} \tag{7}$$

where $\mathbf{G} := \mathbf{H}^T\mathbf{H}$ is a symmetric positive definite[1] matrix, and $\mathbf{v} = \mathbf{H}^T\mathbf{n}$ is a noise vector with covariance matrix $\sigma^2\mathbf{G}$. We then apply one stage of the PDA detector (steps 1–5 in [8]) to the system in (7) and, thus, obtain a vector $\mathbf{p}$ that contains the associated probabilities for the elements of $\mathbf{s}$. Let D denote the subset of bits that satisfy

$$\mathbf{p}(i) \in [0, \tau] \cup [1 - \tau, 1] \tag{8}$$

with τ to be suitably chosen. $\bar{D}$ will henceforth denote the complement of D. We then make hard decisions for the bits in D, that is, set $\hat{s}_i = \text{sign}(\mathbf{p}(i) - 0.5), \forall i \in D$ and collect these decisions in a vector $\hat{\mathbf{s}}_D$. Now, expand (6) as

$$\mathbf{r} = [\mathbf{H}_D \quad \mathbf{H}_{\bar{D}}] \begin{bmatrix} \mathbf{s}_D \\ \mathbf{s}_{\bar{D}} \end{bmatrix} + \mathbf{n}$$

with obvious notation. Assuming perfect decisions for the bits in D (that is, $\hat{\mathbf{s}}_D = \mathbf{s}_D$), the residual subsystem after cancellation is

$$\mathbf{y}_{\bar{D}} := \mathbf{r} - \mathbf{H}_D\mathbf{s}_D = \mathbf{H}_{\bar{D}}\mathbf{s}_{\bar{D}} + \mathbf{n}.$$

After compacting

$$\mathbf{y}_c := \mathbf{H}_{\bar{D}}^T\mathbf{y}_{\bar{D}} = \mathbf{H}_{\bar{D}}^T\mathbf{H}_{\bar{D}}\mathbf{s}_{\bar{D}} + \mathbf{H}_{\bar{D}}^T\mathbf{n} = \mathbf{G}_{\overline{DD}}\mathbf{s}_{\bar{D}} + \mathbf{v}_{\bar{D}}$$

the noise vector $\mathbf{v}_{\bar{D}}$ is colored Gaussian with zero mean and covariance matrix $\sigma^2\mathbf{G}_{\overline{DD}}$. Introduce the Cholesky factorization

$$\mathbf{G}_{\overline{DD}} = \mathbf{L}_{\overline{DD}}^T\mathbf{L}_{\overline{DD}} \tag{9}$$

and premultiply the system with $\mathbf{L}_{\overline{DD}}^{-T}$ to obtain

$$\mathbf{x} := \mathbf{L}_{\overline{DD}}^{-T}\mathbf{y_c} = \mathbf{L}_{\overline{DD}}\mathbf{s}_{\bar{D}} + \mathbf{w} \tag{10}$$

where the noise vector $\mathbf{w}$ is white Gaussian with covariance matrix $\sigma^2\mathbf{I}$. We now apply SD to (10). Let K be the number of elements in $\bar{D}$. As suggested in [5], the initial radius for VB-SD is set to $C = aK\sigma^2$, with a such that

$$\int_0^{aK/2} \frac{x^{(K/2-1)}}{\Gamma(K/2)} e^{-x} dx = 0.99. \tag{11}$$

Alternatively, SE-SD can be used in the second stage of the hybrid algorithm. We try both VB-SD and SE-SD in our simulations.

Threshold Parameter

The threshold parameter τ should be small enough to ensure that the PDA stage makes reliable decisions. On the other hand, τ should not be too small, for otherwise, the inclusion of the PDA stage will yield little if any dimensionality reduction benefit.

While it is clear that τ should be made smaller with increasing SNR, choosing it based on analytical considerations appears intractable. Our experience is that the following choice is reasonable: $\tau = 10^{-p}$ [hard-limited within $(0, 0.45]$], with $p := 3.5((8\sigma^2)/(\rho))^{-1.55}$. This setting is well supported by our simulation results, which are reported next.

IV. Simulation Results

In our simulations, each burst comprises $L = 100$ symbol intervals. Over each symbol interval, n_T 4-QAM symbols $(\pm(1/\sqrt{2}) \pm \jmath(1/\sqrt{2}))$ are simultaneously transmitted. For each burst, a new realization of the Rayleigh channel matrix is generated. For the bit-error rate (BER) plots, we use a dynamic Monte Carlo simulation: For each SNR, the simulation stops when both the number of errors has reached 150, and the

[1]With probability 1, under the i.i.d. Rayleigh assumption.

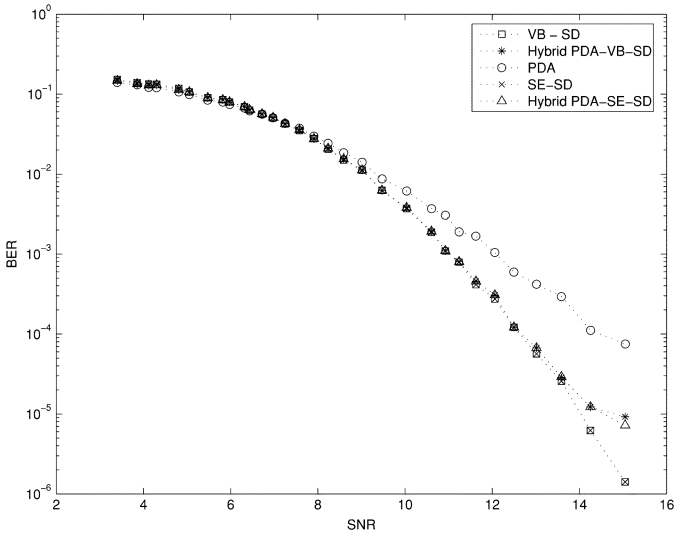


Fig. 1. Probability of error comparison for 4-QAM with $n_T = n_R = 16$. Dynamic Monte Carlo simulation.

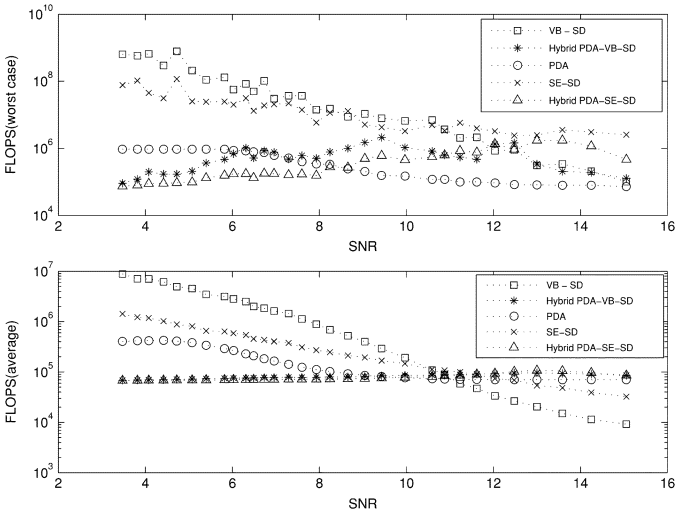


Fig. 2. Computational cost versus SNR, $n_T = n_R = 16$, 4-QAM, 10^4 Monte Carlo runs.

number of bursts has reached five. This ensures sufficient averaging in the low error rate regime while reducing unnecessarily long runs in the high error rate regime. For the computational complexity plots, we use 10^4 (100 bursts of 100 symbol vectors each) Monte Carlo runs per datum reported.

The implementation of PDA does not incorporate the bit-flip stage [8]. The internal threshold parameter of PDA is set to $\epsilon = 10^{-2}/(4\text{SNR})$ as in [8] (note that this is different from our hard decoding threshold τ). The initial radius of SD is set as in Section III; if SD fails to find a point inside the sphere, the radius is increased by one, up to five times (six searches at most). For the SE-SD algorithm, we set the search radius to infinity, which ensures that the ML solution will be found.

Fig. 1 shows the BER performance of PDA, SD, and the hybrid PDA-SD algorithm as a function of SNR $:= 10\log_{10}(\rho/\sigma^2)$, for $n_T = n_R = 16$. Fig. 2 shows the associated average and worst-case computational

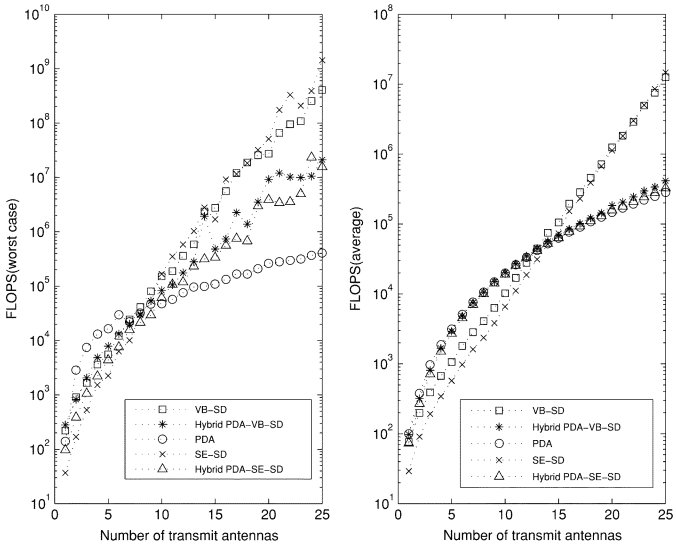


Fig. 3. Computational cost versus $n_T, n_T = n_R$, 4-QAM, SNR $= 10$ dB, 10^4 Monte Carlo runs.

costs per symbol vector, measured in Floating Point Operations (FLOPS). Finally, Fig. 3 shows FLOPS versus n_T, with $n_T = n_R$, for SNR $= 10$ dB.

V. Conclusion

We have presented a two-stage hybrid PDA-SD algorithm for signal detection in MIMO systems. The basic idea is *dimensionality reduction* via hard decoding and cancellation of those symbols that can be quickly and reliably detected via a single PDA stage. In the V-BLAST scenario considered, simulations show that the proposed hybrid algorithm attains performance close to SD, at a complexity close to PDA. The dimensionality reduction idea can also be applied in conjunction with other variants of SD or SDR.

Acknowledgment

The authors would like to thank W. Zhao and G. Giannakis (Univ. Minnesota) for sharing their MATLAB implementation of SE-SD. They also would like to thank the reviewers for their constructive comments.

References

[1] E. Agrell, T. Eriksson, A. Vardy, and K. Zeger, "Closest point search in lattices," *IEEE Trans. Inf. Theory*, vol. 48, no. 8, pp. 2201–2214, Aug. 2002.

[2] A. Chan and I. Lee, "A new reduced-complexity sphere decoder for multiple antenna systems," in *Proc. ICC*, vol. 1, New York, Apr. 28–May 2 2002, pp. 460–464.

[3] G. D. Golden, C. J. Foschini, R. A. Valenzuela, and P. W. Wolniansky, "Detection algorithm and initial laboratory results using V-BLAST space-time communication architecture," *Electron. Lett.*, vol. 35, pp. 14–15, Jan. 1999.

[4] F. Hasegawa, J. Luo, K. Pattipati, P. Willet, and D. Pham, "Speed and accuracy comparison of techniques for multiuser detection in synchronous CDMA," *IEEE Trans. Commun.*, vol. 52, no. 4, pp. 540–545, Apr. 2004.

[5] B. Hassibi and H. Vikalo, "On sphere decoding algorithm, part I: Expected complexity," *IEEE Trans. Signal Process.*, to be published.

[6] J. Jaldén and B. Ottersten, "An exponential lower bound on the expected complexity of sphere decoding," in *Proc. ICASSP*, Montreal, QC, Canada, May 17–21, 2004.

[7] G. Latsoudas, "Comparison and Speedup of Near-Optimal Decoding Algorithms for MIMO Systems," diploma thesis, Dept. Elect. Comput. Eng., Tech. Univ. Crete, May 2004.
[8] J. Luo, K. Pattipati, P. Willett, and F. Hasegawa, "Near-optimal multiuser detection in synchronous CDMA using probabilistic data association," *IEEE Commun. Lett.*, vol. 5, no. 8, pp. 361–363, Sep. 2001.
[9] W.-K. Ma, T. N. Davidson, K. M. Wong, Z.-Q. Luo, and P.-C. Ching, "Quasi-ML multiuser detection using semi-definite relaxation with application to synchronous CDMA," *IEEE Trans. Signal Process.*, vol. 50, no. 4, pp. 912–922, Apr. 2002.
[10] D. Pham, K. R. Pattipati, P. K. Willett, and J. Luo, "A generalized probabilistic data association detector for multiple antenna systems," *IEEE Commun. Lett.*, vol. 8, no. 4, pp. 205–207, Apr. 2004.
[11] E. Viterbo and J. Boutros, "A universal lattice code decoder for fading channels," *IEEE Trans. Inf. Theory*, vol. 45, no. 7, pp. 1639–1642, Jul. 1999.
[12] R. Wang and G. B. Giannakis, "Approaching MIMO capacity with reduced-complexity soft sphere-decoding," in *Proc. WCNC*, Atlanta, GA, Mar. 21–25, 2004.
[13] W. Zhao and G. B. Giannakis, "Sphere decoding algorithms with improved radius search," in *Proc. WCNC*, Atlanta, GA, Mar. 21–25, 2004.

On Downlink Beamforming With Greedy User Selection: Performance Analysis and a Simple New Algorithm

Goran Dimić, *Member, IEEE,* and Nicholas D. Sidiropoulos, *Senior Member, IEEE*

***Abstract*—This paper considers the problem of simultaneous multiuser downlink beamforming. The idea is to employ a transmit antenna array to create multiple "beams" directed toward the individual users, and the aim is to increase throughput, measured by sum capacity. In particular, we are interested in the practically important case of more users than transmit antennas, which requires user selection. Optimal solutions to this problem can be prohibitively complex for online implementation at the base station and entail so-called *Dirty Paper* (DP) precoding for known interference. Suboptimal solutions capitalize on multiuser (selection) diversity to achieve a significant fraction of sum capacity at lower complexity cost. We analyze the throughput performance in Rayleigh fading of a suboptimal greedy DP-based scheme proposed by Tu and Blum. We also propose another user-selection method of the same computational complexity based on simple zero-forcing beamforming. Our results indicate that the proposed method attains a significant fraction of sum capacity and throughput of Tu and Blum's scheme and, thus, offers an attractive alternative to DP-based schemes.**

***Index Terms*—Beamforming, downlink, multiuser diversity.**

I. Introduction

TRANSMIT antenna arrays can be utilized in two basic ways or a combination thereof: space-time coding and spatial multiplexing. The former can be used without Channel State Information (CSI) at the transmitter and allows mitigation of fading and exploitation of transmit-receive diversity. However, if CSI is known at the transmitter, higher throughput can be attained using spatial multiplexing, which can be implemented as multibeam transmit beamforming. Until recently, transmit beamforming was mostly considered for voice services in the context of the cellular downlink. With the emergence of third- and fourth-generation (3G and 4G) systems, higher emphasis is being placed on packet data, which are more delay-tolerant but require much higher throughput. Hence, we have the recent interest in transmit beamforming strategies for the cellular downlink that aim to attain the sum capacity of the wireless channel [1], [11], [13]–[16], [18], [19].

The scenario of interest can be modeled as a nondegraded Gaussian broadcast channel (GBC). Let N be the number of antennas at the transmitter [Base Station (BS) in a cellular context], and consider a cluster of M mobile users, each equipped with a single receive antenna. The channel between each transmit and receive antenna is constant over a certain time interval and is known at the BS. The received signal is corrupted by Additive White Gaussian Noise (AWGN) that is independent across users. The BS may transmit simultaneously, using multiple transmit beams, to more than one user in the cluster.

Since the receivers cannot cooperate, successful transmission critically depends on the transmitter's ability to simultaneously send independent signals with as small interference between them as possible. Caire and Shamai [1] proposed a multiplexing technique based on coding for known interference, known as "Writing on Dirty Paper," Costa precoding [2], or dirty paper (DP) coding. In [2], it is proven that in an AWGN channel with additional additive Gaussian interference, which is known at the transmitter in advance (noncausally), it is possible to achieve the same capacity as if there were no interference. Assuming Costa precoding and known channels at the transmitter, Vishwanath *et al.* [14] and Yu and Cioffi [19] have proposed algorithms that evaluate sum capacity of the GBC along with the associated optimal signal covariance matrix. However, both approaches require convex optimization in (order of) MN variables to find the optimal signal covariance matrix. Jindal *et al.* [7] have recently proposed a more efficient iterative algorithm, which requires $O(M^2N^2)$ operations per iteration.

The complexity of the aforementioned optimal strategies can be problematic for online implementation, especially when M is large. A reduced-complexity suboptimal solution to sum rate maximization is proposed in [1]. It suggests the use of QR decomposition of the channel matrix combined with DP coding at the transmitter. The combined approach nulls interference between data streams, and hence, it is named zero-forcing dirty-paper (ZF-DP) precoding. If $N \geq M$, ZF-DP is proven to be asymptotically optimal at both low and high SNR but suboptimal in general, whereas ZF beamforming without DP coding is optimal in the low SNR regime and yields the same slope of

Manuscript received May 12, 2004; revised November 26, 2004. This work was supported in part by the European Research Office (ERO) of the U.S. Army under Contract N62558-03-C-0012 and in part by the Army Research Laboratory under Cooperative Agreement DADD19-01-2-0011. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of ERO and ARL of the U.S. Army. U.S.–Greek exchange supported in part by a GSRT collaborative exchange grant. An earlier version of part of this work appears in *Proc. IEEE ICASSP 2004*, May 17-21, 2004, Montréal, QC, Canada. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Vikram Krishnamurthy.

G. Dimic was with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis MN 55455 USA. He is now an independent researcher/consultant in Belgrade, Serbia and Montenegro (e-mail: gorandimic@ieee.org).

N. D. Sidiropoulos is with the Department of Electrical and Computer Engineering, Technical University of Crete, Chania, Crete 73100, Greece (e-mail: nikos@ telecom.tuc.gr).

Digital Object Identifier 10.1109/TSP.2005.855401

throughput versus SNR in decibels as the sum capacity curve at high SNR. For the case of $N \geq M$, Spencer and Haardt [11] considered ZF beamforming without DP coding, and Samardzija and Mandayam [10] compared ZF beamforming with QR-decomposition-based spatial prefiltering coupled with DP coding.

If $N < M$, [1] has shown that random selection of $U \leq N$ users incurs significant throughput loss for both ZF-DP and ZF schemes. Tu and Blum [13] have proposed an algorithm based on ZF-DP, with a greedy user-selection procedure, named greedy ZF-DP (gZF-DP). In [13], it is shown by simulations that the throughput of gZF-DP is a significant fraction of the sum capacity. This is achieved by means of *multiuser diversity*. For the case of $N \leq M$, Viswanathan *et al.* [16] considered the problem of achieving any point in the capacity region and not only maximum sum capacity. They proposed ZF beamforming coupled with a user-selection scheme that schedules N users using an exhaustive search over a set of K_T users with the highest *individual* SINR $(N \leq K_T \leq M)$. The throughput of this scheme was compared to the throughput of a DP-coding-based optimal algorithm, and it was reported that as K_T approaches M, the throughput of ZF with exhaustive user selection comes close to the throughput of the optimal algorithm when each receiver has one antenna [16].

An important shortcoming of DP coding is that it requires vector coding, and depending on the SNR, it may require long temporal block lengths to be well approximated in practice. In particular, the required block length decreases as SNR increases, with a block length of one being adequate at sufficiently high SNR. At low and moderate SNR, a good approximation of DP can be computationally demanding with the current state-of-art [8], [18], [20]. For this reason, we advocate herein a more pragmatic approach, based on plain ZF beamforming.

Our goal is to investigate low-complexity downlink beamforming solutions that come close to attaining sum capacity for the practically important case wherein the number of downlink users (M) is larger than the number of transmit antennas (N), which entails user selection. Our aim is three-fold: i) Analyze gZF-DP to better understand the effects of multiuser diversity; ii) propose a simpler greedy alternative, based on ZF beamforming and dubbed ZFS, which does not use DP coding; and iii) assess the performance of both gZF-DP and ZFS relative to sum capacity. The key idea is that multiuser diversity can largely make up for the use of simple linear processing in lieu of more complex schemes. The performance analysis of gZF-DP is useful in system design, and ZFS is appealing from a practical standpoint. In particular, we will show that the complexity of the selection procedure of the proposed algorithm is the same as that of gZF-DP. Our simulation results indicate that at moderate and high SNR, ZFS has equal slope of throughput versus SNR as the gZF-DP and the capacity curve. It achieves a significant fraction of throughput of the gZF-DP algorithm and remains close to sum capacity for all SNR for a small to moderate number of transmit antennas.

We note that an inherent drawback of the maximum sum capacity criterion is the lack of fairness guarantees, at least in the short run. While this could be compensated over a longer timeline due to channel variations, it remains that certain users may be completely shut off during a scheduling epoch. Whether this is appropriate or not depends on the context; on this issue, see also [1], [11], [13]–[16], [18], and [19].

The rest of the paper is organized as follows. The problem of sum rate maximization is formulated in Section II. This is followed by a review of the gZF-DP algorithm, a description of the proposed ZFS algorithm, and a comparison of the complexities of the two algorithms in Section III. In Section IV, the throughput performance of the gZF-DP algorithm in independent Rayleigh fading is analyzed. Simulation-based comparison of the throughput performances of gZF-DP and ZFS is provided in Section V. Conclusions are drawn in Section VI.

II. Problem Formulation

Let $h_{m,n}$ model the quasistatic, flat-fading channel between transmit antenna n and the receive antenna of user m, and denote $\mathbf{h}_m := [\, h_{m,1} \quad h_{m,2} \quad \dots \quad h_{m,N} \,]$. Note that $\mathbf{h}_m$ is a row vector. Thus, the channel matrix $\mathbf{H}$ is

$$\mathbf{H} = [\, \mathbf{h}_1^* \quad \mathbf{h}_2^* \quad \dots \quad \mathbf{h}_M^* \,]^* \tag{1}$$

where $(\cdot)^*$ denotes conjugate-transpose. $\text{rank}(\mathbf{H}) = \min(N, M)$ with probability 1, due to the assumed statistical independence and continuous distribution of the channel vectors. Throughout the paper, we are interested in the case $N < M$ so that we assume that $rank(\mathbf{H}) = N$. Collecting the baseband-equivalent outputs, the received signal vector is

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

where $\mathbf{y}$ is the transmitted signal vector, and $\mathbf{z}$ is the noise vector. The signal covariance matrix is $\mathbf{C}_y = E[\mathbf{y}\mathbf{y}^H]$. The total transmit power is constrained to P. The sum capacity of such a vector Gaussian broadcast channel is [15]

$$C = \sup_{\mathbf{C}_y \in \mathcal{A}} \log\det(\mathbf{I} + \mathbf{H}\mathbf{C}_y\mathbf{H}^*) \tag{3}$$

where $\boldsymbol{\mathcal{A}}$ is the set of N by N non-negative diagonal matrices $\mathbf{C}_y$ with $\text{Trace}[\mathbf{C}_y] \leq P$.

Using only linear spatial processing at the transmitter, which is a suboptimal strategy, we obtain the following model. Let $\mathbf{w}_m = [w_{1,m}\ w_{2,m}\ \dots\ w_{N,m}]^T$ ($(\cdot)^T$ denotes transpose) be the beamforming weight vector for user m. The beamforming weight matrix $\mathbf{W}$ is

$$\mathbf{W} = [\mathbf{w}_1\ \mathbf{w}_2\ \cdots\ \mathbf{w}_M]. \tag{4}$$

Collecting the baseband-equivalent outputs, the received signal vector is

$$\mathbf{x} = \mathbf{H}\mathbf{W}\mathbf{D}\mathbf{s} + \mathbf{z} \tag{5}$$

where $\mathbf{s}$ is the transmitted signal vector containing uncorrelated unit-power entries, and

$$\mathbf{D} = \begin{bmatrix} \sqrt{p_1} & 0 & \cdots & 0 \\ 0 & \sqrt{p_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{p_M} \end{bmatrix} \tag{6}$$

accounts for power loading (the columns of $\mathbf{W}$ are thus normalized to unit norm). Note that the elements of $\mathbf{x}$ are physically distributed across the M mobile terminals. Multiuser decoding

is therefore not feasible; hence, each user treats the signals intended for other users as interference. Noise is assumed to be circular complex Gaussian, zero-mean, and uncorrelated with variance of each complex entry $\sigma^2 = 1$.

The desired signal power received by user m is given by $|\mathbf{h}_m \mathbf{w}_m|^2 p_m$. The Signal-to-Interference plus Noise Ratio (SINR) of user m is

$$\text{SINR}_m = \frac{|\mathbf{h}_m \mathbf{w}_m|^2 p_m}{\sum\limits_{i \neq m} |\mathbf{h}_m \mathbf{w}_i|^2 p_i + \sigma^2}. \tag{7}$$

The linear beamforming problem can now be formulated as

$$\boxed{\begin{array}{c} \max\limits_{\mathbf{W},\mathbf{D}} \sum\limits_{m=1}^{M} \log_2(1 + \text{SINR}_m) \\ \text{subject to} \|\mathbf{WD}\|_F^2 \leq P \end{array}} \tag{8}$$

where $\|\cdot\|_F$ denotes Frobenius norm, and P stands for a bound on average transmitted power.

Attaining capacity requires Gaussian signaling and long codes, yet the logarithmic SINR reward can be motivated from other, more practical perspectives as well. It can be shown that it measures the throughput of QAM-modulated systems over both AWGN and Rayleigh fading channels. The intuition is that SINR improvements eventually yield diminishing throughput returns.

III. Reduced-Complexity Algorithms

A. Greedy Zero-Forcing Dirty-Paper Algorithm

In [1], Caire and Shamai have proposed a suboptimal solution to (3) based on the QR-type decomposition [6] of the channel matrix $\mathbf{H} = \mathbf{LQ}$ obtained by applying Gram–Schmidt orthogonalization to the rows of $\mathbf{H}$. $\mathbf{L}$ is a lower triangular matrix, and $\mathbf{Q}$ has orthonormal rows. Setting $\mathbf{W} = \mathbf{Q}^*$, (5) yields a set of interference channels

$$x_m = l_{m,m}\sqrt{p_m} s_m + \sum_{j<m} l_{m,j}\sqrt{p_j} s_j + z_m, \quad m = 1, \ldots, N \tag{9}$$

while no information is sent to users $m = N+1, \ldots, M$. In order to eliminate the interference term $I_m = \sum_{j<m} l_{m,j}\sqrt{p_j} s_j$, the input signals $\sqrt{p_m} s_m$, for $m = 1, \ldots, N$ are obtained by successive application of DP coding, where for each m, the interference I_m is noncausally known. This particular choice of precoding matrix $\mathbf{W} = \mathbf{Q}^*$ nulls interference caused by users $j > m$ and DP coding nulls interference caused by users $j < m$ so that the scheme forces all interference to zero. Hence, it was dubbed ZF-DP coding. The throughput of the ZF-DP scheme is given by [1]

$$R_{\text{zfdp}} = \sum_{m=1}^{N} [\log_2(\mu d_m)]_+ \tag{10}$$

where $[x]_+ = \max\{0, x\}$, $d_n := |l_{n,n}|^2$, and μ is the solution of the water-filling equation

$$\sum_{m=1}^{N} \left[\mu - \frac{1}{d_m}\right]_+ = P. \tag{11}$$

Then, for $m = 1, \ldots, N$

$$p_m = d_m \left[\mu - \frac{1}{d_m}\right]_+. \tag{12}$$

Note that when $N < M$, one has to select up to N out of M users whose data will be transmitted. In general, different selections yield different values of R_{zfdp} in (10). Furthermore, different ordering within the same set of users yields different sum rate. The ZF-DP scheme does not attempt to optimize the throughput with respect to either user selection or ordering. In [13], Tu and Blum have proposed a greedy algorithm for the selection of N out of M rows of the channel matrix $\mathbf{H}$ and ordering of the selected rows in the Gram-Schmidt orthogonalization, aiming to maximize the throughput. The algorithm is called greedy ZF-DP and is presented here for convenience.

Let $U = \{1, 2, \ldots, M\}$ denote the set of indices of all M users, and let $S_n = \{s_1, \ldots, s_n\} \subset U$ denote the set of n selected users ($|S_n| = n$).

1) **Initialization:**
- Set $n = 1$.
- Let $r_{1,u} = \mathbf{h}_u \mathbf{h}_u^*$. Find a user s_1 such that $s_1 = \arg\max_{u \in U} r_{1,u}$.
- Set $S_1 = \{s_1\}$.

2) **While** $n < N$:
- Increase n by 1.
- Project each remaining channel vector onto the orthogonal complement of the subspace spanned by the channels of the selected users. The projector matrix is

$$\mathbf{P}_n^{\perp} = \mathbf{I}_N - \mathbf{H}(S_{n-1})^* \left(\mathbf{H}(S_{n-1})\mathbf{H}(S_{n-1})^*\right)^{-1} \mathbf{H}(S_{n-1}) \tag{13}$$

where $\mathbf{I}_N$ is the $N \times N$ identity matrix, and $\mathbf{H}(S_{n-1})$ denotes the row-reduced channel matrix consisting of the channel vectors of the users selected in the first $n-1$ steps

$$\mathbf{H}(S_{n-1}) = [\mathbf{h}_{s_1}^* \quad \mathbf{h}_{s_2}^* \quad \ldots \quad \mathbf{h}_{s_{n-1}}^*]^*. \tag{14}$$

Let $r_{n,u} = \left|\mathbf{h}_u \mathbf{P}_n^{\perp}\right|^2$. Due to idempotence of $\mathbf{P}_n^{\perp}$, we have

$$r_{n,u} = \mathbf{h}_u \mathbf{P}_n^{\perp} \mathbf{h}_u^*. \tag{15}$$

- Find a user s_n such that

$$s_n = \arg \max_{u \in U \setminus S_{n-1}} r_{n,u}. \tag{16}$$

- Set $S_n = S_{n-1} \cup \{s_n\}$.

3) **Beamforming:** Let $\mathbf{W} = \mathbf{Q}^*$, where $\mathbf{H}(S_N) = \mathbf{LQ}$ is the QR-type decomposition of $\mathbf{H}(S_N)$.

4) **DP coding**: Applied to the rows of $\mathbf{L}$.

Power Loading: Water-filling.

The rows $\mathbf{q}_m$ of $\mathbf{Q}$ in the QR decomposition of $\mathbf{H}(S_n) = \mathbf{L}(S_n)\mathbf{Q}(S_n)$ are obtained by applying Gram–Schmidt orthogonalization to the ordered rows of $\mathbf{H}(S_n)$: $\mathbf{h}_{s_1}, \ldots, \mathbf{h}_{s_n}$. This yields [1]

$$\mathbf{h}_{s_n} = \sqrt{\mathbf{h}_{s_n} \mathbf{P}_n^{\perp} \mathbf{h}_{s_n}^*}\, \mathbf{q}_{s_n} + \sum_{j \in S_{n-1}} \mathbf{h}_{s_n} \mathbf{q}_j^* \mathbf{q}_j. \tag{17}$$

From $\mathbf{L}(S_n) = \mathbf{H}(S_n)\mathbf{Q}(S_n)^*$, we obtain $l_{n,n} = \mathbf{h}_{s_n}\mathbf{q}_{s_n}^*$. By definition of d_n (10), orthonormality of $\mathbf{Q}(S_n)$, and (17), we have

$$d_n = |l_{n,n}|^2 = \mathbf{h}_{s_n}\mathbf{P}_n^{\perp}\mathbf{h}_{s_n}^*.$$

From (15) and (16), it follows that

$$d_n = \max_{u \in U \setminus S_{n-1}} r_{n,u} \tag{18}$$

for $n = 1, \ldots, N$. In other words, the gZF-DP algorithm maximizes d_n, conditioned on the choice of $d_1, \ldots, d_{n-1}$.

B. ZF With User Selection

ZF beamforming inverts the channel matrix at the transmitter so that orthogonal channels between transmitter and receivers are created. It is then possible to encode users individually, as opposed to the more complex long-block-vector coding generally needed to implement DP. Note that ZF at the transmitter does not enhance noise at the receiver, but it incurs an excess transmission power penalty relative to ZF-DP. If $M \leq N$, and $rank(\mathbf{H}) = M$, then the ZF beamforming matrix is

$$\mathbf{W} = \mathbf{H}^*(\mathbf{H}\mathbf{H}^*)^{-1} \tag{19}$$

which is the Moore-Penrose pseudoinverse of the channel matrix. However, if $M > N$, it is not possible to use (19) because $\mathbf{H}\mathbf{H}^*$ is singular. In that case, one needs to select $n \leq N$ out of M users.

For $M > N$, the problem (8) is reformulated as follows: Given $\mathbf{H} \in \mathbb{C}^{M \times N}$, select $n \leq N$ and a set of channels $\{\mathbf{h}_{s_1}, \ldots, \mathbf{h}_{s_n}\}$, which produce the row-reduced channel matrix

$$\mathbf{H}(S_n) = [\,\mathbf{h}_{s_1}^* \quad \mathbf{h}_{s_2}^* \quad \ldots \quad \mathbf{h}_{s_n}^*\,]^*$$

such that the sum rate is the highest achievable:

$$\max_{1 \leq n \leq N} \quad \max_{S_n} R_{zf}(S_n)$$
$$\text{subject to} \quad \sum_{i \in S_n} \left[\mu - \frac{1}{c_i(S_n)}\right]_+ = P. \tag{20}$$

The throughput of ZF algorithm is given by [1]

$$R_{zf}(S_n) = \sum_{i \in S_n} [\log_2(\mu c_i(S_n))]_+ \tag{21}$$

where

$$c_i(S_n) = \left\{ \left[(\mathbf{H}(S_n)\mathbf{H}(S_n)^*)^{-1} \right]_{i,i} \right\}^{-1} \tag{22}$$

and μ is obtained by solving the water-filling equation in (20). The power-loading then yields

$$p_i = c_i(S_n)\left[\mu - \frac{1}{c_i(S_n)}\right]_+, \quad \forall\, i \in S_n. \tag{23}$$

The problem can be conceptually solved by exhaustive search: For each value of n, find all possible n-tuples S_n and select a pair (n, S_n), which yields maximum $R_{zf}(S_n)$. However, such an algorithm has prohibitive complexity.

We propose a reduced-complexity suboptimal algorithm, dubbed ZF with Selection (ZFS), as outlined next.

1) **Initialization:**
- Set $n = 1$.
- Find a user, s_1, such that

$$s_1 = \arg\max_{u \in U} \mathbf{h}_u\mathbf{h}_u^*.$$

- Set $S_1 = \{s_1\}$ and denote the achieved rate $R_{zf}(S_1)_{\max}$.

2) **While** $n < N$:
- Increase n by 1.
- Find a user, s_n, such that

$$s_n = \arg\max_{u \in U \setminus S_{n-1}} R_{zf}(S_{n-1} \cup \{u\}).$$

- Set $S_n = S_{n-1} \cup \{s_n\}$, and denote the achieved rate $R_{zf}(S_n)_{\max}$.
- If $R_{zf}(S_n)_{\max} \leq R_{zf}(S_{n-1})_{\max}$ **break**, and decrease n by 1

3) **Beamforming:** $\mathbf{W} = \mathbf{H}(S_n)^*(\mathbf{H}(S_n)\mathbf{H}(S_n)^*)^{-1}$
Power Loading: Water-filling.

C. Complexity and Implementation

We consider complexity of the user selection procedure only. The complexity of DP coding, required by the gZF-DP algorithm, depends on its implementation, in particular, the degree of approximation and the associated spatio-temporal block length (which is a function of SNR), cf. [4], [18].

Complexity of the user selection procedure of the gZF-DP algorithm is $O(N^3M)$. To see this, note that for each $n \leq N$, the algorithm evaluates $M - n + 1$ 2-norms $r_{n,u}$. Evaluation of $r_{n,u}$ involves a vector-matrix multiplication, where the vector is $1 \times N$ and the matrix $N \times N$. The complexity of this step is $O(N^2)$. Repeating this over $O(M)$ users in N steps, we obtain $O(N^3M)$.

We will show that the complexity of the user selection procedure of the ZFS algorithm is also $O(N^3M)$. Again, for each $n \leq N$, the ZFS algorithm evaluates $M - n + 1$ rates $R_{zf}(S_{n-1} \cup \{u\})$. The evaluation of $R_{zf}(S_{n-1} \cup \{u\})$ is split into the evaluation of the $c_i(S_{n-1} \cup \{u\})$'s followed by evaluation of μ; cf. (21). An efficient way to evaluate the $c_i(S_{n-1} \cup \{u\})$'s is by using the matrix inversion lemma to invert the matrix $\mathbf{A}(S_{n-1} \cup \{u\}) := \mathbf{H}(S_{n-1} \cup \{u\})\mathbf{H}(S_{n-1} \cup \{u\})^*$. Note that

$$\mathbf{A}(S_{n-1} \cup \{u\}) = \begin{bmatrix} \mathbf{A}(S_{n-1}) & \mathbf{a}_u \\ \mathbf{a}_u^* & a_{u,u} \end{bmatrix}$$

where $\mathbf{a}_u = [\mathbf{h}_{s_1}\mathbf{h}_u^*,\ \mathbf{h}_{s_2}\mathbf{h}_u^*,\ \ldots\ \mathbf{h}_{s_{n-1}}\mathbf{h}_u^*]^T$, and $a_{u,u} = \mathbf{h}_u\mathbf{h}_u^*$. Noting that $\mathbf{A}(S_{n-1})^* = \mathbf{A}(S_{n-1})$ and writing

$$\mathbf{q} = \mathbf{A}(S_{n-1})^{-1}\mathbf{a}_u \tag{24}$$

after some algebraic manipulation, we obtain

$$\mathbf{A}(S_{n-1} \cup \{u\})^{-1} = \begin{bmatrix} \mathbf{A}(S_{n-1})^{-1} & \mathbf{0}_{n-1} \\ \mathbf{0}_{n-1}^T & 0 \end{bmatrix} + (a_{u,u} - \mathbf{a}_u^*\mathbf{q})^{-1} \begin{bmatrix} \mathbf{q}\mathbf{q}^* & -\mathbf{q} \\ -\mathbf{q}^* & 1 \end{bmatrix} \tag{25}$$

where $\mathbf{0}_{n-1}^T = [0 \quad 0 \quad \ldots \quad 0]_{1 \times (n-1)}$. It can be verified that each time n is increased, $\mathbf{A}(S_{n-1})^{-1}$ and $a_{i,u},\ i \in S_{n-2}$ are

known before the search over $u \in U \setminus S_{n-1}$ starts. Hence, evaluation of $\mathbf{A}(S_{n-1} \cup \{u\})^{-1}$ from (24) and (25) has complexity proportional to $O(n^2)$. Repeating this over $O(M)$ users in each of $n \leq N$ steps, we obtain the overall complexity of the user-selection procedure of the ZFS algorithm to be $O(N^3M)$.

It can be shown that the per-iteration complexity of the sum power iterative water-filling algorithm proposed by Jindal *et al.* [7] is $O(N^2M^2)$. Therefore, the gZF-DP and ZFS algorithms have significantly lower computational complexity than the sum power iterative water-filling algorithm if $M \gg N$.

In the following, we pay attention to the substeps in step 2) of the ZFS algorithm. Given a set S_n, we have [1]

$$c_i(S_n) = |\mathbf{h}_{s_i}\mathbf{P}(S_n \setminus \{s_i\})^{\perp}|^2 \tag{26}$$

where $\mathbf{P}(S_n)^{\perp}$ denotes the projector onto the orthogonal complement of $\Omega(S_n) = \text{span}\{\mathbf{h}_{s_l} : s_l \in S_n\}$. Note that $c_j(S_{n-1} \cup \{u\}) \leq c_j(S_{n-1})$ for every user $j \in S_{n-1}$. This is due to (26) and $S_{n-1} \subset S_{n-1} \cup \{u\}$. Therefore, if (20) and (23) yield $p_u = 0$, then $R_{zf}(S_{n-1} \cup \{u\}) < R_{zf}(S_{n-1})$. We discard such u. We also discard u if (20) and (23) yield $p_{s_i} = 0$ for some $s_i \in S_{n-1}$. This is done to keep complexity at bay for otherwise, combinatorial search might effectively emerge. Hence, user u is a candidate for S_n if $p_i > 0, \forall\, i \in S_{n-1} \cup \{u\}$. From the properties of water-filling, this holds if

$$\frac{n}{c_{i_{\min}}(S_{n-1} \cup \{u\})} < P + \sum_{i \in S_{n-1} \cup \{u\}} \frac{1}{c_i(S_{n-1} \cup \{u\})} \tag{27}$$

where $c_{i_{\min}}(S_{n-1} \cup \{u\}) = \min_{i \in S_{n-1} \cup \{u\}} c_i(S_{n-1} \cup \{u\})$. Then, we have

$$\mu = \frac{1}{n}\left[P + \sum_{i \in S_{n-1} \cup \{u\}} \frac{1}{c_i(S_{n-1} \cup \{u\})}\right]. \tag{28}$$

If (27) is not satisfied, we skip to the next u.

We note that the **break** in Step 2 is necessary when ZFS is used but redundant when ZF-DP is used; it is shown in [1] and [13] that in the latter case, maximum sum rate can always be achieved with N active users if $P > 0$ [1]. On the other hand, when ZF alone is used, the optimum number of active users is $n_{opt} \leq N$ and decreases as P decreases, so that for $P \to 0$, the ZF scheme reduces to maximum ratio combining (MRC) $n_{opt} = 1$ [1]. This also holds for the proposed ZFS algorithm, which follows from the water-filling equation in (20) and the fact that $c_1(S_1) = \max_{i \in U} a_{i,i}$.

IV. Performance Analysis in Independent Rayleigh Fading

In this section, we evaluate the throughput of the greedy ZF-DP algorithm [13] in independent Rayleigh fading when channels remain constant over the duration of a transmission of a block of symbols. The channels of all M users are assumed to have i.i.d. entries, which are circularly symmetric, zero-mean, complex Gaussian random variables (r.v.s) with unit variance $h_{m,n} \sim \mathcal{CN}(0,1)$. In [1], the average throughput of the ZF-DP and ZF schemes in independent Rayleigh fading under a long-term power constraint for general N and M is evaluated. As noted earlier, the simple ZF-DP and ZF algorithms in [1] do not attempt to optimize throughput with respect to user selection and ordering when $M > N$. Instead, users are selected and ordered randomly.

A. gZF-DP Sum Rate Under Long-Term Power Constraint

We model the greedy ZF-DP algorithm [13] under a long-term power constraint. We are interested in evaluating

$$R_{gZF\text{-}DP} = E\left[\sum_{i=1}^{N}[\log(\mu_o d_i)]_+\right] \tag{29}$$

where μ_o is the solution of the water-filling equation, stemming from the long-term (LT) power constraint

$$E\left[\sum_{i=1}^{N}\left[\mu - \frac{1}{d_i}\right]_+\right] = P. \tag{30}$$

Note that the optimum μ_o determined by (30) will be a *deterministic function* of the statistics of the d_i's and not a function of the random variables themselves. By this and linearity of expectation, we can rewrite (29) as

$$\begin{aligned} R_{gZF\text{-}DP} &= \sum_{i=1}^{N} E[\log(\mu_o d_i)]_+ \\ &= \sum_{i=1}^{N}\int_0^{\infty}[\log(\mu_o x)]_+ f_{d_i}(x)dx. \end{aligned}$$

Therefore

$$R_{gZF\text{-}DP} = \sum_{i=1}^{N}\int_{1/\mu_o}^{\infty}\log(\mu_o x) f_{d_i}(x)dx \tag{31}$$

where $f_{d_i}(x)$ denotes the probability density function (pdf) of d_i. Similarly, (30) becomes

$$\sum_{i=1}^{N}\left(\mu\int_{1/\mu}^{\infty} f_{d_i}(x)dx - \int_{1/\mu}^{\infty}\frac{1}{x} f_{d_i}(x)dx\right) = P. \tag{32}$$

In order to evaluate R, we need to evaluate the pdfs of d_i's based on the knowledge of channel statistics and selection procedure. Our derivation below draws in part from performance analysis tools in [5], [17], which we tailor to fit the context of gZF-DP. In particular, our analysis accounts for and exploits the specific selection procedure employed in gZF-DP.

B. Probability Density Functions

It is instructive to consider the modeling of the pdf of d_1 first, followed by modeling the pdf of d_2, and then generalizing to compute the pdf of d_n for general $n \leq N$. First, let us determine the distribution of $r_{1,u} = \mathbf{h}_u\mathbf{h}_u^*$. Note that $r_{1,u}$ is a sum of N squared magnitudes of circularly symmetric, zero-mean, unit-variance complex Gaussian random variables. Therefore, it has Chi-squared distribution with $2N$ degrees of freedom ($r_{1,u} \sim \chi^2_{2N}$), whose pdf is

$$f_{r_1}(x_1) = \frac{1}{\Gamma(N)}x_1^{N-1}\exp(-x_1). \tag{33}$$

$\Gamma(N)$ denotes the Gamma function, and $\Gamma(N) = (N-1)!$ for a positive integer N. According to the selection algorithm

$$d_1 = \max_{u \in U} r_{1,u}. \tag{34}$$

From order statistics, e.g., [3, (2.1.1)], we obtain the pdf of d_1 as

$$f_{d_1}(x_1) = M\left[F_{r_1}(x_1)\right]^{M-1} f_{r_1}(x_1) \tag{35}$$

where $F_{r_1}(x_1)$ is the cumulative distribution function (cdf) of $r_{1,u}$. We say that the distribution of $r_{1,u}$ is the *parent distribution* of the order statistics $r_{1,(1)} \geq r_{1,(2)} \geq \cdots \geq r_{1,(M)}$, where $r_{1,(i)}$ is the ith largest $r_{1,u}$ for $u \in U$.

Noting that $r_{1,u} \leq d_1$, for all of the remaining users ($u \in U \setminus S_1$), it follows that the posterior distribution of $r_{1,u}$ of the remaining users (after selecting user s_1) depends on the realization of d_1. In the sequel, we will need to use the conditional pdf of $r_{1,u}$ of the remaining users given a realization of d_1. According to (34) and, e.g., [3, Th. 2.7], the parent distribution of the order statistics of the remaining users $u \in U \setminus S_1$ is equal to $f_{r_1}(x_1)$ truncated on the right at the value of d_1

$$f_{r_1|d_1}(x_1|y_1) = \begin{cases} \frac{f_{r_1}(x_1)}{F_{r_1}(y_1)}, & \text{if } x_1 \in [0, y_1] \\ 0, & \text{otherwise.} \end{cases} \tag{36}$$

After setting $n = 2$, the selection algorithm proceeds by projecting the channel vectors of all of the remaining users onto the orthogonal complement of the subspace spanned by the channel vector of user s_1. From (15), we have $r_{2,u} = \mathbf{h}_u \mathbf{P}_2^{\perp} \mathbf{h}_u^*$, for $u \in U \setminus S_1$, where $\mathbf{P}_2^{\perp}$ is given in (13). The distribution of $r_{2,u}$ given d_1, which is denoted $f_{r_2|d_1}(x_2|y_1)$, then becomes the parent distribution of the order statistics $r_{2,(i)}$, given d_1 for $i \geq 2$. Therefore, we need a mapping from $f_{r_1|d_1}(x_1|y_1)$ to $f_{r_2|d_1}(x_2|y_1)$ that models the projection step

$$f_{r_2|d_1}(x_2|y_1) = \int_0^{\infty} f_{r_2|r_1,d_1}(x_2|v, y_1) f_{r_1|d_1}(v|y_1) dv. \tag{37}$$

Here, $f_{r_2|r_1,d_1}(x_2|v, y_1)$ denotes the pdf of $r_{2,u}$, given realizations of $r_{1,u}$ and d_1. Note that $r_{2,u} \leq r_{1,u} \leq d_1$. $\mathbf{h}_{s_1}$ is statistically independent of $\mathbf{h}_u$, for $u \in U \setminus S_1$, so that from the point of view of the users in $U \setminus S_1$, $\mathbf{P}_2^{\perp}$ appears to be a randomly selected projector matrix. However, the first user has been selected after considering the channels of *all* users, and thus, there might be mild dependence between the channels of the remaining users in $U \setminus S_1$ and $\mathbf{P}_2^{\perp}$. For analytical tractability, we will ignore this dependence. Our simulation results will fully corroborate this approximation: The difference is not even noticeable in simulations.

Assumption 1: We therefore assume that d_1 conveys no information about $\mathbf{P}_2^{\perp}$, i.e., $f_{r_2|r_1,d_1}(x_2|v, y_1)$ has the Markovian property

$$f_{r_2|r_1,d_1}(x_2|v, y_1) = f_{r_2|r_1}(x_2|v). \tag{38}$$

The pdf $f_{r_2|r_1}(x_2|v)$ is obtained from the following.

Claim 1: Let $\mathbf{h} = [h_1 \ \ldots \ h_N]$ and $\mathbf{p} = [p_1 \ \ldots \ p_N]$ denote independent N-dimensional random (row-) vectors with i.i.d., circularly symmetric, zero-mean, complex Gaussian entries with unit variance $h_n \sim \mathcal{CN}(0,1)$ and $p_n \sim \mathcal{CN}(0,1)$. Let $Y := \mathbf{h}\mathbf{h}^*$ and $X := \mathbf{h}\mathbf{P}\mathbf{h}^*$, where $\mathbf{P} = \mathbf{I}_N - \mathbf{p}^* \left(\mathbf{p}\mathbf{p}^*\right)^{-1} \mathbf{p}$ [cf. (13)] is an $N \times N$ projector matrix with $N-1$ eigenvalues equal to 1 and one eigenvalue equal to 0. Then, the cdf of X, given Y, is given by

$$F_{X|Y}(x|y) = \begin{cases} \left(\frac{x}{y}\right)^{N-1}, & \text{for } x \in [0, y] \\ 0, & \text{elsewhere.} \end{cases} \tag{39}$$

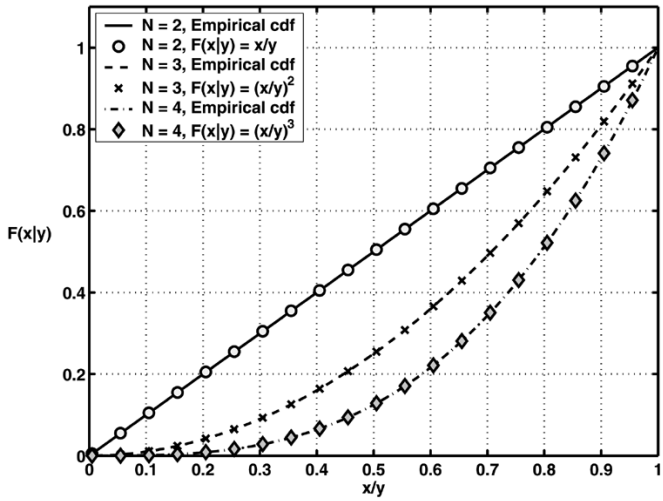


Fig. 1. cdf $F_{r_{n+1}|r_n}(x|y)$ when $y : 1 \times N$ channel vector.

Remark 1: The rigorous proof of this claim turned out to be elusive, but it is *very* well supported by simulations. Fig. 1 depicts $F_{X|Y}(x|y)$ versus x/y for $N = 2$, 3, and 4. Lines show empirical cdfs obtained by Monte Carlo (MC) simulations, and markers show samples of analytic curves given by (39). In MC simulations, for each value of N, there were 2×10^5 random realizations of $\mathbf{P}$ given $\mathbf{h}$, for 10^2 realizations of $\mathbf{h}$. The empirical $F_{X|Y}(x|y)$ is discrete. Its support $x/y \in [0, 1]$ is divided into 200 intervals of length 1/200. The match in Fig. 1 is very accurate.

From (39), we obtain

$$f_{r_2|r_1}(x_2|v) = \begin{cases} \frac{N-1}{v} \left(\frac{x_2}{v}\right)^{N-2}, & \text{for } x_2 \in [0, v], \\ 0, & \text{otherwise.} \end{cases} \tag{40}$$

From (18), it follows that d_2, conditioned on a realization of d_1, is the maximum of $M-1$ r.v.s with the parent distribution given by the pdf $f_{r_2|d_1}(x_2|x_1)$ from (37). Using order statistics, we obtain [3]

$$f_{d_2|d_1}(x_2|x_1) = (M-1)\left[F_{r_2|d_1}(x_2|x_1)\right]^{M-2} f_{r_2|d_1}(x_2|x_1). \tag{41}$$

Since $f_{r_2|d_1}(x_2|x_1) = 0$ for $x_2 > x_1$, it follows that $d_2 \leq d_1$. Finally

$$f_{d_2}(x_2) = \int_{x_1=0}^{\infty} f_{d_2|d_1}(x_2|x_1) f_{d_1}(x_1) dx_1 \tag{42}$$

for $x_1 \geq x_2$.

Armed with these insights, we can now generalize to the computation of the pdf of d_n for $n \leq N$. The associated derivation

is deferred to the Appendix. Using the results of Section IV-A, the pdf of d_n is obtained as a marginal distribution:

$$f_{d_n}(x_n) = \int_{x_1=x_n}^{\infty} \int_{x_2=x_n}^{x_1} \cdots \int_{x_{n-1}=x_n}^{x_{n-2}} \cdot f_{d_n|d_{n-1},\ldots,d_1}(x_n|x_{n-1},\ldots,x_1) \cdot \prod_{k=2}^{n-1} f_{d_k|d_{k-1},\ldots,d_1}(x_k|x_{k-1},\ldots,x_1) \cdot f_{d_1}(x_1)dx_n \ldots dx_1 \quad (43)$$

for $x_1 \geq x_2 \geq \cdots x_n$.

The pdfs of d_n for $n = 1,\ldots,N$ can be written in a more compact form, facilitating analysis and numerical integration.

Proposition 1: Define

$$\phi_1(x_1) = f_{r_1}(x_1) \quad (44)$$

and

$$\phi_n(x_n, x_{n-1},\ldots,x_1) = \frac{1}{\Gamma(N-n+1)} x_n^{N-n} \cdot \int_{v_{n-1}=x_n}^{x_{n-1}} \int_{v_{n-2}=v_{n-1}}^{x_{n-2}} \cdots \int_{v_1=v_2}^{x_1} \exp(-v_1) \cdot dv_1 \ldots dv_{n-1}. \quad (45)$$

Then, we have (46), shown at the bottom of the page. The proof is given in the Appendix. We will use the forms in the above proposition in the Proof of Theorem 1, whose statement appears in Section IV-C.

Fig. 2 depicts an example of pdfs of d_n for $N = 4$ and $M = 8$. Full lines depict analytically obtained pdfs. Markers show samples of the empirically obtained pdfs through Monte Carlo (MC) simulations. There are 10^6 MC samples. For every d_n, the support of the empirical pdf is truncated where the tail becomes insignificant. Then, the empirical pdf is discretized by dividing the truncated support into 100 equal intervals. These results justify the approximation (Assumption 1) made in the course of an analytical derivation for tractability considerations.

C. Throughput of gZF-DP at High SNR

Let $R_{gZF\text{-}DP}$ denote the average throughput of the gZF-DP algorithm. Let $\rho = 10\log_{10} P$ denote the SNR, where the noise variance of each user is assumed equal to 1. We have the following result.

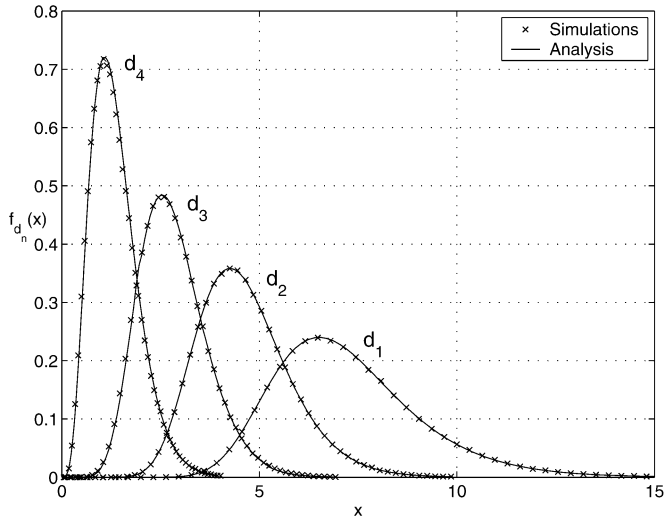


Fig. 2. Family of pdfs of d_n for $N = 4$, $M = 8$.

Theorem 1: Let $N < M$, and let P be the power limit. Then, under our working assumptions

$$\lim_{P\to\infty} \frac{\partial}{\partial \rho} R_{gZF\text{-}DP} = N \frac{\log_2 10}{10} \left[\frac{\text{bits}}{\text{dB}}\right]. \quad (47)$$

The proof is given in the Appendix. The above theorem shows that the throughput versus SNR slope of the gZF-DP algorithm in the high SNR regime is proportional to the number of antennas at the transmitter N. Note that this is the theoretical limit of the capacity versus SNR slope for a multiple-input multiple-output (MIMO) system with N transmit and $M > N$ receive antennas [9].

V. Comparison of Greedy ZF-DP and ZFS

The throughputs of the gZF-DP and ZFS algorithms are presented in Figs. 3 and 4. The y-axis shows sum capacity and sum rate in bits per channel use. The x-axis shows total power P in decibels. The noise level of every user is 1. The sum capacity and sum rates are averaged over 100 channels. Channels are complex-valued, drawn from an i.i.d. Rayleigh distribution with unit-variance for each channel entry. The sum capacity is obtained using the approach proposed in [14].

For the gZF-DP algorithm, analysis (obtained under a long-term power constraint) yields throughput very close to that obtained via simulations (under a short-term power constraint). This can be explained as follows. Capitalizing on multiuser diversity, gZF-DP selects and orders channels (users) from a large pool of statistically independent candidates. The result is that the ensuing d_i's are far more stable than they would have been

$$f_{d_n}(x_n) = \frac{M!}{(M-n)!} \int_{x_1=x_n}^{\infty} \int_{x_2=x_n}^{x_1} \cdots \int_{x_{n-1}=x_n}^{x_{n-2}} \left[\int_{y=0}^{x_n} \phi_n(y, x_{n-1},\ldots,x_1)dy\right]^{M-n} \cdot \prod_{k=1}^{n} \phi_k(x_k,\ldots,x_1)dx_{n-1}\ldots dx_1. \quad (46)$$

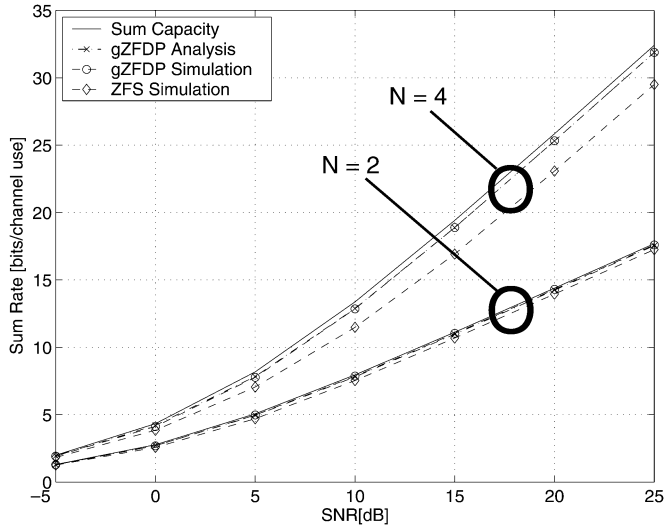


Fig. 3. ZFS versus Greedy ZF-DP versus Sum capacity: $M = 8$ users, $N = 2$, and $N = 4$.

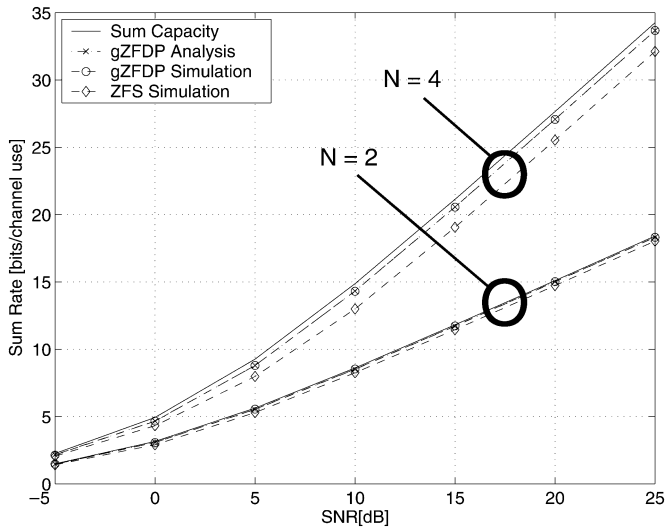


Fig. 4. ZFS versus Greedy ZF-DP versus Sum Capacity: $M = 16$ users, $N = 2$, and $N = 4$.

without user selection and ordering. This justifies the use of a long-term power constraint for analysis, as opposed to the short-term power constraint originally proposed in the algorithm and used in simulations.

In these scenarios ($N = 2$ or 4 and $M = 8$ or 16), both gZF-DP and ZFS algorithms achieve throughput close to sum capacity. Note that ZFS exhibits the same slope of rate increase per decibel of SNR as the gZF-DP algorithm and the sum capacity curve at moderate and high SNR.

Fig. 5 shows the throughput of the ZFS algorithm as a fraction of the throughput of the gZF-DP algorithm for various pairs N, M at 20 dB SNR. The curves are obtained by simulations, averaging over 2×10^4 channels for each pair N, M. For all N, M considered, this fraction stays between 0.875 and 0.985. For a given M, the gap between gZF-DP and ZFS increases as N increases, but even for $N = 8$, the gap is uniformly less than

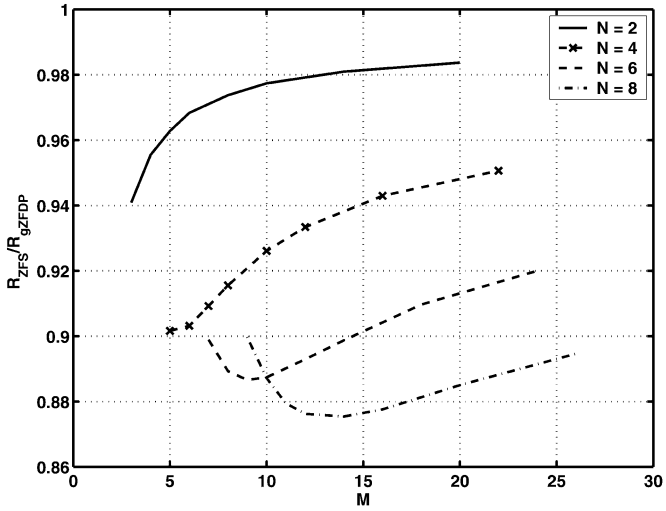


Fig. 5. $R_{ZFS}/R_{gZF\text{-}DP}$ for various numbers of antennas, N, and users, M, at 20 dB SNR.

13% of the gZF-DP throughput. Note that a realistic implementation of DP coding will incur a certain rate loss for the gZF-DP algorithm, so that the gap would be smaller in reality.

Given N and for sufficiently large M, Fig. 5 shows that the gap between ZFS and gZF-DP decreases with M. This is due to multiuser diversity—the more users that contend for transmission, the higher the probability that N of them will be almost orthogonal. This in turn reduces the advantage of DP-coding-based schemes over ZFS. Depending on N, the fraction of sum rate of ZFS over the sum rate of gZF-DP may first exhibit a dip before starting to increase steadily with M. While the dip is small (less than 3%), it is noticeable, and we do not have an explanation for it. We have observed that, as SNR increases, more transmit antennas are required for this dip to occur.

VI. Conclusions

We have considered two algorithms that capitalize on multiuser diversity to achieve a significant fraction of the multiantenna downlink sum capacity when the number of users M is greater than the number of antennas N. We have analyzed the throughput performance of the greedy ZF-DP algorithm in independent Rayleigh fading and characterized the pdfs of certain key parameters of interest. Determining the proper number of samples required for accurate Monte Carlo estimates is a difficult issue without a baseline. While the end result of gZF-DP performance analysis requires sequential numerical integration and is admittedly cumbersome, it provides such a baseline and thus corroborates the results of Monte Carlo estimation. In addition, numerical integration is simpler than Monte Carlo simulation for a small number of transmit antennas. Furthermore, our analysis allowed us to establish that at high SNR, the throughput versus SNR slope of the gZF-DP algorithm is proportional to N.

We have also proposed another low-complexity algorithm, dubbed ZFS, which does not require DP coding at the transmitter. We have shown that the selection procedures of gZF-DP and ZFS algorithms have the same complexity order $O(N^3M)$, which is significantly smaller than the complexity of the optimal algorithms when $M \gg N$. We have evaluated the throughput

performance of the ZFS algorithm via simulations. The results show that for a realistic number of transmit antennas, ZFS achieves a significant fraction of the throughput of gZF-DP and sum capacity at a low coding and online computation cost. The simulation results also indicate that at high SNR, ZFS achieves the same slope of throughput per decibel of SNR as the capacity-achieving strategy based on the use of DP coding for known interference cancellation and convex optimization.

Due to its simplicity, low complexity, and close to optimal performance, the proposed ZFS method offers an attractive alternative to earlier DP-based methods when $M \gg N$.

APPENDIX A
DERIVATION OF THE PDF OF d_n

Note that there are three basic steps in deriving $f_{d_n}(x_n)$:

1) *Truncation of the parent pdf after selecting user s_{n-1}*: Find the conditional pdf of $r_{n-1,u}$ of the remaining users $(u \in U \setminus S_{n-1})$ given realizations of $d_{n-1} \leq \cdots \leq d_1$. From order statistics [3], we obtain (48), shown at the bottom of the page.
2) *Mapping of $f_{r_{n-1}|d_{n-1},\ldots,d_1}(x_{n-1}|y_{n-1},\ldots,y_1)$ into $f_{r_n|d_{n-1},\ldots,d_1}(x_n|y_{n-1},\ldots,y_1)$*: Given realizations of d_i for $i = 1,\ldots,n-1$, where $n \leq N$, there are $n-1$ quadratic-form equations

$$d_i = \mathbf{h}_{s_i}\mathbf{P}_i^{\perp}\mathbf{h}_{s_i}^*.$$

Let the eigenvalue decomposition of $\mathbf{P}_i^{\perp}$ be

$$\mathbf{P}_i^{\perp} = \mathbf{U}_i\mathbf{\Theta}_i\mathbf{U}_i^*.$$

From (13), it follows that there are $N-i+1$ eigenvalues equal to 1 and $i-1$ eigenvalues equal to zero. Then, we can write

$$d_i = \sum_{j=1}^{N-i+1} |\,(\mathbf{h}_{s_i}\mathbf{U}_i)_j\,|^2.$$

As per Assumption 1, we neglect the (mild) dependence of the projector matrices $\mathbf{P}_i^{\perp}$ on the d_i's for $i = 1,\ldots,n-1$. This yields

$$f_{r_n|r_{n-1},d_{n-1},\ldots,d_1}(x_n|v,y_{n-1},\ldots,y_1) = f_{r_n|r_{n-1}}(x_n|v). \tag{49}$$

Since the projection $\mathbf{h}_u\mathbf{P}_n^{\perp}$ is a vector in an $N-n+1$-dimensional subspace, it follows from Claim 1 that

$$f_{r_n|r_{n-1}}(x_n|v) = \begin{cases} \frac{N-n+1}{v}\left(\frac{x_n}{v}\right)^{N-n}, & \text{for } x_n \in [0,v] \\ 0, & \text{otherwise.} \end{cases} \tag{50}$$

Then, the pdf of the parent distribution of $r_{n,u}$ of the remaining users given $d_{n-1} \leq \cdots \leq d_1$ is

$$f_{r_n|d_{n-1},\ldots,d_1}(x_n|y_{n-1},\ldots,y_1) = \int_0^{\infty} f_{r_n|r_{n-1}}(x_n|v) \cdot f_{r_{n-1}|d_{n-1},\ldots,d_1}(v|y_{n-1},\ldots,y_1)dv \tag{51}$$

where $x_n \leq v \leq y_{n-1} \leq \cdots \leq y_1$.

3) d_n conditioned on $d_{n-1},\ldots,d_1$ is the maximum of $M-n+1$ r.v.s with pdf given in (51). Using order statistics [3], we obtain

$$\begin{aligned} &f_{d_n|d_{n-1},\ldots,d_1}(x_n|x_{n-1},\ldots,x_1) \\ &\quad = (M-n+1) \\ &\qquad \cdot \left[F_{r_n|d_{n-1},\ldots,d_1}(x_n|x_{n-1},\ldots,x_1)\right]^{M-n} \\ &\qquad \cdot f_{r_n|d_{n-1},\ldots,d_1}(x_n|x_{n-1},\ldots,x_1). \end{aligned} \tag{52}$$

APPENDIX B
PROOFS

Proof of Proposition 1: Let us first prove the following:

$$f_{r_n|d_{n-1},\ldots,d_1}(x_n|x_{n-1},\ldots,x_1) = \frac{\phi_n(x_n,x_{n-1},\ldots,x_1)}{\prod_{j=1}^{n-1} F_{r_j|d_{j-1},\ldots,d_1}(x_j|x_{j-1},\ldots,x_1)} \tag{53}$$

where $\phi_n(x_n,x_{n-1},\ldots,x_1)$ is given in (45).

This is proven by induction. For $n=2$, we have

$$f_{r_2|d_1}(x_2|x_1) = \int_{y=x_2}^{x_1} f_{r_2|r_1}(x_2|y) f_{r_1|d_1}(y|x_1)dy.$$

From (33), (36), and (40), we obtain

$$f_{r_2|d_1}(x_2|x_1) = \frac{1}{F_{r_1}(x_1)}\frac{1}{\Gamma(N-1)}x_2^{N-2}\int_{v_1=x_2}^{x_1}\exp(-v_1)\,dv_1.$$

From (45), it follows that

$$f_{r_2|d_1}(x_2|x_1) = \frac{\phi_2(x_2,x_1)}{F_{r_1}(x_1)}.$$

Induction hypothesis: (53).

Induction Step:

$$f_{r_{n+1}|d_n,\ldots,d_1}(x_{n+1}|x_n,\ldots,x_1) = \int_{v_n=x_{n+1}}^{x_n} f_{r_{n+1}|r_n}(x_{n+1}|v_n) f_{r_n|d_n,\ldots,d_1}(v_n|x_n,\ldots x_1)dv_n.$$

$$f_{r_{n-1}|d_{n-1},\ldots,d_1}(x_{n-1}|y_{n-1},\ldots,y_1) = \begin{cases} \frac{f_{r_{n-1}|d_{n-2},\ldots,d_1}(x_{n-1}|y_{n-2},\ldots,y_1)}{F_{r_{n-1}|d_{n-2},\ldots,d_1}(y_{n-1}|y_{n-2},\ldots,y_1)}, & x_{n-1} \leq y_{n-1} \leq \cdots \leq y_1 \\ 0, & \text{otherwise.} \end{cases} \tag{48}$$

From (48) and (50), we obtain

$$\begin{aligned} &f_{r_{n+1}|d_n,\ldots,d_1}(x_{n+1}|x_n,\ldots,x_1) \\ &\quad = F_{r_n|d_{n-1},\ldots,d_1}(x_n|x_{n-1},\ldots,x_1)^{-1} \\ &\qquad \cdot \int_{v_n=x_{n+1}}^{x_n} \frac{N-n}{v_n}\left(\frac{x_{n+1}}{v_n}\right)^{N-n-1} \\ &\qquad \cdot f_{r_n|d_{n-1},\ldots,d_1}(v_n|x_{n-1},\ldots,x_1)dv_n. \end{aligned}$$

By the induction hypothesis, we have

$$\begin{aligned} &f_{r_{n+1}|d_n,\ldots,d_1}(x_{n+1}|x_n,\ldots,x_1) \\ &\quad = \left[F_{r_n|d_{n-1},\ldots,d_1}(x_n|x_{n-1},\ldots,x_1)\right]^{-1} \\ &\qquad \cdot \left[\prod_{j=1}^{n-1} F_{r_j|d_{j-1},\ldots,d_1}(x_j|x_{j-1},\ldots,x_1)\right]^{-1} \\ &\qquad \cdot \int_{v_n=x_{n+1}}^{x_n} \frac{N-n}{v_n}\left(\frac{x_{n+1}}{v_n}\right)^{N-n-1} \\ &\qquad \cdot \phi_n(v_n,x_{n-1},\ldots x_1)dv_n. \end{aligned}$$

From (45), it follows that

$$\begin{aligned} &f_{r_{n+1}|d_n,\ldots,d_1}(x_{n+1}|x_n,\ldots,x_1) \\ &\quad = \left[\prod_{j=1}^{n} F_{r_j|d_{j-1},\ldots,d_1}(x_j|x_{j-1},\ldots,x_1)\right]^{-1} \\ &\qquad \cdot \frac{N-n}{\Gamma(N-n+1)}x_{n+1}^{N-n-1} \\ &\qquad \cdot \int_{v_n=x_{n+1}}^{x_n}\int_{v_{n-1}=v_n}^{x_{n-1}}\cdots\int_{v_1=v_2}^{x_1}\exp(-v_1) \\ &\qquad \cdot dv_1\ldots dv_{n-1}dv_n. \end{aligned}$$

Applying (45) again, we have

$$\begin{aligned} &f_{r_{n+1}|d_n,\ldots,d_1}(x_{n+1}|x_n,\ldots,x_1) \\ &\qquad = \frac{\phi_{n+1}(x_{n+1},x_n,\ldots,x_1)}{\prod_{j=1}^{n} F_{r_j|d_{j-1},\ldots,d_1}(x_j|x_{j-1},\ldots,x_1)}. \end{aligned}$$

Now, we use the above result to prove Proposition 1. For $n=1$, from (44), we obtain

$$f_{d_1}(x_1) = M\left[\int_{y=0}^{x_1}\phi_1(y)dy\right]^{M-1}\phi_1(x_1).$$

For $1 < n \le N$ and substituting (52) into (43), we obtain the equation shown at the bottom of the page. Applying (53), we obtain the equation at the top of the next page. Dividing the left fraction and rearranging the right one, we obtain

$$\begin{aligned} &G_{n-1}(x_{n-1},\ldots,x_1) \\ &\quad = \frac{\prod_{k=2}^{n-1}\left[F_{r_k|d_{k-1},\ldots,d_1}(x_k|x_{k-1},\ldots,x_1)\right]^{n-1-k}}{\prod_{j=1}^{n-2}\left[F_{r_j|d_{j-1},\ldots,d_1}(x_j|x_{j-1},\ldots,x_1)\right]^{n-1-j}} \\ &\qquad \cdot \frac{1}{[F_{r_1}(x_1)]^{M-n+1}} \\ &G_{n-1}(x_{n-1},\ldots,x_1) \\ &\quad = \frac{1}{[F_{r_1}(x_1)]^{M-1}}. \end{aligned}$$

Therefore

$$\begin{aligned} &f_{d_n}(x) \\ &= \frac{M!}{(M-n)!} \\ &\cdot \int_{x_1=x}^{\infty}\int_{x_2=x}^{x_1}\int_{x_{n-1}=x}^{x_{n-2}}\left[\int_{y=0}^{x_n}\phi_n(y,x_{n-1},\ldots,x_1)dy\right]^{M-n} \\ &\cdot \prod_{k=1}^{n}\phi_k(x_k,\ldots,x_1)dx_{n-1}\ldots dx_1. \end{aligned}$$

■

Proof of Theorem:

$$\frac{\partial}{\partial\rho}R_{gZF\text{-}DP} = \sum_{i=1}^{N}\frac{\partial}{\partial\rho}R_i, \quad \text{where } \frac{\partial}{\partial\rho}R_i = \frac{\partial\mu}{\partial\rho}\frac{\partial}{\partial\mu}R_i.$$

$\rho = 10\log_{10}(P)$ so that from (32), we have

$$\begin{aligned} \frac{\partial\mu}{\partial\rho} &= \frac{\partial\mu}{\partial P}\frac{\partial P}{\partial\rho} \\ &= \frac{1}{N-\sum_{i=1}^{N}F_{d_i}\left(\frac{1}{\mu}\right)}\frac{\ln 10}{10}P. \end{aligned}$$

Using the Leibnitz rule, from (31), we have

$$\frac{\partial}{\partial\mu}R_i = \frac{1}{\mu\ln 2}\left(1-F_{d_i}\left(\frac{1}{\mu}\right)\right).$$

It follows that

$$\frac{\partial}{\partial\rho}R_{gZF\text{-}DP} = \frac{\log_2 10}{10}\frac{P}{\mu}.$$

In order to determine $\lim_{\rho\to\infty}(\partial/\partial\rho)R_{gZF\text{-}DP}$, we will determine $\lim_{\rho\to\infty}(P/\mu)$. Note that $\rho\to\infty$ is equivalent to $P\to\infty$. In addition, $(\partial P/\partial\mu) = N-\sum_{i=1}^{N}F_{d_i}(1/\mu) > 0$ for

$$\begin{aligned} f_{d_n}(x_n) = \frac{M!}{(M-n)!}\int_{x_1=x_n}^{\infty}\int_{x_2=x_n}^{x_1}\cdots\int_{x_{n-1}=x_n}^{x_{n-2}}&\left[F_{r_n|d_{n-1},\ldots,d_1}(x_n|x_{n-1},\ldots,x_1)\right]^{M-n} \\ &\cdot f_{r_n|d_{n-1},\ldots,d_1}(x_n|x_{n-1},\ldots,x_1)\prod_{k=2}^{n-1}\left[F_{r_k|d_{k-1},\ldots,d_1}(x_k|x_{k-1},\ldots,x_1)\right]^{M-k} \\ &\cdot f_{r_k|d_{k-1},\ldots,d_1}(x_k|x_{k-1},\ldots,x_1)\left[F_{r_1}(x_1)\right]^{M-1}f_{r_1}(x_1)dx_{n-1}\ldots dx_1. \end{aligned}$$

$$f_{d_n}(x_n) = \frac{M!}{(M-n)!} \int_{x_1=x_n}^{\infty} \int_{x_2=x_n}^{x_1} \cdots \int_{x_{n-1}=x_n}^{x_{n-2}} \left[\frac{\int_{y=0}^{x_n} \phi_n(y, x_{n-1}, \ldots, x_1) dy}{\prod_{j=1}^{n-1} F_{r_j|d_{j-1},\ldots,d_1}(x_j|x_{j-1}, \ldots, x_1)} \right]^{M-n}$$
$$\cdot \frac{\phi_n(x_n, \ldots, x_1)}{\prod_{j=1}^{n-1} F_{r_j|d_{j-1},\ldots,d_1}(x_j|x_{j-1}, \ldots, x_1)}$$
$$\cdot \prod_{k=2}^{n-1} \left\{ \left[F_{r_k|d_{k-1},\ldots,d_1}(x_k|x_{k-1}, \ldots, x_1)\right]^{M-k} \frac{\phi_k(x_k, \ldots, x_1)}{\prod_{j=1}^{k-1} F_{r_j|d_{j-1},\ldots,d_1}(x_j|x_{j-1}, \ldots, x_1)} \right\}$$
$$\cdot \left[F_{r_1}(x_1)\right]^{M-1} \phi_1(x_1) dx_{n-1} \ldots dx_1$$
$$f_{d_n}(x_n) = \frac{M!}{(M-n)!}$$
$$\cdot \int_{x_1=x_n}^{\infty} \int_{x_2=x_n}^{x_1} \cdots \int_{x_{n-1}=x_n}^{x_{n-2}} \left[\int_{y=0}^{x_n} \phi_n(y, x_{n-1}, \ldots, x_1) dy \right]^{M-n}$$
$$\cdot \prod_{k=1}^{n} \phi_k(x_k, \ldots, x_1) G_{n-1}(x_{n-1}, \ldots, x_1) \left[F_{r_1}(x_1)\right]^{M-1} dx_{n-1} \ldots dx_1$$

where

$$G_{n-1}(x_{n-1}, \ldots, x_1) = \frac{\prod_{k=2}^{n-1} \left[F_{r_k|d_{k-1},\ldots,d_1}(x_k|x_{k-1}, \ldots, x_1)\right]^{M-k}}{\prod_{j=1}^{n-1} \left[F_{r_j|d_{j-1},\ldots,d_1}(x_j|x_{j-1}, \ldots, x_1)\right]^{M-n+1}} \frac{1}{\prod_{k=2}^{n-1} \prod_{j=1}^{k-1} F_{r_j|d_{j-1},\ldots,d_1}(x_j|x_{j-1}, \ldots, x_1)}.$$

$\mu > 0$ so that $P \to \infty$ is equivalent to $\mu \to \infty$. We will prove that $\lim_{\mu\to\infty} (P/\mu) = N$. From (32), we have

$$\frac{P}{\mu} = \sum_{i=1}^{N} \int_{1/\mu}^{\infty} f_{d_i}(z) dz - \frac{1}{\mu} \sum_{i=1}^{N} \int_{1/\mu}^{\infty} \frac{1}{z} f_{d_i}(z) dz.$$

Then

$$\lim_{\mu\to\infty} \frac{P}{\mu} = N - \lim_{\mu\to\infty} \frac{1}{\mu} g_N(\mu)$$

where

$$g_N(\mu) = \sum_{i=1}^{N} \int_{1/\mu}^{\infty} \frac{1}{z} f_{d_i}(z) dz.$$

Note that if we demonstrate that

$$\lim_{\mu\to\infty} \frac{\partial}{\partial\mu} g_N(\mu) = 0$$

the desired result will follow because

$$\lim_{\mu\to\infty} \frac{\partial}{\partial\mu} g_N(\mu) = 0 \Rightarrow \lim_{\mu\to\infty} g_N(\mu) = O(1)$$
$$\Rightarrow \lim_{\mu\to\infty} \frac{1}{\mu} g_N(\mu) = 0$$
$$\Rightarrow \lim_{\mu\to\infty} \frac{P}{\mu} = N.$$

It is easy to check that $(\partial/\partial\mu) g_N(\mu) = \sum_{i=1}^{N} (1/\mu) f_{d_i}(1/\mu)$ so that it suffices to prove that $\lim_{\mu\to\infty} f_{d_i}(1/\mu) = 0$ or, equivalently, $\lim_{x\to 0} f_{d_i}(x) = 0$, for $i = 1, 2, \ldots, N$, where $N \geq 2$, and $N < M$.

From (33) and (44), it follows that $\phi_1(0) = 0$. Then, from (35), it follows that $f_{d_1}(0) = 0$.

In order to prove that $f_{d_n}(0) = 0$ for $1 < n \leq N$, we will prove that $\phi_n(0, x_{n-1}, \ldots, x_1)$ is bounded for any $0 \leq x_{n-1} \leq \cdots \leq x_1$. In order to prove that $\phi_n(0, x_{n-1}, \ldots, x_1)$ is bounded, consider the multiple integral [cf. (45)]

$$I_n = \int_{v_{n-1}=0}^{x_{n-1}} \int_{v_{n-2}=v_{n-1}}^{x_{n-2}} \cdots \int_{v_1=v_2}^{x_1} \exp(-v_1) dv_1 \ldots dv_{n-1}.$$

Integrating over dv_1, we obtain

$$I_n = \int_{v_{n-1}=0}^{x_{n-1}} \int_{v_{n-2}=v_{n-1}}^{x_{n-2}} \cdots \int_{v_2=v_3}^{x_2} \exp(-v_2) dv_2 \ldots dv_{n-1}$$
$$- \exp(-x_1) \int_{v_{n-1}=0}^{x_{n-1}} \int_{v_{n-2}=v_{n-1}}^{x_{n-2}} \cdots \int_{v_2=v_3}^{x_2} dv_2 \ldots dv_{n-1}$$
$$I_n = I_{n,1} - \exp(-x_1) B_{n,2}(x_{n-1}, \ldots, x_2).$$

Observe that the first multiple integral on the right-hand side (RHS), which is denoted $I_{n,1}$, has the same form as I_n. Due to $x_1 \geq x_2 \geq \cdots \geq x_{n-1} \geq 0$, we have

$$B_{n,1}(x_{n-1}, \ldots, x_2) \leq \int_{v_{n-1}=0}^{x_1} \int_{v_{n-2}=0}^{x_1} \cdots \int_{v_2=0}^{x_1} dv_2 \ldots dv_{n-1}.$$

Therefore

$$B_{n,1}(x_{n-1}, \ldots, x_2) \leq x_1^{n-2}.$$

Note that $\exp(-x_1)x_1^{n-2}$ is bounded for all $x_1 \geq 0$ so that $\exp(-x_1)B_{n,1}(x_{n-1},\ldots,x_2)$ is also bounded. Integrating over all dummy variables, we obtain

$$I_n = \exp(-x_{n-1}) - 1 - \sum_{j=1}^{n-2} \exp(-x_j) B_{n,j}(x_{n-j},\ldots,x_2)$$

where

$$B_{n,j}(x_{n-j},\ldots,x_{j+1}) = \int_{v_{n-1}=0}^{x_{n-1}} \cdots \int_{v_{j+1}=v_{j+2}}^{x_{j+1}} dv_{j+1}\ldots dv_{n-1}.$$

It can be shown that $\exp(-x_j)\ B_{n,j}(x_{n-1},\ldots,x_2)$ is bounded by the same argument as for $\exp(-x_1)\ B_{n,1}(x_{n-1},\ldots,x_2)$. Therefore, $\phi_n(0,x_{n-1},\ldots,x_1)$ is bounded for all $x_1 \geq x_2 \geq \cdots \geq x_{n-1} \geq 0$.

Then, from (45), it follows that

$$\phi_n(0,x_{n-1},\ldots,x_1) = \begin{cases} \frac{\int_{v_{n-1}=0}^{x_{n-1}} \cdots \int_{v_1=v_2}^{x_1} \exp(-v_1) dv_1 \ldots dv_{n-1}}{\Gamma(N-n+1)}, & n = N \\ 0, & n < N. \end{cases}$$

If $n < N$, then from (43), it follows that $f_{d_n}(0) = 0$. If $n = N$, then applying the mean-value theorem, we obtain

$$\lim_{x_n \to 0} \int_{y=0}^{x_n} \phi_n(y,x_{n-1},\ldots,x_1)dy = \lim_{x_n \to 0} \phi_n(0,x_{n-1},\ldots,x_1)x_n.$$

Since $\phi_n(0,x_{n-1},\ldots,x_1)$ is bounded

$$\lim_{x_n \to 0} \int_{y=0}^{x_n} \phi_n(y,x_{n-1},\ldots,x_1)dy = 0.$$

Finally, from (43) and $n = N < M$, it follows that $f_{d_N}(0) = 0$. ■

REFERENCES

[1] G. Caire and S. Shamai (Shitz), "On the achievable throughput of a multi-antenna Gaussian broadcast channel," *IEEE Trans. Inf. Theory*, vol. 49, no. 7, pp. 1691–1706, Jul. 2003.

[2] M. H. M. Costa, "Writing on dirty paper," *IEEE Trans. Inf. Theory*, vol. IT-29, no. 3, pp. 439–441, May 1983.

[3] H. A. David, *Order Statistics*, 2nd ed. New York: Wiley, 1981.

[4] G. D. Forney Jr, "Trellis shaping," *IEEE Trans. Inf. Theory*, vol. 38, no. 2, pp. 281–300, Mar. 1992.

[5] D. Gore, R. W. Heath Jr, and A. Paulraj, "Transmit selection in spatial multiplexing systems," *IEEE Commun. Lett.*, vol. 6, no. 11, pp. 491–493, Nov. 2002.

[6] R. A. Horn and C. R. Johnson, *Matrix Analysis*. New York: Cambridge Univ. Press, 1985.

[7] N. Jindal, W. Rhee, S. Vishwanath, S. Jafar, and A. Goldsmith, "Sum power iterative waterfilling for Gaussian vector broadcast channels," *IEEE Trans. Inf. Theory*, submitted for publication.

[8] C. B. Peel, "On dirty paper coding," *Signal Process. Mag.*, vol. 20, no. 3, pp. 112–113, May 2003.

[9] W. Rhee and J. M. Cioffi, "On the capacity of multiuser wireless channels with multiple antennas," *IEEE Trans. Inf. Theory.*, vol. 49, no. 10, pp. 2580–2595, Oct. 2003.

[10] D. Samardzija and N. Mandayam, "Multiple antenna transmitter optimization schemes for multiuser systems," in *Proc. IEEE Veh. Technol. Conf.*, Orlando, FL, 2003, pp. 399–403.

[11] Q. Spencer and M. Haardt, "Capacity and downlink transmission algorithms for a multi-user MIMO channel," in *Proc. 36th Asilomar Conf. Signals, Syst., Comput.*, Pacific Grove, CA, Nov. 2002.

[12] I. E. Telatar, "Capacity of multi-antenna Gaussian channels," *Eur. Trans. Telecomm.*, vol. 10, no. 6, pp. 585–596, Nov.–Dec. 1999.

[13] Z. Tu and R. S. Blum, "Multiuser diversity for a dirty paper approach," *IEEE Commun. Lett.*, vol. 7, no. 8, pp. 370–372, Aug. 2003.

[14] S. Vishwanath, N. Jindal, and A. Goldsmith, "Duality, achievable rates and sum-rate capacity of Gaussian MIMO broadcast channels," *IEEE Trans. Inf. Theory*, vol. 49, no. 10, pp. 2658–2668, Oct. 2003.

[15] P. Viswanath and D. Tse, "Sum capacity of the vector Gaussian broadcast channel and uplink-downlink duality," *IEEE Trans. Inf. Theory*, vol. 49, no. 8, pp. 1912–1921, Aug. 2003.

[16] H. Viswanathan, S. Venkatesan, and H. Huang, "Downlink capacity evaluation of cellular networks with known interference cancellation," *IEEE J. Sel. Areas Commun.*, vol. 21, no. 5, pp. 802–811, Jun. 2003.

[17] J. H. Winters, J. Salz, and R. D. Gitlin, "The impact of antenna diversity on the capacity of wireless communication systems," *IEEE Trans. Commun.*, vol. 42, no. 2/3/4, pp. 1740–1751, Feb./Mar./Apr. 1994.

[18] W. Yu and J. M. Cioffi, "Trellis precoding for the broadcast channel," in *Proc. Globecom*, San Antonio, TX, Nov. 2001, pp. 1344–1348.

[19] ——, "Sum capacity of a Gaussian broadcast channel," in *Proc. IEEE Int. Symp. Inf. Theory*, Lausanne, Switzerland, Jul. 2002.

[20] R. Zamir, S. Shamai (Shitz), and U. Erez, "Nested linear/lattice codes for structured multiterminal binning," *IEEE Trans. Inf. Theory*, vol. 48, no. 6, pp. 1250–1276, Jun. 2002.

Goran Dimić (M'05) received the Diploma in electrical engineering from the University of Belgrade, Belgrade, Serbia and Montenegro, and the M.S. and Ph.D. degrees in electrical engineering from the University of Minnesota, Minneapolis, in 1999, 2001, and 2004, respectively.

His research interests are in the area of signal processing for communications and networking.

Nicholas D. Sidiropoulos (M'92–SM'99) received the Diploma in electrical engineering from the Aristotelian University of Thessaloniki, Thessaloniki, Greece, and the M.S. and Ph.D. degrees in electrical engineering from the University of Maryland, College Park (UMCP), in 1988, 1990 and 1992, respectively.

From 1988 to 1992, he was a Fulbright Fellow and a Research Assistant at the Institute for Systems Research (ISR), UMCP. From September 1992 to June 1994, he served his military service as a Lecturer in the Hellenic Air Force Academy. From October 1993 to June 1994, he also was a member of the technical staff, Systems Integration Division, G-Systems Ltd., Athens, Greece. He was a Postdoctoral Fellow (1994 to 1995) and Research Scientist (1996 to 1997) at ISR-UMCP, an Assistant Professor with the Department of Electrical Engineering, University of Virginia, Charlottesville, from 1997 to 1999, and an Associate Professor with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, from 2000 to 2002. He is currently a Professor with the Telecommunications Division of the Department of Electronic and Computer Engineering, Technical University of Crete, Chania, Crete, Greece, and Adjunct Professor at the University of Minnesota. His current research interests are primarily in signal processing for communications, and multi-way analysis. He is an active consultant for industry in the areas of frequency hopping systems and signal processing for xDSL modems.

Dr. Sidiropoulos is a member of both the Signal Processing for Communications (SPCOM) and Sensor Array and Multichannel Signal Processing (SAM) Technical Committees of the IEEE Signal Processing Society and currently serves as an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING. From 2000 to 2002, he also served as Associate Editor for the IEEE SIGNAL PROCESSING LETTERS. He received the NSF/CAREER award (Signal Processing Systems Program) in June 1998 and an IEEE Signal Processing Society Best Paper Award in 2001.

TIME-FREQUENCY ANALYSIS USING PARTICLE FILTERING: CLOSED-FORM OPTIMAL IMPORTANCE FUNCTION AND SAMPLING PROCEDURE FOR A SINGLE TIME-VARYING HARMONIC

E. E. Tsakonas
Dept. ECE, Tech. Univ. of Crete
73100 Chania - Crete, Greece
etsakwnas@gmail.com

N. D. Sidiropoulos[*]
Dept. ECE, Tech. Univ. of Crete
73100 Chania - Crete, Greece
nikos@telecom.tuc.gr

A. Swami
Army Research Laboratory
Adelphi, MD, 20783, U.S.A.
a.swami@ieee.org

ABSTRACT

We consider the problem of tracking the frequency and complex amplitude of a time-varying (TV) harmonic signal using particle filtering (PF) tools. Similar to previous PF approaches to TV spectral analysis, we assume that the frequency and complex amplitude evolve according to a Gaussian AR(1) model; but we concentrate on the important special case of a single TV harmonic. For this case, we show that the optimal importance function (that minimizes the variance of the particle weights) can be computed in closed form. We also develop a suitable procedure to sample from the optimal importance function. The end result is a custom PF solution that is more efficient than generic ones, and can be used in a broad range of important applications that postulate a single TV harmonic component, e.g., TV Doppler estimation in communications and radar.

1. INTRODUCTION

Spectral analysis and time-frequency analysis are core tools in signal processing research (e.g., [10, 3]). Time-varying (TV) spectra arise in a broad range of important applications: from speech, to radar, to wireless channel modeling and estimation.

TV spectral analysis tools range from basic non-parametric approaches such as the spectrogram, to the Wigner-Ville and other time-frequency distributions, and on to parametric ones such as polynomial basis expansion models, and TV line spectra mixture models.

Line spectra mixtures (whether stationary or TV) entail a nonlinear observation equation, which complicates parameter estimation. When the evolution of model parameters can be captured in state-space form, particle filtering (PF) tools become particularly appealing for tracking the model parameters. For a multicomponent TV harmonic mixture model, PF approaches have been pursued in [1, 7]. In [1], the evolution of harmonic parameters (frequencies, complex amplitudes, possibly also decay rates) is modeled using a Gaussian auto-regressive (AR) process, and an improved auxiliary particle filtering algorithm is applied to track the parameters. In [7], a similar Gaussian random walk model is used for the evolution of the parameters. Unlike [1], temporal slices of the spectrogram are used in the measurement equation of [7] (which apparently limits the attainable time-frequency resolution), and an unscented PF algorithm is adapted to track the model parameters.

Gaussian AR models of the evolution of harmonic mixture parameters are plausible and convenient in many situations - e.g., they can capture smoothness due to inertia or other physical constraints. Following [1, 7], we also assume that the frequency and complex amplitude evolve according to a Gaussian AR(1) model; but we concentrate on the important special case of a single TV harmonic. For this case, we show that the optimal importance function (that minimizes the variance of the particle weights) can be computed in closed form. We also develop a suitable procedure to sample from the optimal importance function. The end result is a custom PF solution that is more efficient than generic ones, and can be used in a broad range of important applications that postulate a single TV harmonic component, e.g., TV Doppler estimation in communications and radar.

[*]Corresponding author. Supported in part by the Army Research Laboratory (ARL) through participation in the ARL Collaborative Technology Alliance (ARL-CTA) for Communications and Networks under Cooperative Agreement DADD19-01-2-0011, and in part by ARO under ERO Contract N62558-03-C-0012.

2. DATA MODEL

Let $\mathbf{x}_k := [\omega_k, A_k]^T$ denote the state at time k, where $\omega_k \in \Re$ and $A_k \in \mathbb{C}$ denote instantaneous frequency and complex amplitude. The state evolves according to the following AR(1) model:

$$\mathbf{x}_k = \mathbf{H}\mathbf{x}_{k-1} + [u_{k-1} \;\; w_{k-1}]^T$$

where $\mathbf{H}$ is 2×2 diagonal, $\mathbf{H} = diag\left([b_1, b_2]^T\right)$, with b_ℓ equal to $1 - \epsilon_\ell$ (e.g., 0.999). The process noise sequence is i.i.d. The process noise vector at time k consists of two independent random variables with the following marginal statistics:

$$[u_{k-1} \;\; w_{k-1}]^T \sim \left[\mathcal{N}\left(0, \sigma_\omega^2\right), \mathcal{CN}(0, 2\sigma_A^2)\right]^T,$$

where $\mathcal{N}, \mathcal{CN}$ stand for the (real) normal and circularly symmetric complex normal distribution, respectively. The measurements are related to the state via the measurement equation

$$y_k = \mathbf{x}_k(2)e^{j\mathbf{x}_k(1)k} + v_k,$$

where v_k denotes i.i.d. $\mathcal{CN}(0, 2\sigma_n^2)$ measurement noise .

Given a sequence of observations $\{y_k\}_{k=1}^T$, the problem of interest is to estimate the sequence of posterior densities, that is $p\left(\mathbf{x}_k | \{y_l\}_{l=1}^k\right)$, $k \in \{1, \cdots, T\}$. Given $p\left(\mathbf{x}_k | \{y_l\}_{l=1}^k\right)$, one can estimate $\mathbf{x}_k$ via the associated (posterior) mean, or mode.

3. PARTICLE FILTERING

Particle filtering has emerged as an important sequential state estimation method for stochastic non-linear and/or non-Gaussian state-space models, for which it provides a powerful alternative to the commonly used extended Kalman filter. See [2, 5, 6] for recent tutorial overviews.

In particle filtering, continuous distributions are approximated by discrete random measures, comprising "particles" and associated weights. That is, a certain continuous distribution of interest, say $p(\mathbf{x})$, is approximated as

$$p(\mathbf{x}) \approx \sum_{n=1}^{N} w_n \delta(\mathbf{x} - \mathbf{x}_n),$$

where $\delta(\cdot)$ denotes the Dirac delta functional. A useful simplification stemming from this approximation is that the computation of pertinent expectations and conditional probabilities reduces to summation, as opposed to integration. While this can also be accomplished via direct discretization over a fixed grid, the use of a random measure affords flexibility in adapting the particle locations to better fit the distribution of interest.

Different types of particle filters may be applied to a given state-space model. The various particle filters primarily differ in the choice of so-called *importance* (or, *proposal*) function. Different importance functions yield different estimation performance - complexity trade-offs. From the viewpoint of minimizing the variance of the weights, the optimal importance function is given by [2, 5]

$$p(\mathbf{x}_k|\mathbf{x}_{n,k-1}, y_k) = \frac{p(y_k|\mathbf{x}_k)p(\mathbf{x}_k|\mathbf{x}_{n,k-1})}{\int_{\mathbf{x}} p(y_k|\mathbf{x})p(\mathbf{x}|\mathbf{x}_{n,k-1})d\mathbf{x}},$$

where $\mathbf{x}_{n,k} := [\omega_{n,k}, A_{n,k}]^T$ denotes the n-th particle at time k. The optimal importance function usually strikes a better performance - complexity trade-off than other alternatives. There are, however, two difficulties associated with the use of the optimal importance function. First and foremost, it requires multidimensional integration to compute the normalization factor, which is usually intractable. Second, sampling from the optimal importance function is a rather complicated process. Thankfully, for our particular model, it turns out that it is possible to carry out the integration analytically. This is explained next.

Define a dummy variable $\mathbf{x} := [\omega, A]^T$, and let $D(y_k, \mathbf{x}_{n,k-1}) := \int_{\mathbf{x}} p(y_k|\mathbf{x})p(\mathbf{x}|\mathbf{x}_{n,k-1})d\mathbf{x}$. Then

$$D(y_k, \mathbf{x}_{n,k-1}) = \int_{\omega\in\Re}\int_{A\in\mathbb{C}} \frac{1}{2\pi\sigma_n^2} e^{-\frac{|y_k - Ae^{j\omega k}|^2}{2\sigma_n^2}} \times$$

$$\left[\frac{1}{\sqrt{2\pi}\sigma_\omega} e^{-\frac{(\omega - b_1\omega_{n,k-1})^2}{2\sigma_\omega^2}} \frac{1}{2\pi\sigma_A^2} e^{-\frac{|A - b_2 A_{n,k-1}|^2}{2\sigma_A^2}}\right] dAd\omega$$

Letting $m_A := b_2 A_{n,k-1}$, $m_\omega := b_1\omega_{n,k-1}$, $v := \phi_A - \phi_{y_k}$, it can be shown that

$$D(y_k, \mathbf{x}_{n,k-1}) = \frac{1}{2\pi(\sigma_A^2 + \sigma_n^2)} e^{-\frac{|y_k|^2 + |m_A|^2}{2(\sigma_A^2 + \sigma_n^2)}} \times \mathcal{B},$$

with the multiplicative factor $\mathcal{B}$ given by

$$\mathcal{B} = \mathbf{I}_0(-\frac{|m_A||y_k|}{\sigma_A^2 + \sigma_n^2}) +$$

$$+2\sum_{m=1}^{m=+\infty} (-1)^m \mathbf{I}_m(-\frac{|m_A||y_k|}{\sigma_A^2 + \sigma_n^2}) e^{-\frac{(k\sigma_\omega)^2 m^2}{2}} \cos(mkm_\omega - mv)$$

where $\mathbf{I}_m(\cdot)$ denotes the modified Bessel function of the first kind. The sum term above is quite interesting. Due to the negative exponential dependence on the time index k and the properties of Bessel functions, it vanishes quickly with k - only the zero-order Bessel term remains.

We use rejection [4, pp. 40-42] to generate samples from the optimal importance function $p(\mathbf{x}_k|\mathbf{x}_{n,k-1}, y_k) =$

$$= \frac{\frac{1}{2\pi\sigma_n^2} e^{-\frac{|y_k - A_k e^{j\omega_k k}|^2}{2\sigma_n^2}} \frac{1}{\sqrt{2\pi}\sigma_\omega} e^{-\frac{(\omega_k - m_\omega)^2}{2\sigma_\omega}} \frac{1}{2\pi\sigma_A^2} e^{-\frac{|A_k - m_A|^2}{2\sigma_A^2}}}{\frac{1}{2\pi(\sigma_A^2+\sigma_n^2)} e^{-\frac{|y_k|^2 + |m_A|^2}{2(\sigma_A^2+\sigma_n^2)}} \mathcal{B}}.$$

Let $\sigma^2 := \frac{\sigma_A^2\sigma_n^2}{\sigma_A^2+\sigma_n^2}$ and $\mu := \frac{\sigma_A^2|y_k| + \sigma_n^2|m_A|}{\sigma_A^2+\sigma_n^2}$. Using the triangle inequality, it can be shown that a suitable dominating density is

$$g(\mathbf{x}_k|\mathbf{x}_{n,k-1}, y_k) = \frac{e^{-\frac{(\omega_k - m_\omega)^2}{2\sigma_\omega^2}} e^{-\frac{(|A_k| - \mu)^2}{2\sigma^2}}}{(2\pi)^2 Q_o\ \sigma_\omega\sigma},$$

$$c := \frac{\sqrt{2\pi}Q_o}{e^{-\frac{|m_A||y_k|}{\sigma_A^2+\sigma_n^2}} \mathcal{B}\sigma}$$

$$Q_o := \int_{r=0}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(r-\mu)^2}{2\sigma^2}}\ dr = \frac{1}{2}\ \text{erfc}(-\frac{\sigma_A^2|y_k| + \sigma_n^2|m_A|}{\sigma_A\sigma_n\sqrt{2(\sigma_A^2+\sigma_n^2)}})$$

Fig. 1 shows a typical plot of the dominated and dominating densities, illustrating the tightness of the bounding step. The overall algorithm is summarized in Table 1.

4. CRAMER-RAO LOWER BOUND

The Cramér-Rao Lower Bound (CRLB) for our model can be computed using the recursive formula of Tichavsky *et al* [11] for the calculation of the Fisher information matrix, $\mathbf{J}_k$. The state equation in our particular model is linear, Gaussian; this allows considerable simplification of the general result in [11], thus yielding

$$\mathbf{J}_k = \mathbf{D}_{k-1}^{22} - \mathbf{D}_{k-1}^{21}(\mathbf{J}_{k-1} + \mathbf{D}_{k-1}^{11})^{-1}\mathbf{D}_{k-1}^{12}, k \geqslant 0$$

with

$$\mathbf{D}_{k-1}^{11} := -\mathbf{E}\{\nabla_{\mathbf{x}_{k-1}} \left[\nabla_{\mathbf{x}_{k-1}} \log \mathbf{p}(\mathbf{x_k}|\mathbf{x_{k-1}})\right]^T\},$$

$$\mathbf{D}_{k-1}^{12} := \left[\mathbf{D}_{k-1}^{21}\right]^T = -\mathbf{E}\{\nabla_{\mathbf{x}_k} \left[\nabla_{\mathbf{x}_{k-1}} \log \mathbf{p}(\mathbf{x_k}|\mathbf{x_{k-1}})\right]^T\},$$

and

$$\mathbf{D}_{k-1}^{22} := -\mathbf{E}\{\nabla_{\mathbf{x}_k} [\nabla_{\mathbf{x}_k} \log \mathbf{p}(\mathbf{x}_k|\mathbf{x}_{k-1})]^T\} -$$
$$\mathbf{E}\{\nabla_{\mathbf{x}_k} [\nabla_{\mathbf{x}_k} \log \mathbf{p}(\mathbf{y}_k|\mathbf{x}_k)]^T\}.$$

At this point, it is convenient to rewrite our model in real-valued form. Upon defining $\mathbf{x}'_k := [\omega_k, \Re(A_k), \Im(A_k)]^T$, where $\Re(\cdot), \Im(\cdot)$ extract the real, resp. imaginary part, we have

$$\mathbf{x}'_k = \mathbf{H}'\mathbf{x}'_{k-1} + \mathbf{u}_{k-1}$$

$$\mathbf{y}_k = \left[\ \Re\{A_k e^{j\omega_k k}\}\ \ \Im\{A_k e^{j\omega_k k}\}\ \right]^T\ +\ \mathbf{v}_k$$

where $\mathbf{H}' = diag\left(\left[b_1, b_2, b_3\right]^T\right)$, with b_ℓ being $1-\epsilon_\ell$, $\mathbf{u}_{k-1} \sim \mathcal{N}(\mathbf{0}, \mathbf{Q})$ with $\mathbf{Q} = diag\left(\left[\sigma_\omega^2, \sigma_A^2, \sigma_A^2\right]^T\right)$, and $\mathbf{v}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$ with $\mathbf{R} = diag\left(\left[\sigma_n^2, \sigma_n^2\right]^T\right)$. Then

$$\mathbf{D}_{k-1}^{11} = \mathbf{H}'^T \mathbf{Q}^{-1} \mathbf{H}',$$

$$\mathbf{D}_{k-1}^{12} = \left[\mathbf{D}_{k-1}^{21}\right]^T := -\mathbf{H}'^T \mathbf{Q}^{-1},$$

$$\mathbf{D}_{k-1}^{22} = \mathbf{Q}^{-1} + \mathbf{E}\{\tilde{\mathbf{F}_\mathbf{k}}^T \mathbf{R}^{-1} \tilde{\mathbf{F}_\mathbf{k}}\},$$

with $\tilde{\mathbf{F}_\mathbf{k}}$ being the $\mathbf{2} \times \mathbf{3}$ matrix

$$\tilde{\mathbf{F}_\mathbf{k}} = \nabla_{\mathbf{x}_k'} \left[\begin{array}{cc} \Re\{A_k e^{j\omega_k k}\} & \Im\{A_k e^{j\omega_k k}\} \end{array} \right]^T.$$

For $\mathbf{D}_{k-1}^{11}$ and $\mathbf{D}_{k-1}^{12}$, note that the expectation operator was dropped because the respective Jacobians are independent of the target state. The expectation operator in the expression for $\mathbf{D}_{k-1}^{22}$ can be easily estimated using MC integration; it can also be calculated analytically, albeit the resulting formula appears cumbersome. Putting terms together yields

$$\mathbf{J}_k = \mathbf{Q}^{-1} + \mathbf{E}\{\tilde{\mathbf{F}_\mathbf{k}}^T \mathbf{R}^{-1} \tilde{\mathbf{F}_\mathbf{k}}\} - \mathbf{Q}^{-1}\mathbf{H}' \times$$

$$(\mathbf{J}_{k-1} + \mathbf{H}'^T \mathbf{Q}^{-1} \mathbf{H}')^{-1} \mathbf{H}'^T \mathbf{Q}^{-1}, k \geqslant 0$$

The initial density $\mathbf{p}(\mathbf{x}_0)$ is taken to be $\mathcal{N}(\bar{\mathbf{x}}_0, \mathbf{Q}_0)$, in which case $\mathbf{J}_0 = \mathbf{Q}_0^{-1}$.

5. SIMULATIONS

In our simulations, we benchmark the performance of our optimal particle filter against the CRLB and two additional particle filters: an Auxiliary PF, and a regularized PF. The three alternative particle filters are briefly discussed next.

5.1. Regularized PF (RPF)

This algorithm is identical to the Sampling Importance Resampling (SIR) algorithm, which uses the prior importance function, except for a "jittering" of the resampled particles (using a normal distribution kernel) in order to protect the filter from sample impoverishment; see, e.g., [2]. Since the process noise involved in our model is relatively small, this modification is expected to improve the performance over the standard SIR. However, this filter also has well known disadvantages - the samples are no longer guaranteed to approximate the posterior density asymptotically in the number of particles.

5.2. Auxiliary SIR (ASIR) Filter

The particular algorithm used is the Auxiliary SIR filter introduced by Pitt and Shephard (see [9]). This filter tries to explore the state-space in a more sophisticated way than the SIR filter. This is done by resampling at the "previous" time step based on certain point estimates that capture the essential features of the posterior density. This approximation can be inefficient when the process noise is large, or when the auxiliary index varies a lot for a fixed prior. When process noise is small enough, though, the ASIR filter is reported to improve the performance over the standard SIR.

5.3. PF Using Optimal Importance Function (PF-OIF)

For our particular model and choice of sampling procedure, an implementation is given in Table 1. Note that this algorithm allows both the weight update and the resampling step to be performed prior to sampling from the optimal importance function. An additional regularization step can be incorporated, if necessary, to improve the filter's diversity after resampling.

5.4. Estimation performance results

In the following, we focus on the frequency estimation performance of the three aforementioned filters in a tracking mode, wherein the initial state is assumed to be known exactly - corresponding to a Dirac delta initial distribution. The associated CRLB, however, assumes that the initial density is a Gaussian. This mismatch is dealt with by using a very tight density (very small initial variance) to approximate a delta distribution. The expectation appearing in the CRLB was approximated using 100 realizations of the state vector. The error curves corresponding to the three filters were produced by averaging over 200 independent runs, each comprising 80 temporal samples. The conditional mean was used to generate point state estimates. System parameters were set to $b_\ell = 0.999$, $\forall \ell$, $\sigma_\omega = 0.01$, $\sigma_A = 0.01$, $\sigma_n^2 = 0.2$, and multinomial resampling was employed. The number of particles, N, was 1000 for RPF, 800 for ASIR, and 30 for PF-OIF. The results are summarized in Fig. 2. It is satisfying to see that all three filters operate close to the CRLB, and PF-OIF in particular performs that well with order-of-magnitude less particles. This being a three-dimensional state-space, such good performance with only 30 particles is not at all obvious. RPF and ASIR filters perform very poorly with less than a few hundred particles in this context. A small number of particles implies small memory requirements, but on the other hand the use of rejection in our present implementation of PF-OIF entails a random delay, which can be significant, depending on system parameters. We are presently looking at possible ways of speeding up the sampling step.

6. CONCLUSIONS

We revisited the important problem of tracking a single time-varying harmonic, whose frequency and complex amplitude evolve according to a linear Gaussian separable AR(1) model. A key difficulty in treating this model comes from the nonlinear measurement equation. For this model, we derived the optimal importance function in closed form. This yields interesting insights and opens up the possibility of designing particle filters that are more efficient than generic ones. We also derived a procedure to sample from this optimal importance function, using rejection and the concept of a dominating density. Our preliminary numerical experiments comparing the resulting filter to standard particle filters and the CRLB show that the proposed PF-OIF algorithm has merits, particularly in terms of reducing the number of particles, and therefore memory requirements as well. Our present implementation of PF-OIF can be slow, due to the use of rejection. We are currently looking at other alternatives as well as extensions to more general signal models.

7. REFERENCES

[1] C. Andrieu, M. Davy, A. Doucet, "Improved Auxiliary Particle Filtering: Applications to Time-Varying Spectral Analysis", in *Proc. IEEE SSP 2001* Workshop, Singapore, Aug. 2001.

[2] M.S. Arulampalam, S. Maskell, N. Gordon, T. Clapp, "A tutorial on particle filters for nonlinear/non-Gaussian Bayesian tracking," *IEEE Trans. Signal Processing*, vol. 50, no. 2, pp. 174–188, Feb. 2002.

[3] L. Cohen, *Time-Frequency Analysis*, Prentice-Hall, 1994

[4] L. Devroye, *Non-uniform random variate generation*, Springer-Verlag, New York, 1986.

[5] P. Djuric, J.H. Kotecha, J. Zhang, Y. Huang, T. Ghirmai, M. Bugallo, J. Miguez, "Particle Filtering,", *IEEE Signal Processing Magazine*, pp. 19-38, Sep. 2003.

[6] A. Doucet, X. Wang, "Monte Carlo Methods for Signal Processing: A Review in the Statistical Signal Processing Context", *IEEE Signal Processing Magazine*, vol. 22, no. 6, pp. 152-170, Nov. 2005.

[7] C. Dubois, M. Davy, J. Idier, "Tracking of Time-Frequency Components Using Particle Filtering", in *Proc. IEEE ICASSP 2005*, March 18-23, 2005, Philadelphia, PA, U.S.A.

[8] I.S. Gradshteyn, I.M. Ryzhik (A. Jeffrey, Ed.), *Tables of Integrals, Series, and Products*, Academic Press, 5^{th} ed., 1994.

[9] M. Pitt and N. Shephard, "Filtering via simulation: Auxiliary Particle filters", in *Journal of the American Statistical Association*, vol. 94, no. 446, pp. 590-599, 1999.

[10] P. Stoica, R.L. Moses, *Spectral Analysis of Signals*, Prentice-Hall, 2005.

[11] P. Tichavsky, C.H. Muravchik, and A. Nehorai, "Posterior Cramer-Rao bounds for discrete-time dynamical systems", in *IEEE Trans. on Signal Processing*, vol. 46, no. 5, pp. 1386-1396, May 1998.

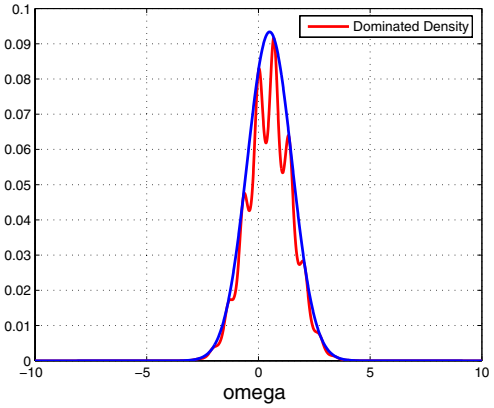


Fig. 1. Illustration of dominated (optimal importance function) and dominating densities as a function of frequency for fixed complex amplitude.

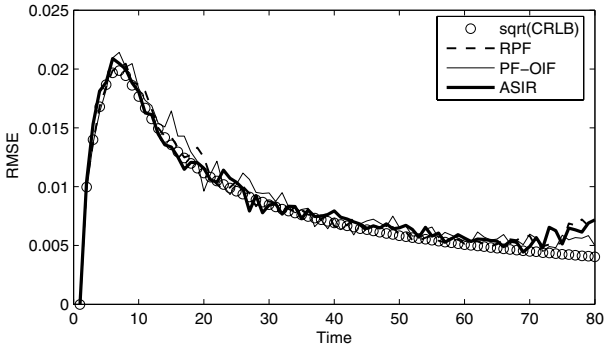


Fig. 2. Comparison of the three particle filters and the CRLB

Table 1. PF using OIF for Tracking A Single Time-Varying Harmonic (see text for definition of constants)

$$\underline{\left[\{\mathbf{x}_k^i\}_{i=1}^N\right] = PF-OIF\left[\{\mathbf{x}_{k-1}^i\}_{i=1}^N, \mathbf{y}_k\right]}$$

1. Compute normalized importance weights
   - FOR i=1:N,
     $\tilde{\mathbf{w}}_k^i = \frac{1}{2\pi(\sigma_A^2+\sigma_n^2)} e^{-\frac{|y_k|^2+|b_2 A_{k-1}^i|^2}{2(\sigma_A^2+\sigma_n^2)}} \times \mathcal{B}$
   - END FOR
   - FOR i=1:N,
     - Normalize : $\mathbf{w}_k^i = \tilde{\mathbf{w}}_k^i / sum\left[\{\tilde{\mathbf{w}}_k^i\}_{i=1}^N\right]$
   - END FOR
2. Resample $\rightarrow$ equally weighted particles
   $\left[\{\mathbf{x}_{k-1}^i\}_{i=1}^N\right] = RESAMPLE\left[\{\mathbf{x}_{k-1}^i, \mathbf{w}_k^i\}_{i=1}^N\right]$
3. Sample from the optimal importance density :
   - FOR i=1:N,
     - Calculate $\mathbf{C} := \sqrt{2\pi} Q_o / e^{-\frac{|b_2 A_{k-1}^i||y_k|}{\sigma_A^2+\sigma_n^2}} \mathcal{B}\sigma$
     - Set $U := 1/eps$ and $\tau := 1/eps$
   - WHILE $(U\tau > 1)$
     - Draw candidate sample $\sim$ dominating density:
       $\mathbf{x}_k^i \sim \frac{e^{-\frac{(\omega_k - b_1\omega_{k-1}^i)^2}{2\sigma_\omega^2}} e^{-\frac{(|A_k|-\mu)^2}{2\sigma^2}}}{(2\pi)^2 Q_o\ \sigma_\omega \sigma}$
     - Set $\tau := \mathbf{C}\frac{\text{Dominating}(\mathbf{x}_k^i)}{\text{Optimal}(\mathbf{x}_k^i)}$
     - Draw a sample $U \sim \text{Uniform}[0,1]$
   - END WHILE
   - END FOR

TRACKING A FREQUENCY HOPPED SIGNAL USING PARTICLE FILTERING

N. D. Sidiropoulos[*] *A. Swami* *A. Valyrakis*

Dept. ECE, Tech. Univ. of Crete
73100 Chania - Crete, Greece
nikos@telecom.tuc.gr

Army Research Laboratory
Adelphi, MD, 20783, U.S.A.
a.swami@ieee.org

Dept. ECE, Tech. Univ. of Crete
73100 Chania - Crete, Greece
alevali@telecom.tuc.gr

ABSTRACT

The problem of tracking the frequency and complex amplitude of a frequency-hopped complex sinusoid is considered, using a novel stochastic state-space formulation and particle filtering tools. The problem is of considerable interest for interference mitigation in frequency-hopped wireless networks, and in military communications. The proposed particle filtering approach has a number of desirable features. It affords high-resolution estimates of carrier frequency and hop timing, manageable complexity (linear in the number of processed samples), and flexibility in tracking signals with irregular hopping patterns due to intentional timing jitter. The proposed state-space model is not only parsimonious, but fortuitous as well: it turns out that the associated optimal importance function can be computed in closed form, and thus samples from it can be drawn using rejection techniques. Both prior and optimal importance sampling versions are developed and illustrated in pertinent simulations.

Keywords: Frequency hopping, spectral analysis, estimation of time-varying line spectra, sequential importance sampling, particle filtering

1. INTRODUCTION

Tracking the frequency of a time-varying complex sinusoid is an important problem which arises in numerous applications. In speech processing, for example, one is often interested in tracking formant frequencies. In wireless communications, it arises in the context of frequency hopping, when the receiver has no prior knowledge of the hopping pattern, or is simply out of sync with the transmitter's hopping pattern generator [2, 8, 6, 7].

Both non-parametric time-frequency analysis, and parametric techniques have been developed for the more general problem of tracking a time-varying sinusoid, and can be applied to the problem of tracking a frequency-hopped sinusoid as well. However, existing methods have limitations, especially when used to track a frequency-hopped signal. Non-parametric methods, like the spectrogram, or coarse channelization [2] suffer from limited frequency- and temporal-resolution due to leakage. It is possible to employ time-frequency distributions that are better-adapted to frequency hopping [3], but the results are still not very satisfactory. Parametric methods for frequency hopping explicitly model the frequency as piecewise-constant, assume a "budget" on the number of hops within a given observation interval, and employ dynamic programming to track the sought frequency and complex amplitude parameters [6, 7]. Other than an upper bound on the number of hops, the methods in [6, 7] do not assume anything else about the frequencies or complex amplitudes, which are treated as deterministic unknowns.

A different viewpoint is adopted in this paper. A stochastic non-linear, non-Gaussian state-space formulation is proposed, which captures frequency hopping dynamics in a probabilistic sense. The proposed formulation is naturally well-suited for the application of particle filtering for state estimation. Compared to the prior state-of-art in [6, 7], the new approach has a number of desirable features:

• Computational complexity: The complexity of particle filtering is $O(NT)$, where N is the number of particles and T is the number of temporal samples. The complexity of dynamic programming, on the other hand, is roughly $O(T^4)$. This means that only short segments can be processed by dynamic programming, and then one has to rely on hop periodicity to segment the rest of the data. This has two disadvantages: first, the more samples are processed the better from an estimation performance perspective; second, hop timing is often intentionally randomized as a countermeasure.

• Flexibility: The state-space model in the particle filtering formulation can be easily tailored to match a given scenario (e.g., spread bandwidth and modulation).

The proposed state-space model is simple and fortuitous: the associated optimal importance function can be computed in closed form, and thus samples from it can be drawn using rejection techniques. Both prior and optimal importance sampling versions are developed and compared in pertinent simulations.

2. DATA MODEL AND PROBLEM STATEMENT

We propose the following non-linear non-Gaussian stochastic state-space model of a frequency-hopped complex sinusoid. Let $\mathbf{x}_k := [\omega_k, A_k]^T$ denote the state at time k, where $\omega_k \in [-\pi, \pi)$ and $A_k \in \mathbb{C}$ denote instantaneous frequency and complex amplitude. Let $\mathbf{u}_k := [b_k, \tilde{\omega}_k, \tilde{A}_k]^T$ denote an auxiliary sequence of independent and identically distributed (i.i.d.) vectors with independent components and the following marginal statistics: b_k is a binary random variable with $Pr(b_k = 1) = h$; $\tilde{\omega}_k$ is uniformly distributed over $[-\pi, \pi)$, denoted $\mathcal{U}([-\pi, \pi))$; and $\tilde{A}_k$ is $\mathcal{CN}(0, \sigma_A^2)$, i.e., complex circular Gaussian of variance σ_A^2. Then

$$\mathbf{x}_k = f(\mathbf{x}_{k-1}, \mathbf{u}_k) = \begin{cases} \mathbf{x}_{k-1} & , \mathbf{u}_k(1) = 0 \\ [\mathbf{u}_k(2),\ \mathbf{u}_k(3)]^T & , \mathbf{u}_k(1) = 1 \end{cases}$$

[*]Corresponding author. Supported in part by the Army Research Laboratory (ARL) through participation in the ARL Collaborative Technology Alliance (ARL-CTA) for Communications and Networks under Cooperative Agreement DADD19-01-2-0011, and in part by ARO under ERO Contract N62558-03-C-0012.

$$= \begin{cases} \mathbf{x}_{k-1} & , w.p.\ 1-h \\ \left[\mathcal{U}\left([-\pi,\pi)\right),\ \mathcal{CN}(0,\sigma_A^2)\right]^T & , w.p.\ h \end{cases},$$

$$y_k = \mathbf{x}_k(2) e^{j\mathbf{x}_k(1)k} + v_k,$$

where v_k denotes i.i.d. $\mathcal{CN}(0,\sigma_n^2)$ measurement noise, and $u_k(1)$ the hop variable.

The above state-space formulation models frequency hopping in a probabilistic fashion. Hops are random, i.i.d., with hop probability h per sample interval. This is different from traditional models of frequency hopping, which assume that the frequency hops periodically, and is motivated by the following considerations:

- In military communications, intentional jitter is often introduced in the hop timing in order to reduce the probability of detection by unintended receivers and improve immunity to jamming. Timing jitter yields a pseudo-random quasi-periodic, or even seemingly aperiodic hop timing sequence.
- The above probabilistic model captures information about the average hop rate in a "soft" ensemble sense: the expected number of hops over a long observation interval T is hT. While less accurate if the exact hop period is known, probabilistic modeling is more robust with respect to hop period inaccuracies. Finally,
- The proposed probabilistic model is ideally suited for on-line sequential estimation via particle filtering.

It is worth elaborating on some of the implicit assumptions of the proposed state-space model.

1. When the (discrete-time, baseband-equivalent) frequency hops, it hops anywhere within $[-\pi,\pi)$ with a uniform density. This is well-suited for carrier hopping, which is usually discontinuous. Modulation-induced variations can (and should) be neglected when the objective is to estimate carrier frequency, but could also be explicitly modeled using, e.g., a smooth auto-regressive frequency variation model in-between hops, in lieu of the simplified constant model postulated above. This extension is relatively simple.
2. When the frequency hops, the complex amplitude also changes according to an i.i.d. complex Gaussian distribution. This is also well-motivated for carrier hopping, for every time the carrier frequency hops beyond the coherence bandwidth of the channel, a new channel realization is encountered.

The problem, then, can be stated as follows: Given a sequence of observations $\{y_k\}_{k=1}^T$, estimate the sequence of system states $\{\mathbf{x}_k\}_{k=1}^T$ - that is, the unknown carrier frequencies and complex amplitudes.

3. PARTICLE FILTERING SOLUTIONS

Particle filtering has emerged as an important sequential state estimation method for stochastic non-linear and/or non-Gaussian state-space models, for which it provides a powerful alternative to the commonly used extended Kalman filter. See [1, 5] for recent tutorial overviews. In particle filtering, continuous distributions are approximated by discrete random measures, comprising "particles" and associated weights. That is, a certain continuous distribution of interest, say $p(\mathbf{x})$, is approximated as

$$p(\mathbf{x}) \approx \sum_{n=1}^{N} w_n \delta(\mathbf{x} - \mathbf{x}_n),$$

where $\delta(\cdot)$ denotes the Dirac delta functional. A useful simplification stemming from this approximation is that the computation of pertinent expectations and conditional probabilities reduces to summation, as opposed to integration. While this can also be accomplished via direct discretization over a fixed grid, the use of a random measure affords flexibility in adapting the particle locations to better fit the distribution of interest.

3.1. Basics of particle filtering

If we aim for an on-line filtering algorithm, in which the state at time k should be estimated from measurements up to and including time k, the key distribution of interest is the posterior density $p\left(\mathbf{x}_k \mid \{y_l\}_{l=1}^k\right)$. Given this density, one can estimate the state at time k, e.g., via the associated (posterior) mean, or mode. The basic idea of particle filtering, then, is to begin with a random measure approximation of the initial state distribution, and, as measurements become available, derive updated random measure approximations of $p\left(\mathbf{x}_k \mid \{y_l\}_{l=1}^k\right)$, $k \in \{1, 2, \cdots\}$. That is, we seek random measure approximations

$$\hat{p}\left(\mathbf{x}_k \mid \{y_l\}_{l=1}^k\right) = \sum_{n=1}^{N} w_{n,k} \delta(\mathbf{x}_k - \mathbf{x}_{n,k})$$

In particle filtering, the updates - the derivation of $\hat{p}\left(\mathbf{x}_k \mid \{y_l\}_{l=1}^k\right)$ from $\hat{p}\left(\mathbf{x}_{k-1} \mid \{y_l\}_{l=1}^{k-1}\right)$ - are based on the Bayes rule [1, 5].

A random measure approximation comprises two components: the particles (locations) and the associated weights. If we could sample from the sought posterior $p\left(\mathbf{x}_k \mid \{y_l\}_{l=1}^k\right)$, then all particle weights would have been equal. Unfortunately, such direct sampling is not possible in most cases, and thus we resort to sampling from a so-called *importance function* that "resembles" the desired posterior, and from which samples can be drawn with relative ease. The mismatch between the sought density and the importance function is compensated in the calculation of weights, chosen proportional to their ratio evaluated at each particle [1, 5]. The choice of importance function is a very important step in the design of a particle filtering algorithm. Two common choices are discussed next.

3.2. Prior importance function

Perhaps the most intuitive choice of importance function is the *prior importance function* $p(\mathbf{x}_k \mid \mathbf{x}_{n,k-1})$; i.e., the n-th particle is updated by propagating it through the state-evolution part of the system: $\mathbf{x}_{n,k} = f(\mathbf{x}_{n,k-1}, \mathbf{u}_n)$. This is an often-made choice, for simplicity considerations. The drawback is that particles evolve without regard to the latest measurement, which only comes into play in the ensuing weight update. When using the prior importance function, the said weight update at time instant k is given by $w_{n,k} = w_{n,k-1} p(y_k \mid \mathbf{x}_{n,k})$, followed by normalization to enforce $\sum_{n=1}^{N} w_{n,k} = 1$.

Regardless of the particular importance function employed, a common problem in particle filtering is *degeneracy*: the weights of all but a few particles tend to become negligible after a few iterations [1, 5]. Degeneracy can be detected via degeneracy measures, and mitigated via *resampling* techniques [1, 5]. Resampling the discrete measure replicates particles with large weights and removes those with negligible weights. All particle weights become

equal after resampling. There exist several computationally efficient ($O(N)$) resampling schemes that can be used to avoid the quadratic cost of brute-force resampling [1, 5].

3.3. Optimal importance function

From the viewpoint of minimizing the variance of the weights, the optimal importance function is given by [1, 5]

$$p(\mathbf{x}_k|\mathbf{x}_{n,k-1}, y_k) = \frac{p(y_k|\mathbf{x}_k)p(\mathbf{x}_k|\mathbf{x}_{n,k-1})}{\int_{\mathbf{x}} p(y_k|\mathbf{x})p(\mathbf{x}|\mathbf{x}_{n,k-1})d\mathbf{x}}.$$

Notice that, in contrast to the prior importance function, the above takes into account the newly available measurement in the particle update itself. While both the prior importance function and the optimal one yield consistent algorithms[1], the optimal one usually works well with much smaller N, and is therefore preferable from a performance point of view. There are, however, two difficulties associated with the use of the optimal importance function. First, it requires multidimensional integration to compute the normalization factor, which is often intractable. Second, sampling from the optimal importance function is more complicated than sampling from the prior. The smaller number of particles needed to attain satisfactory performance with the optimal importance function usually more than offsets the cost of drawing samples from it; the integration problem remains the bottleneck in most cases [1]. Thankfully, for our particular model, it turns out that it is possible to carry out this integration analytically. This is explained next.

Denote $\mathbf{x}_k := [\omega_k, A_k]^T$, where $\omega_k \in [-\pi, \pi)$, and $A_k \in \mathbb{C}$; likewise $\mathbf{x}_{n,k-1} := [\omega_{n,k-1}, A_{n,k-1}]^T$, and a dummy variable $\mathbf{x} := [\omega, A]^T$. Let $D(y_k, \mathbf{x}_{n,k-1}) := \int_{\mathbf{x}} p(y_k|\mathbf{x})p(\mathbf{x}|\mathbf{x}_{n,k-1})d\mathbf{x}$. Then

$$D(y_k, \mathbf{x}_{n,k-1}) = \int_{\omega\in[-\pi,\pi)} \int_{A\in\mathbb{C}} \frac{1}{2\pi\sigma_n^2} e^{-\frac{|y_k - Ae^{j\omega k}|^2}{2\sigma_n^2}} \times$$

$$\left[(1-h)\delta(\omega - \omega_{n,k-1})\delta(A - A_{n,k-1}) + \frac{h}{2\pi}\frac{1}{2\pi\sigma_A^2} e^{-\frac{|A|^2}{2\sigma_A^2}}\right] dAd\omega$$

This integral can be computed by completing the squares, yielding

$$D(y_k, \mathbf{x}_{n,k-1}) = \frac{1}{2\pi}\frac{h}{\sigma_n^2 + \sigma_A^2} e^{-\frac{|y_k|^2}{2(\sigma_n^2+\sigma_A^2)}} +$$

$$\frac{1}{2\pi}\frac{1-h}{\sigma_n^2} e^{-\frac{|y_k - A_{n,k-1}e^{j\omega_{n,k-1}k}|^2}{2\sigma_n^2}}.$$

For the above optimal choice of the importance function, the weight update is given by

$$w_{n,k} \propto w_{n,k-1} p(y_k|\mathbf{x}_{n,k-1}) = w_{n,k-1} D(y_k, \mathbf{x}_{n,k-1}),$$

followed by normalization to 1. What is missing is a way to sample from the optimal importance function. As a first step towards this end, note that $p(\mathbf{x}_k|\mathbf{x}_{n,k-1}, y_k)$ can be written as a mixture of two pdfs

$$p(\mathbf{x}_k|\mathbf{x}_{n,k-1}, y_k) = (1-\tilde{h})p_0(\mathbf{x}_k|\mathbf{x}_{n,k-1}, y_k) + \tilde{h}p_1(\mathbf{x}_k|\mathbf{x}_{n,k-1}, y_k),$$

where

$$p_0(\mathbf{x}_k|\mathbf{x}_{n,k-1}, y_k) := \delta(\omega_k - \omega_{n,k-1})\delta(A_k - A_{n,k-1}),$$

$$p_1(\mathbf{x}_k|\mathbf{x}_{n,k-1}, y_k) := \frac{\frac{1}{2\pi}\frac{1}{2\pi\sigma_n^2}\frac{1}{2\pi\sigma_A^2} e^{-\frac{|y_k - A_k e^{j\omega_k k}|^2}{2\sigma_n^2}} e^{-\frac{|A_k|^2}{2\sigma_A^2}}}{\frac{1}{2\pi}\frac{1}{\sigma_n^2+\sigma_A^2} e^{-\frac{|y_k|^2}{2(\sigma_n^2+\sigma_A^2)}}},$$

and

$$\tilde{h} := h\frac{\frac{1}{2\pi}\frac{1}{\sigma_n^2+\sigma_A^2} e^{-\frac{|y_k|^2}{2(\sigma_n^2+\sigma_A^2)}}}{D(y_k, \mathbf{x}_{n,k-1})}.$$

It follows that with probability $1 - \tilde{h}$ we simply copy the previous particle, else we draw a particle from $p_1(\mathbf{x}_k|\mathbf{x}_{n,k-1}, y_k)$. We will use rejection sampling techniques for this latter step, as explained next.

3.4. Sampling from the optimal importance function: Rejection

The basic idea of rejection-based sampling can be summarized as follows [4, pp. 40-42]. Suppose we wish to draw samples from a density $\phi(\mathbf{x})$, for which there exists a *dominating density* $g(\mathbf{x})$ and a known constant c such that $\phi(\mathbf{x}) \leq cg(\mathbf{x}), \forall \mathbf{x}$. In practice, we choose $g(\mathbf{x})$ to be easy to sample from, and such that c is as small as possible. The rejection method then works as follows. We i) draw a sample $\mathbf{x}$ from $g(\cdot)$ and an independent sample $U \sim \mathcal{U}([0, 1])$; ii) set $\tau := c\frac{g(\mathbf{x})}{\phi(\mathbf{x})}$; iii) test whether $U\tau \leq 1$; if so, we accept the sample $\mathbf{x}$; else we reject it and repeat the process.

It can be shown that the above rejection method generates samples from the desired density $\phi(.)$, and the mean number of iterations until a sample is accepted is c (thus the desire to keep $c \geq 1$ as small as possible). Furthermore, the distribution of the number of trials is geometic with parameter $1 - \frac{1}{c}$, which means that the probabilities of longer trials decay exponentially [4, p. 42].

In our present context, we wish to sample from the density $p_1(\mathbf{x}_k|\mathbf{x}_{n,k-1}, y_k)$. Define

$$\mu := \frac{|y_k|\sigma_A^2}{\sigma_n^2 + \sigma_A^2}, \quad \sigma^2 := \frac{\sigma_n^2\sigma_A^2}{\sigma_n^2 + \sigma_A^2}.$$

Using the triangle inequality, it can be shown that the following is a suitable dominating density:

$$g(\mathbf{x}_k|\mathbf{x}_{n,k-1}, y_k) = \frac{e^{-\frac{(|A_k| - \mu)^2}{2\sigma^2}}}{(2\pi)^{5/2}Q_0\sigma},$$

for which it holds that $p_1(\mathbf{x}_k|\mathbf{x}_{n,k-1}, y_k) \leq cg(\mathbf{x}_k|\mathbf{x}_{n,k-1}, y_k)$, with

$$c := \sqrt{2\pi}Q_0/\sigma \geq 1,$$

$$Q_0 := \int_{r=0}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(r-\mu)^2}{2\sigma^2}} dr = \frac{1}{2}\text{erfc}(-\frac{|y_k|\sigma_A}{\sigma_n\sqrt{2(\sigma_n^2 + \sigma_A^2)}}).$$

[1] In the sense that the pertinent discrete measure approximations converge to the sought continuous distributions as $N \to \infty$, see [1] and references therein.

Through experimentation, we have found that even better results can be attained using an *outer rejection loop*, which declines candidates $\mathbf{x}_{n,k}$ generated through rejection when the following metric exceeds a certain small value (set to 3×10^{-3} in our experiments):

$$\tilde{h}(y_k, \mathbf{x}_{n,k}) := h \frac{\frac{1}{2\pi} \frac{1}{\sigma_n^2 + \sigma_A^2} e^{-\frac{|y_k|^2}{2(\sigma_n^2 + \sigma_A^2)}}}{D(y_k, \mathbf{x}_{n,k})},$$

where $D(\cdot, \cdot, \cdot)$ was defined in Sec. 3.3. This outer rejection loop selects particles that are consistent with the new measurement (cf. the functional form of the denominator) and, at the same time, have large weight after the associated update. We do not have a full explanation at this point, yet this version of the algorithm appears to yield the best results - in particular, better than the one based on the optimal importance function. Note that the latter is optimal with respect to minimizing the variance of the weights after the update (and typically works better than the one based on the prior importance function), but it is not necessarily optimal in terms of the performance - complexity trade-off.

4. SIMULATIONS

We now present simulation results for the three algorithms: the basic one using the prior importance function (denoted P), the one using the optimal importance function (O), and the one using the outer rejection loop as above (V). Fig. 1 shows a plot of a typical simulation run, using the posterior mean to form instantaneous frequency estimates and multinomial resampling for all three algorithms. Monte-Carlo (MC) simulation results are presented in Fig. 2. The Root Mean Square Error (RMSE) frequency estimation performance of the three algorithms is assessed using the following parameters: $h = 0.01$, $T = 100$, $\sigma_A^2 = 1$, $\sigma_n^2 = 0.2$, and the number of MC trials is 300. The execution time for P is $O(NT)$, whereas for O and V the execution time is also an increasing function of h. As a result, O and/or V can be faster than P, even for the same number of particles. For our simulation setup above, P, O, and V, each with 1K particles, have about the same average execution time, yet V does much better in terms of RMSE as shown in Fig. 2. It takes 3K particles for O and 5K particles for P to reach the performance of V with 1K particles.

5. CONCLUSIONS

We have developed three new particle filtering algorithms for tracking a frequency-hopped complex sinusoid, based on a novel stochastic state-space formulation. The algorithms range from a plain-vanilla version that uses the prior importance function (P), to a more advanced version that employs the optimal importance function (O), and, finally, an improvement of the latter using a problem-specific outer rejection loop (V). The two latter algorithms afford considerably better performance - complexity trade-offs.

6. REFERENCES

[1] M.S. Arulampalam, S. Maskell, N. Gordon, T. Clapp, "A tutorial on particle filters for nonlinear/non-Gaussian Bayesian tracking," *IEEE Trans. Signal Processing*, vol. 50, no. 2, pp. 174–188, Feb. 2002.

[2] L. Aydin and A. Polydoros, "Hop-timing estimation for FH signals using a coarsely channelized receiver," *IEEE Trans. Communcations*, vol. 44, no. 4, pp. 516–526, Apr. 1996.

[3] S. Barbarossa and A. Scaglione, "Parameter estimation of spread spectrum frequency-hopping signals using time-frequency distributions," in *Proc. Signal Proc. Advances in Wireless Communications*, pp. 213–216, Apr. 1997.

[4] L. Devroye, *Non-uniform random variate generation*, Springer-Verlag, New York, 1986. Available on-line at http://cgm.cs.mcgill.ca/∼luc/rnbookindex.html

[5] P. Djuric, J.H. Kotecha, J. Zhang, Y. Huang, T. Ghirmai, M. Bugallo, J. Miguez, "Particle Filtering,", *IEEE Signal Processing Magazine*, pp. 19–38, Sep. 2003.

[6] X. Liu, N. D. Sidiropoulos, and A. Swami, "Blind high resolution localization and tracking of multiple frequency hopped signals," *IEEE Trans. Signal Processing*, vol. 50, no. 4, pp. 889–901, Apr. 2002.

[7] X. Liu, N. D. Sidiropoulos, and A. Swami, "Joint Hop Timing and Frequency Estimation for Collision Resolution in Frequency Hopped Networks," *IEEE Trans. Wireless Communications*, to appear, Nov. 2005.

[8] M. K. Simon, U. Cheng, L. Aydin, A. Polydoros, and B. K. Levitt, "Hop timing estimation for noncoherent frequency-hopped M-FSK intercept receivers," *IEEE Trans. Communications*, vol. 43, no. 2/3/4, pp. 1144–1154, Feb./Mar./Apr. 1995.

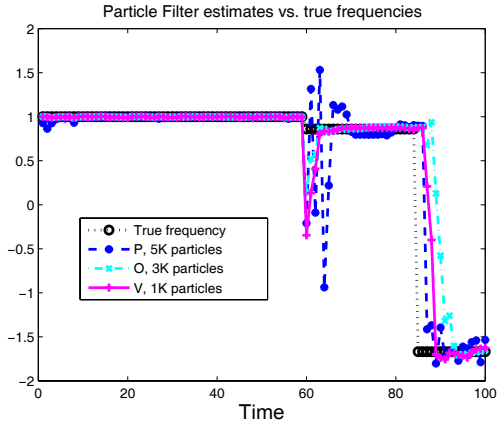


Fig. 1. Typical sample run of the three algorithms using different number of particles for each.

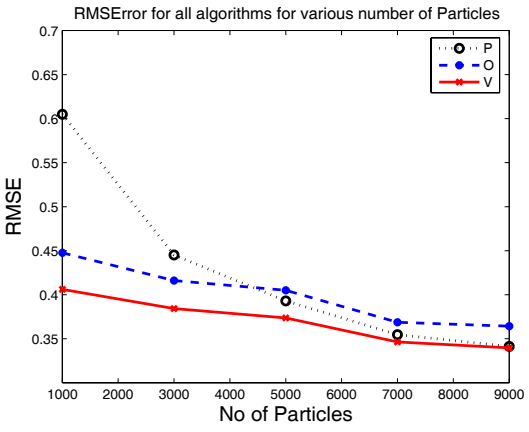


Fig. 2. MC simulation results for the three algorithms.

CONVEX TRANSMIT BEAMFORMING FOR DOWNLINK MULTICASTING TO MULTIPLE CO-CHANNEL GROUPS

Eleftherios Karipidis, Nicholas D. Sidiropoulos **

Dept. of ECE, Tech. Univ. of Crete
73100 Chania - Crete, Greece

Zhi-Quan Luo †

Dept. of ECE, Univ. of Minnesota
Minneapolis, MN 55455, U.S.A.

ABSTRACT

We consider the problem of transmit beamforming to multiple co-channel multicast groups. Since the direct minimization of transmit power while guaranteeing a prescribed minimum signal to interference plus noise ratio (SINR) at each receiver is nonconvex and NP-hard, we present convex SDP relaxations of this problem and study when such relaxations are tight. Our results show that when the steering vectors for all receivers are of Vandermonde type (such as in the case of a uniform linear array and line-of-sight propagation), a globally optimum solution to the corresponding transmit beamforming problem can be obtained via an equivalent SDP reformulation. We also present various robust formulations for the problem of single-group multicasting, when the steering vectors are only approximately known. Simulation results are presented to illustrate the effectiveness of our SDP relaxations and reformulations.

1. INTRODUCTION

Consider a downlink transmission scenario where the transmitter is equipped with N antennas and there are M receivers. Let $\mathbf{h}_i$ denote the $N \times 1$ complex channel vector from each transmit antenna to the single receive antenna of user $i \in \{1, \ldots, M\}$. Let there be a total of $1 \leq G \leq M$ multicast groups, $\{\mathcal{G}_1, \ldots, \mathcal{G}_G\}$, where $\mathcal{G}_k$ is the index set for receivers participating in multicast group k, and $k \in \{1, \ldots, G\}$. Assume that $\mathcal{G}_k \cap \mathcal{G}_l = \emptyset$, $l \neq k$, $\cup_k \mathcal{G}_k = \{1, \ldots, M\}$, and, denoting $G_k := |\mathcal{G}_k|$, $\sum_{k=1}^{G} G_k = M$.

Let $\mathbf{w}_k^H$ denote the beamforming weight vector applied to the N transmitting antenna elements to transmit multicast stream k. The signal transmitted by the antenna array is equal to $\sum_{k=1}^{G} \mathbf{w}_k^H s_k(t)$, where $s_k(t)$ is the temporal information-bearing signal directed to receivers in multicast group k. This setup includes the case of *broadcasting* ($G = 1$) [6], and the case of individual user transmissions ($G = M$) [2]) as special cases. If each $s_k(t)$ is zero-mean white with unit variance, and the waveforms $\{s_k(t)\}_{k=1}^{G}$ are mutually uncorrelated, then the total power radiated is equal to $\sum_{k=1}^{G} ||\mathbf{w}_k||_2^2$.

The joint design of transmit beamformers subject to received SINR constraints can then be posed as follows:

$$\mathcal{P}: \quad \min_{\{\mathbf{w}_k \in \mathbb{C}^N\}_{k=1}^{G}} \sum_{k=1}^{G} \|\mathbf{w}_k\|_2^2$$
$$s.t.: \quad \frac{|\mathbf{w}_k^H \mathbf{h}_i|^2}{\sum_{l \neq k} |\mathbf{w}_l^H \mathbf{h}_i|^2 + \sigma_i^2} \geq c_i, \ \forall i \in \mathcal{G}_k, \ \forall k \in \{1, \ldots, G\}.$$

Problem $\mathcal{P}$ was considered in [5] and it was found to be NP-hard, in the case of general steering vectors, based on arguments proved in earlier work [6]. Therefore, a two step approach was proposed and shown to yield high-quality approximate solutions at manageable complexity cost. Specifically, in the first step, the original non-convex quadratically constrained quadratic programming (QCQP) problem $\mathcal{P}$ is relaxed to a semidefinite program (SDP) (denoted as $\mathcal{R}$), by changing the optimization variables to $\mathbf{X}_k := \mathbf{w}_k \mathbf{w}_k^H$ and dropping the associated non-convex constraints $\{\mathrm{rank}(\mathbf{X}_k) = 1\}_{k=1}^{G}$. In the second step, a randomization procedure is employed to generate candidate beamforming vectors from the solution of $\mathcal{R}$. For each candidate set of vectors, a multi-group power control ($\mathcal{MGPC}$) linear programming (LP) problem is solved to ensure that the constraints of the original problem $\mathcal{P}$ are met. The final solution of this algorithm is the set of beamforming vectors yielding the smallest $\mathcal{MGPC}$ objective. The overall complexity of the algorithm is manageable, since the SDP and LP problems can be solved efficiently using interior point methods and the randomization procedure is designed so that its computational cost is negligible compared to the aforementioned problems.

2. EXACT GLOBALLY OPTIMAL SOLUTION IN THE VANDERMONDE CASE

When a uniform linear array (ULA) is used for far-field transmit beamforming, the $N \times 1$ complex vectors which model the phase shift from each transmit antenna to the receive antenna of user $i \in \{1, \ldots, M\}$ are Vandermonde $\mathbf{h}_i = [1 \ e^{j\theta_i} \ e^{j2\theta_i} \ \cdots \ e^{j(N-1)\theta_i}]^T$. In this scenario, we observed that when the relaxed SDP problem $\mathcal{R}$ in [5] is feasible, its optimal solution, i.e., the blocks $\{\mathbf{X}_k^{\mathrm{opt}}\}_{k=1}^{G}$, are all consistently rank-one. This means that problem $\mathcal{R}$ is then equivalent to, and not a relaxation of, the original problem $\mathcal{P}$. Thus, the second step of the proposed algorithm, comprising the randomization - multicast power control loop, turns out being redundant and the set of the optimum beamforming vectors $\{\mathbf{w}_k^{\mathrm{opt}}\}_{k=1}^{G}$ can be formed simply using the principal components of the blocks $\{\mathbf{X}_k^{\mathrm{opt}}\}_{k=1}^{G}$. This observation suggests that, in the case of Vandermonde channel vectors, the original problem $\mathcal{P}$ is no longer NP-hard and can be equivalently posed as a convex optimization problem.

Towards this end, note that for the special case of Vandermonde steering vectors, the signal power received at each user can be rewrit-

*Tel: +302821037227, Fax: +302821037542, E-mail: (karipidis,nikos)@telecom.tuc.gr. Supported in part by the U.S. ARO under ERO Contract N62558-03-C-0012, and the E.U. under FP6 U-BROAD STREP # 506790

†E-mail: luozq@ece.umn.edu. Supported in part by the National Science Foundation, Grant No. DMS-0312416, and by the Natural Sciences and Engineering Research Council of Canada, Grant No. OPG0090391.

ten as

$$\left|\mathbf{w}_k^H \mathbf{h}_i\right|^2 = \sum_{\ell=-(N-1)}^{N-1} r_k(\ell) e^{j\theta_i \ell}, \quad (1)$$

where $\ell := n - m$ and $r_k(\ell) := \sum_{m=\max(1-\ell,1)}^{\min(N-\ell,N)} w_k(m) w_k^*(m+\ell)$. Let us consider $r_k(\ell)$ for $0 < \ell \leq N-1$, i.e., $r_k(\ell) = \sum_{m=1}^{N-\ell} w_k(m) w_k^*(m+\ell)$. Then $r_k^*(-\ell) = r_k(\ell)$, i.e., $r_k(\ell)$ is conjugate-symmetric about the origin. Define the $(2N-1) \times 1$ vector

$$\mathbf{r}_k := [r_k(-N+1), \cdots, r_k(-1), r_k(0), r_k(1), \cdots, r_k(N+1)]^T, \quad (2)$$

and the associated $(2N-1) \times 1$ "extended" steering vector

$$\mathbf{f}_i := [e^{-j\theta_i(N-1)}, \ \cdots, \ e^{-j\theta_i}, \ 1, \ e^{j\theta_i}, \ \cdots, e^{j\theta_i(N-1)}]^T. \quad (3)$$

Then $\left|\mathbf{w}_k^H \mathbf{h}_i\right|^2 = \mathbf{f}_i^T \mathbf{r}_k$. Furthermore, note that $r_k(0) = \mathbf{r}_k(N) = \sum_{m=1}^{N} w_k(m) w_k^*(m) = ||\mathbf{w}_k||_2^2$. It therefore follows that the original problem $\mathcal{P}$ can be equivalently written as follows

$$\min_{\{\mathbf{r}_k\}_{k=1}^G} \ \sum_{k=1}^{G} \mathbf{r}_k(N)$$

$$\textit{s.t.}: \quad \mathbf{f}_i^T \mathbf{r}_k \geq c_i \sum_{\ell \neq k} \mathbf{f}_i^T \mathbf{r}_\ell + c_i \sigma_i^2, \ \forall i \in \mathcal{G}_k, \ \forall k \in \{1, \ldots, G\},$$

$$\mathbf{r}_k : \text{autocorrelation vector}, \ \forall k \in \{1, \ldots, G\},$$

where the fact that the terms in the denominator are all non-negative has also been taken into account.

This is a problem comprising a linear cost, M linear inequality constraints, and autocorrelation constraints. Each of the latter is equivalent to a linear matrix inequality (LMI) constraint [1]. Specifically, $r_k(m), \ \forall m \in \{-N+1, \ldots, N-1\}$ belongs to the set of finite autocorrelation sequences if and only if $r_k(m) = \text{trace}(\mathbf{E}^m \mathbf{Y}_k)$, $\forall m \in \{-N+1, \ldots, N-1\}$, for some positive semidefinite matrix $\mathbf{Y}_k \in \mathbb{C}^{N \times N}$, where $\mathbf{E}$ is the $N \times N$ unit-shift matrix with ones in the first lower sub-diagonal and zeros elsewhere.

Thus, introducing G positive semidefinite $N \times N$ "slack" matrices, one for each autocorrelation vector $\mathbf{r}_k$, the autocorrelation constraints are equivalently converted to linear equality constraints plus positive semidefinite constraints as follows

$$\begin{array}{|l|} \hline \mathcal{V}: \\ \qquad \min_{\{\mathbf{r}_k\}_{k=1}^G, \ \{\mathbf{Y}_k\}_{k=1}^G} \ \sum_{k=1}^{G} \mathbf{r}_k(N) \\ \textit{s.t.}: \quad \mathbf{f}_i^T \mathbf{r}_k - c_i \sum_{\ell \neq k} \mathbf{f}_i^T \mathbf{r}_\ell \geq c_i \sigma_i^2, \\ \qquad \forall i \in \mathcal{G}_k, \ \forall k \in \{1, \ldots, G\}, \\ \qquad \mathbf{r}_k(m) = \text{trace}(\mathbf{E}^m \mathbf{Y}_k), \\ \qquad \forall m \in \{-N+1, \ldots, N-1\}, \ \forall k \in \{1, \ldots, G\} \\ \qquad \mathbf{Y}_k \succeq \mathbf{0}, \ \forall k \in \{1, \ldots, G\}. \\ \hline \end{array}$$

Problem $\mathcal{V}$ is an SDP problem which can be efficiently solved by any standard SDP solver, such as SeDuMi [7], by means of interior point methods. Once the optimum autocorrelation sequences $\left\{\mathbf{r}_k^{\text{opt}}\right\}_{k=1}^G$ are found, they can be factored to obtain the respective optimum beamforming vectors $\left\{\mathbf{w}_k^{\text{opt}}\right\}_{k=1}^G$, using spectral factorization techniques [9].

A simple simulation experiment illustrates the equivalence of the aforementioned algorithm to the one proposed in [5]. Figures 1 and 2 show the optimized transmit beam patterns generated by algorithm 1 (SDP relaxation problem $\mathcal{R}$ and randomization - multicast power control problem $\mathcal{MGPC}$) and algorithm 2 (SDP problem $\mathcal{V}$ and spectral factorization), respectively. The ULA consists of $N = 4$ transmit antenna elements spaced $\lambda/2$ apart. The $M = 24$ users are considered evenly clustered in $G = 2$ groups, at an angle of 0.5 degrees to their neighboring ones. The angular cluster separation (defined as the minimum angle between any 2 users belonging to different groups) is set to 10 degrees. The received SINR constraints are set to 10dB for all users and the noise variance to $\sigma^2 = 1$ for all channels.

3. ROBUST RELAXATION OF SINGLE-GROUP MULTICAST BEAMFORMING

In this section we provide a robust relaxation to the problem of downlink transmit beamforming towards a single multicast group, which was considered in [6]. The key difference here is that full channel state information (CSI) is no longer available; instead, the channel vectors are assumed to lie in a ball with known center and radius. Specifically, letting $\tilde{\mathbf{h}}_i := \mathbf{h}_i / \sqrt{c_i \sigma_i^2}$ denote the normalized channel vectors, we assume that $\tilde{\mathbf{h}}_i \in \text{B}_\epsilon(\bar{\mathbf{h}}_i) := \{\tilde{\mathbf{h}}_i | \tilde{\mathbf{h}}_i = \bar{\mathbf{h}}_i + \mathbf{e}, \|\mathbf{e}\| \leq \epsilon\}$. The robust design of the beamformer that minimizes the transmitted power, subject to constraints on the received SNR can be written as

$$\begin{array}{|l|} \hline \mathcal{RB}: \\ \qquad \min_{\mathbf{w} \in \mathbb{C}^N} \ \|\mathbf{w}\|_2^2 \\ \textit{s.t.}: \quad |\mathbf{w}^H \tilde{\mathbf{h}}_i|^2 \geq 1, \ \forall \tilde{\mathbf{h}}_i \in \text{B}_\epsilon(\bar{\mathbf{h}}_i), \ \forall i \in \{1, \ldots, M\}. \\ \hline \end{array}$$

The constraints in problem $\mathcal{RB}$ guarantee that the received signal power in all M users will be larger than unity in the *worst case*, i.e. for the particular channel vector $\tilde{\mathbf{h}}_i$ that corresponds to the smallest value of $|\mathbf{w}^H \tilde{\mathbf{h}}_i|^2$. Each one of these constraints is equivalent to the semi-infinite nonconvex constraint

$$|\mathbf{w}^H \tilde{\mathbf{h}}_i| \geq 1, \ \forall \tilde{\mathbf{h}}_i \in \text{B}_\epsilon(\bar{\mathbf{h}}_i), \quad (4)$$

which admits a convex (SOC) reformulation, as it was shown in [8]. First note that equation (4) can be equivalently written as

$$\min_{\tilde{\mathbf{h}}_i \in \text{B}_\epsilon(\bar{\mathbf{h}}_i)} |\mathbf{w}^H \tilde{\mathbf{h}}_i| \geq 1. \quad (5)$$

Under the natural constraint $|\mathbf{w}^H \bar{\mathbf{h}}_i| \geq \epsilon \|\mathbf{w}\|_2$, it can be shown [8] that

$$\min_{\tilde{\mathbf{h}}_i \in \text{B}_\epsilon(\bar{\mathbf{h}}_i)} |\mathbf{w}^H \tilde{\mathbf{h}}_i| = |\mathbf{w}^H \bar{\mathbf{h}}_i| - \epsilon \|\mathbf{w}\|_2, \quad (6)$$

and we can recast equation (5) as

$$|\mathbf{w}^H \bar{\mathbf{h}}_i| - \epsilon \|\mathbf{w}\|_2 \geq 1 \Leftrightarrow |\mathbf{w}^H \bar{\mathbf{h}}_i| \geq 1 + \epsilon \|\mathbf{w}\|_2. \quad (7)$$

The robust beamforming problem $\mathcal{RB}$ is thus equivalently written as

$$\begin{array}{|l|} \hline \mathcal{RB}': \\ \qquad \min_{\mathbf{w} \in \mathbb{C}^N} \ \|\mathbf{w}\|_2^2 \\ \textit{s.t.}: \quad |\mathbf{w}^H \tilde{\mathbf{h}}_i| \geq 1 + \epsilon \|\mathbf{w}\|_2, \ \forall i \in \{1, \ldots, M\}. \\ \hline \end{array}$$

Let us also consider the corresponding original non-robust beamforming (ONRB) problem:

$$\min_{\mathbf{w}\in\mathbb{C}^N} \|\mathbf{w}\|_2^2$$

$$s.t.: \quad |\mathbf{w}^H\tilde{\mathbf{h}}_i| \geq 1, \quad \forall i \in \{1,\ldots,M\}.$$

Our main result in this section is the following:

Claim 1 *Let $\mathbf{w}'$ be an exact solution of $\mathcal{RB}'$. Then $\mathbf{w}'/(1+\epsilon\|\mathbf{w}'\|)$ is an exact solution of ONRB. Conversely, if $\mathbf{w}_o$ is an exact solution of ONRB, then $\mathbf{w}_o/(1-\epsilon\|\mathbf{w}_o\|)$ is an exact solution of $\mathcal{RB}'$.*

Proof: Forward: The proof is based on two Lemmas. The first is the following *Scaling Lemma*:

Lemma 1 *$\mathbf{w}_o$ is an exact solution of ONRB if and only if $t\mathbf{w}_o$ is an exact solution of*

$$\min_{\mathbf{w}\in\mathbb{C}^N} \|\mathbf{w}\|_2^2$$

$$s.t.: \quad |\mathbf{w}^H\tilde{\mathbf{h}}_i| \geq t, \quad \forall i \in \{1,\ldots,M\}.$$

Proof: $|\mathbf{w}_o^H\tilde{\mathbf{h}}_i| \geq 1 \Longrightarrow |t\mathbf{w}_o^H\tilde{\mathbf{h}}_i| \geq t$. Suppose there exists $\mathbf{w}_1$ with $|\mathbf{w}_1^H\tilde{\mathbf{h}}_i| \geq t$, $\forall i$, and $\|\mathbf{w}_1\|_2^2 < t^2\|\mathbf{w}_o\|_2^2$. Consider $\mathbf{w}_2 := \mathbf{w}_1/t$. It satisfies $|\mathbf{w}_2^H\tilde{\mathbf{h}}_i| \geq 1$, and

$$\|\mathbf{w}_2\|_2^2 = \frac{1}{t^2}\|\mathbf{w}_1\|_2^2 < \frac{1}{t^2}t^2\|\mathbf{w}_o\|_2^2 = \|\mathbf{w}_o\|_2^2, \tag{8}$$

which contradicts optimality of $\mathbf{w}_o$ for ONRB. The converse is obvious. □

Lemma 2 *Let $\mathbf{w}'$ be an exact solution of $\mathcal{RB}'$. Then, $\mathbf{w}'$ is an exact solution of the following non-robust beamforming problem (NRB)*

$$\min_{\mathbf{w}\in\mathbb{C}^N} \|\mathbf{w}\|_2^2$$

$$s.t.: \quad |\mathbf{w}^H\tilde{\mathbf{h}}_i| \geq 1+\epsilon\|\mathbf{w}'\|_2, \quad \forall i \in \{1,\ldots,M\}.$$

Proof: Clearly, $\mathbf{w}'$ is a feasible solution of NRB, since it satisfies the constraints. Suppose there exists $\mathbf{w}''$ that also satisfies the constraints of NRB, but with $\|\mathbf{w}''\|_2^2 < \|\mathbf{w}'\|_2^2$. Then $1+\epsilon\|\mathbf{w}'\|_2 > 1+\epsilon\|\mathbf{w}''\|_2^2$, and thus $\mathbf{w}''$ also satisfies the constraints of problem $\mathcal{RB}'$, with $\|\mathbf{w}''\|_2^2 < \|\mathbf{w}'\|_2^2$. This contradicts optimality of $\mathbf{w}'$ for $\mathcal{RB}'$. □

Now suppose that $\mathbf{w}'$ is an exact solution of $\mathcal{RB}'$. It follows from the last Lemma that it is also an exact solution of NRB. Then, from the *Scaling Lemma*, it follows that $\mathbf{w}'/(1+\epsilon\|\mathbf{w}'\|)$ is an exact solution of ONRB. This completes the forward part of the proof of Claim 1. □

Converse: Let $\mathbf{w}_o$ be a solution of ONRB. Then, according to the Scaling Lemma

$$\mathbf{w}' = \frac{\mathbf{w}_o}{1-\epsilon\|\mathbf{w}_o\|_2} \tag{9}$$

is a solution of the modified NRB (MNRB) problem

$$\min_{\mathbf{w}\in\mathbb{C}^N} \|\mathbf{w}\|_2^2$$

$$s.t.: \quad |\mathbf{w}^H\tilde{\mathbf{h}}_i| \geq \frac{1}{1-\epsilon\|\mathbf{w}_o\|_2}, \quad \forall i \in \{1,\ldots,M\}.$$

We will show that $\mathbf{w}'$ is also a solution of $\mathcal{RB}'$. Since $\mathbf{w}'$ is a solution of MNRB, it follows that

$$|\mathbf{w}'^H\tilde{\mathbf{h}}_i| \geq \frac{1}{1-\epsilon\|\mathbf{w}_o\|_2}. \tag{10}$$

However, from (9), it follows (provided that $1-\epsilon\|\mathbf{w}_o\|_2 \geq 0$, i.e., $\epsilon \leq \frac{1}{\|\mathbf{w}_o\|_2}$) that

$$\|\mathbf{w}'\|_2 = \frac{\|\mathbf{w}_o\|_2}{1-\epsilon\|\mathbf{w}_o\|_2} \Leftrightarrow \|\mathbf{w}_o\|_2 = \frac{\|\mathbf{w}'\|_2}{1+\epsilon\|\mathbf{w}'\|_2}.$$

Hence

$$\frac{1}{1-\epsilon\|\mathbf{w}_o\|_2} = \frac{1}{1-\frac{\epsilon\|\mathbf{w}'\|_2}{1+\epsilon\|\mathbf{w}'\|_2}} = 1+\epsilon\|\mathbf{w}'\|_2, \tag{11}$$

so $\mathbf{w}'$ indeed satisfies the constraints of $\mathcal{RB}'$. Suppose there exists $\mathbf{w}''$, such that $\|\mathbf{w}''\|_2 < \|\mathbf{w}'\|_2$ which also satisfies the constraints of $\mathcal{RB}'$. From the forward proof it follows that $\frac{\mathbf{w}''}{1+\epsilon\|\mathbf{w}''\|_2}$ satisfies the constraints of ONRB, with norm $\frac{\|\mathbf{w}''\|_2}{1+\epsilon\|\mathbf{w}''\|_2}$. On the other hand, $\mathbf{w}_o$ in (9) is an exact solution of ONRB, and $\|\mathbf{w}'\|_2 = \frac{\|\mathbf{w}_o\|_2}{1-\epsilon\|\mathbf{w}_o\|_2}$ yielding $\|\mathbf{w}_o\|_2 = \frac{\|\mathbf{w}'\|_2}{1+\epsilon\|\mathbf{w}'\|_2}$. But $\frac{x}{1+x}$ is monotone increasing in $x > 0$. Therefore, $\|\mathbf{w}''\| < \|\mathbf{w}'\|$ implies that

$$\frac{\|\mathbf{w}''\|_2}{1+\epsilon\|\mathbf{w}''\|_2} < \frac{\|\mathbf{w}'\|_2}{1+\epsilon\|\mathbf{w}'\|_2} = \|\mathbf{w}_o\|_2, \tag{12}$$

which contradicts optimality of $\mathbf{w}_o$ for ONRB. Thus, the proof of Claim 1 is complete. □

Claim 1 implies that we can derive an exact solution of the robust beamforming problem $\mathcal{RB}'$ by a simple scaling of a solution to ONRB. Since both problems are NP-hard in general, in practice this translates to the following algorithm:

1. Compute a good feasible solution $\mathbf{w}_o$ for ONRB using the SDP relaxation approach in [6].
2. A good feasible solution of $\mathcal{RB}'$ is then $\mathbf{w}_o/(1-\epsilon\|\mathbf{w}_o\|_2)$.

Letting c_o and c' denote the norms of the optimal solutions of ONRB and $\mathcal{RB}'$, respectively, we also have

$$c_o = \frac{c'}{1+\epsilon c'} \Leftrightarrow c' = \frac{c_o}{1-\epsilon c_o}. \tag{13}$$

Claim 1 further suggests that if we set $\epsilon > 1/\|\mathbf{w}_o\|_2$, then the robust problem would be infeasible.

4. EXACT ROBUST SOLUTION IN THE SINGLE-GROUP VANDERMONDE CASE

Let us consider again the case when the steering vectors are Vandermonde. Then, the single-group ($G = 1$) version of problem $\mathcal{V}$ can be written as

$$\begin{array}{ll} \mathcal{V}1: & \\ & \min_{\mathbf{r}\in\mathbb{R}\times\mathbb{C}^{N-1}} \mathbf{e}_1^T\mathbf{r} \\ s.t.: & \mathrm{Re}[\mathbf{h}_i^H\tilde{\mathbf{I}}\mathbf{r}] \geq c_i\sigma_i^2, \ \forall\, i \in \{1,\ldots,M\}, \\ & r_\ell = \mathrm{trace}(\mathbf{E}^\ell\mathbf{Y}), \ \forall \ell \in \{0,\ldots,N-1\}, \\ & \mathbf{Y} \succeq \mathbf{0}. \end{array}$$

where $\mathbf{e}_1$ is the first column of the $N \times N$ identity matrix,

$$r_\ell = \sum_{m=1}^{N-\ell} w_m^* w_{m+\ell}, \quad \forall \ell \in \{0, \ldots, N-1\}, \tag{14}$$

$$\mathbf{r} = [r_0 \; r_1 \; \cdots r_{N-1}]^T \in \mathbb{R} \times \mathbb{C}^{N-1}, \tag{15}$$

and

$$\tilde{\mathbf{I}} = \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & 2\mathbf{I}_{N-1} \end{bmatrix} \in \mathbb{R}^N. \tag{16}$$

A robust extension of the problem $\mathcal{V}1$ would be to ask that the SNR constraints are still met, when the angles $\{\theta_i\}_{i=1}^M$ are not known exactly, but allowing an estimation error up to Δ, i.e., they are assumed to lie within the intervals $\theta_i \in [\bar{\theta}_i - \Delta, \bar{\theta}_i + \Delta]$. In such scenario, the SNR constraints are defined as

$$\mathrm{Re}[\mathbf{h}_i^H \tilde{\mathbf{I}} \mathbf{r}] \geq c_i \sigma_i^2, \; \forall \, i \in \{1, \ldots, M\}, \; \forall \theta_i \in [\bar{\theta}_i - \Delta, \bar{\theta}_i + \Delta]. \tag{17}$$

An interpretation of these constraints is that they require (the real part of) certain trigonometric polynomials to be nonnegative over a segment of the unit circle. As it is shown in [4], constraints of this form can be equivalently reformulated to the LMI constraints

$$\tilde{\mathbf{I}} \mathbf{r} - (c_i \sigma_i^2 + j\xi_i)\mathbf{e}_1 = \mathbf{L}^*(\mathbf{X}_i) + \mathbf{\Lambda}^*(\mathbf{Z}_i; \bar{\theta}_i - \Delta, \bar{\theta}_i + \Delta), \tag{18}$$

$\forall \, i \in \{1, \ldots, M\}$, where $\mathbf{X}_i \in \mathbb{C}^{N \times N} \succeq \mathbf{0}$, $\mathbf{Z}_i \in \mathbb{C}^{(N-1)\times(N-1)} \succeq \mathbf{0}$, $\xi_i \in \mathbb{R}$ is unconstrained, and the linear operators $\mathbf{L}^*$ and $\mathbf{\Lambda}^*$ are defined by equations (35) and (36)(along with (16)) in [4], respectively. Hence, the problem encountered in this section is an SDP problem, since it consists of a linear cost, MN linear equality constraints and $2M$ positive semidefinite constraints.

5. CONCLUSIONS

Whereas multi-group multicast transmit beamforming under SINR constraints is NP-hard in general [5, 6], we have shown that, in the special case of Vandermonde steering vectors it is in fact a semidefinite problem, which can be efficiently solved. We have also considered robust beamforming solutions under channel uncertainty for the case of a single multicast group. For general steering vectors, we have shown that exact solutions of the robust and non-robust versions of the problem are related via a simple one-to-one scaling transformation. Since both problems are NP-hard, this suggests an algorithm to generate a quasi-optimal solution for one given a quasi-optimal solution for the other. In the important special case of Vandermonde steering vectors, we have shown that the robust version of the problem is convex as well. This robust solution can be extended to the multi-group Vandermonde case.

6. REFERENCES

[1] B. Alkire, and L. Vandenberghe, "Convex optimization problems involving finite autocorrelation sequences," in *Mathematical Programming, Series A*, vol. 93, no. 3, pp. 331–359, 2002.

[2] M. Bengtsson, and B. Ottersten, "Optimal and suboptimal transmit beamforming," ch. 18 in *Handbook of Antennas in Wireless Communications*, L. C. Godara, Ed., CRC Press, Aug. 2001.

[3] S. Boyd, and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[4] T.N. Davidson, Z.-Q. Luo, and J.F. Sturm, "Linear Matrix Inequality Formulation of Spectral Mask Constraints With Applications to FIR Filter Design," *IEEE Trans. on Signal Processing*, vol. 50, no. 11, Nov. 2002.

[5] E. Karipidis, N.D. Sidiropoulos, and Z.-Q. Luo, "Transmit Beamforming to Multiple Co-channel Multicast Groups," in *IEEE CAMSAP 2005*, to appear.

[6] N.D. Sidiropoulos, T.N. Davidson, and Z.-Q. Luo, "Transmit Beamforming for Physical Layer Multicasting", *IEEE Trans. on Signal Processing*, to appear; see also *Proc. IEEE SAM 2004*.

[7] J.F. Sturm, "Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones," *Optimization Methods and Software*, vol. 11-12, pp. 625–653, 1999.

[8] S.A. Voroboyov, A.B. Gershman, Z.-Q. Luo, "Robust Adaptive Beamforming Using Worst-Case Performance Optimization: A Solution to the Signal Mismatch Problem," *IEEE Trans. on Signal Processing*, vol. 51, no. 2, pp. 313–324, Feb. 2003.

[9] S.-P.Wu, S. Boyd, and L. Vandenberghe, "FIR filter design via spectral factorization and convex optimization," ch. 5 in *Applied and Computational Control, Signals and Circuits,* B. Datta, Ed., Boston, MA: Birkhauser, vol. 1, pp. 215–245, 1998.

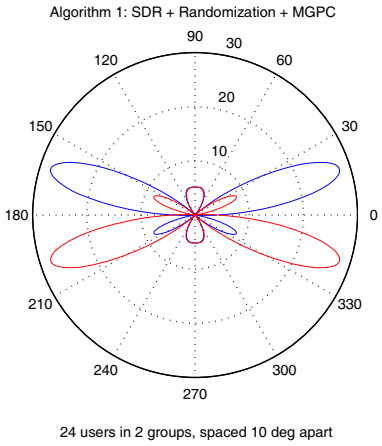


Fig. 1. SDP Relaxation + Randomization result for ULA, $N = 4$, $M = 2 \times 12$, $SINR = 10$dB

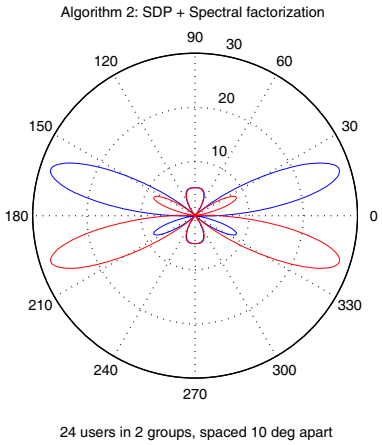


Fig. 2. Exact SDP + Spectral Factorization result for ULA, $N = 4$, $M = 2 \times 12$, $SINR = 10$dB

TRANSMIT BEAMFORMING TO MULTIPLE CO-CHANNEL MULTICAST GROUPS

Eleftherios Karipidis, Nicholas D. Sidiropoulos **

Dept. of ECE, Tech. Univ. of Crete
73100 Chania - Crete, Greece
(karipidis,nikos)@telecom.tuc.gr

Zhi-Quan Luo

Dept. of ECE, Univ. of Minnesota
Minneapolis, MN 55455, U.S.A.
luozq@ece.umn.edu

ABSTRACT

The problem of transmit beamforming to multiple co-channel multicast groups is considered, from the viewpoint of guaranteing a prescribed minimum signal-to-interference-plus-noise-ratio (SINR) at each receiver. The problem is a multicast generalization of the SINR-constrained multiuser downlink beamforming problem: the difference is that each transmitted stream is directed to multiple receivers, each with its own channel. Such generalization is relevant and timely, e.g., in the context of 802.16 wireless networks. Based on earlier results for a single multicast group, the joint problem is easily shown to be NP-hard, a fact that motivates the pursuit of quasi-optimal computationally efficient solutions. It is shown that Lagrangian relaxation coupled with a randomization / co-channel multicast power control loop yields a computationally efficient high-quality approximate solution. For a significant fraction of problem instances, the solutions generated this way are exactly optimal. Carefully designed and extensive simulation results are presented to support the main findings.

1. DATA MODEL AND PROBLEM STATEMENT

Consider a wireless scenario incorporating a single transmitter with N antenna elements and M receivers, each with a single antenna. Let $\mathbf{h}_i$ denote the $N \times 1$ complex vector that models the propagation loss and phase shift of the frequency-flat quasi-static channel from each transmit antenna to the receive antenna of user $i \in \{1, \ldots, M\}$. Let there be a total of $1 \leq G \leq M$ multicast groups, $\{\mathcal{G}_1, \ldots, \mathcal{G}_G\}$, where $\mathcal{G}_k$ contains the indices of receivers participating in multicast group k, and $k \in \{1, \ldots, G\}$. Each receiver listens to a single multicast; thus $\mathcal{G}_k \cap \mathcal{G}_l = \emptyset$, $l \neq k$, $\cup_k \mathcal{G}_k = \{1, \ldots, M\}$, and, denoting $G_k := |\mathcal{G}_k|$, $\sum_{k=1}^{G} G_k = M$.

Let $\mathbf{w}_k^H$ denote the beamforming weight vector applied to the N transmitting antenna elements to generate the spatial channel for transmitting to group k. Then the signal transmitted by the antenna array is equal to $\sum_{k=1}^{G} \mathbf{w}_k^H s_k(t)$, where $s_k(t)$ is the temporal information-bearing signal directed to receivers in multicast group k. Note that the above setup includes the case of *broadcasting* (a *single* multicast group, $G = 1$) [6], as well as the case of individual information transmission to each receiver ($G = M$) by means of spatial multiplexing (see, e.g., [1]). If each $s_k(t)$ is zero-mean white with unit variance, and the waveforms $\{s_k(t)\}_{k=1}^{G}$ are mutually uncorrelated, then the total power radiated by the transmitting antenna array is equal to $\sum_{k=1}^{G} ||\mathbf{w}_k||_2^2$.

*Supported in part by the U.S. ARO under ERO Contract N62558-03-C-0012, the E.U. under FP6 U-BROAD STREP # 506790

The joint design of transmit beamformers can then be posed as the problem of minimizing the total radiated power subject to meeting prescribed SINR constraints c_i at each of the M receivers

$$\mathcal{I}: \quad \min_{\{\mathbf{w}_k \in \mathbb{C}^N\}_{k=1}^{G}} \sum_{k=1}^{G} \|\mathbf{w}_k\|_2^2$$
$$\text{s.t.:} \quad \frac{|\mathbf{w}_k^H \mathbf{h}_i|^2}{\sum_{l \neq k} |\mathbf{w}_l^H \mathbf{h}_i|^2 + \sigma_i^2} \geq c_i, \ \forall i \in \mathcal{G}_k, \ \forall k \in \{1, \ldots, G\}.$$

Problem $\mathcal{I}$ contains the associated broadcasting problem as a special case; from this and [6], it immediately follows that

Claim 1 *Problem $\mathcal{I}$ is NP-hard.*

This motivates (cf. [4]) the pursuit of sensible approximate solutions to problem $\mathcal{I}$.

2. RELAXATION

Towards this end, define $\mathbf{Q}_i := \mathbf{h}_i \mathbf{h}_i^H$ and $\mathbf{X}_k := \mathbf{w}_k \mathbf{w}_k^H$, and note that $|\mathbf{w}_k^H \mathbf{h}_i|^2 = \mathbf{h}_i^H \mathbf{w}_k \mathbf{w}_k^H \mathbf{h}_i = \mathrm{trace}(\mathbf{h}_i^H \mathbf{w}_k \mathbf{w}_k^H \mathbf{h}_i) = \mathrm{trace}(\mathbf{h}_i \mathbf{h}_i^H \mathbf{w}_k \mathbf{w}_k^H) = \mathrm{trace}(\mathbf{Q}_i \mathbf{X}_k)$. Then, problem $\mathcal{I}$ can be *equivalently* reformulated as

$$\min_{\{\mathbf{X}_k \in \mathbb{C}^{N \times N}\}_{k=1}^{G}} \sum_{k=1}^{G} \mathrm{trace}(\mathbf{X}_k)$$
$$\text{s.t.:} \quad \mathrm{trace}(\mathbf{Q}_i \mathbf{X}_k) \geq c_i \sum_{l \neq k} \mathrm{trace}(\mathbf{Q}_i \mathbf{X}_l) + c_i \sigma_i^2,$$
$$\forall i \in \mathcal{G}_k, \ \forall k \in \{1, \ldots, G\},$$
$$\mathbf{X}_k \succeq \mathbf{0}, \ \forall k \in \{1, \ldots, G\},$$
$$\mathrm{rank}(\mathbf{X}_k) = 1, \ \forall k \in \{1, \ldots, G\},$$

where the fact that the terms in the denominator are all non-negative has also been taken into account. Dropping the rank-one constraints, we arrive at the following relaxation of problem $\mathcal{I}$

$$\mathcal{R}: \quad \min_{\{\mathbf{X}_k \in \mathbb{C}^{N \times N}\}_{k=1}^{G}, \ \{s_i \in \mathbb{R}\}_{i=1}^{M}} \sum_{k=1}^{G} \mathrm{trace}(\mathbf{X}_k)$$
$$\text{s.t.:} \quad \mathrm{trace}(\mathbf{Q}_i \mathbf{X}_k) - c_i \sum_{l \neq k} \mathrm{trace}(\mathbf{Q}_i \mathbf{X}_l) - s_i = c_i \sigma_i^2,$$
$$\forall i \in \mathcal{G}_k, \ \forall k \in \{1, \ldots, G\},$$
$$s_i \geq 0, \ \forall i \in \{1, \ldots, M\},$$
$$\mathbf{X}_k \succeq \mathbf{0}, \ \forall k \in \{1, \ldots, G\},$$

where M non-negative real "slack" variables s_i have been introduced, in order to convert the inequality constraints to equality constraints, plus non-negativity constraints. Problem $\mathcal{R}$ is a *Semi-Definite Program* (SDP), expressed in the primal standard form used by SDP solvers, such as SeDuMi [7]. SeDuMi uses interior point methods to solve efficiently such SDP problems, at a complexity cost that is at most $O((GN^2+M)^{3.5})$, and usually much less.

3. OBTAINING AN APPROXIMATE SOLUTION TO PROBLEM $\mathcal{I}$

Problem $\mathcal{I}$ may not admit a feasible solution (counter-examples may be easily constructed), but if it does, the aforementioned approach will yield a solution to problem $\mathcal{R}$. Due to relaxation, this solution will not, in general, consist of rank-one blocks. In order to obtain a high-quality approximate solution of problem $\mathcal{I}$, the concept of *randomization* can be employed to generate candidate beamforming vectors in the span of the respective transmit covariance matrices; see, for example, [6]. The main difference relative to the simpler broadcast case ($G=1$) considered in [6], is that here we cannot simply "scale up" the candidate beamforming vectors generated during randomization to satisfy the hard constraints of problem $\mathcal{I}$. The reason is that, in contrast to [6], we herein deal with an interference scenario, and boosting one group's beamforming vector also increases interference to nodes in other groups. Whether it is feasible to satisfy the constraints for a given set of candidate beamforming vectors is also an issue here. Towards resolving this situation, let $a_{k,i} := |\mathbf{w}_k^H \mathbf{h}_i|^2$ denote the signal power received at receiver i from the stream directed towards users in multicast group k. Let $\beta_k := ||\mathbf{w}_k||^2$, and p_k denote the power boost factor for multicast group k. Then the following *Multi-Group Power Control* ($\mathcal{MGPC}$) problem emerges in converting candidate beamforming vectors to a candidate solution of problem $\mathcal{I}$

$$
\begin{array}{ll}
\mathcal{MGPC}: & \\
 & \min\limits_{\{p_k\in\mathbb{R}\}_{k=1}^{G}} \sum\limits_{k=1}^{G} \beta_k p_k \\
s.t.: & \frac{p_k a_{k,i}}{\sum_{l\neq k} p_l a_{l,i}+\sigma_i^2} \geq c_i, \\
 & \forall i \in \mathcal{G}_k,\ \forall k \in \{1,\ldots,G\}, \\
 & p_k \geq 0,\ \forall k \in \{1,\ldots,G\}.
\end{array}
$$

As in Section 2, taking advantage of the fact that the terms in the denominator are all non-negative and introducing M non-negative real "slack" variables s_i, problem $\mathcal{MGPC}$ can be reformulated as

$$
\begin{array}{ll}
\mathcal{MGPC}: & \\
 & \min\limits_{\{p_k\in\mathbb{R}\}_{k=1}^{G},\ \{s_i\in\mathbb{R}\}_{i=1}^{M}} \sum\limits_{k=1}^{G} \beta_k p_k \\
s.t.: & p_k a_{k,i} - c_i \sum\limits_{l\neq k} p_l a_{l,i} - s_i = c_i \sigma_i^2, \\
 & \forall i \in \mathcal{G}_k,\ \forall k \in \{1,\ldots,G\}, \\
 & p_k \geq 0,\ \forall k \in \{1,\ldots,G\}. \\
 & s_i \geq 0,\ \forall i \in \{1,\ldots,M\},
\end{array}
$$

Problem $\mathcal{MGPC}$ is a *Linear Program* (LP), since the cost function and all constraints are linear. SeDuMi can be used again to solve it efficiently. Note that SeDuMi will also yield an infeasibility certificate in case the $\mathcal{MGPC}$ problem is not solvable for a particular beamforming configuration, which is nice.

For $G=M$ (independent information transmission to each receiver), problem $\mathcal{R}$ is equivalent to and not a relaxation of $\mathcal{I}$, see [1], and problem $\mathcal{MGPC}$ reduces to the well-known multiuser downlink power control problem, which can be solved using simpler means (e.g., [3]): matrix inversion, but also iterative descent algorithms. In this special case, (in)feasibility can be determined from the spectral radius of a certain "connectivity" matrix. Similar simplifications for the general instance of $\mathcal{MGPC}$ are perhaps possible, but appear highly non-trivial. At any rate, LP routines are very efficient.

The overall algorithm for obtaining an approximate solution to problem $\mathcal{I}$ can thus be summarized as follows:

1. Relaxation: Solve problem $\mathcal{R}$, using SDP. Denote the solution $\{\mathbf{X}_k\}_{k=1}^{G}$.
2. Randomization / Scaling Loop: For each k, generate a vector in the span of $\mathbf{X}_k$, using the Gaussian randomization technique (randC) in [6]. If, for some k, $\mathrm{rank}(\mathbf{X}_k)=1$, then use the principal component instead. Next, feed the resulting set of candidate beamforming vectors $\{\mathbf{w}_k\}_{k=1}^{G}$ into problem $\mathcal{MGPC}$ and solve it using LP. If the particular instance of $\mathcal{MGPC}$ is infeasible, discard the proposed set of candidate beamforming vectors; else, see if it yields smaller $\mathcal{MGPC}$ objective than previously checked candidates. If so, record solution and associated objective value.

The *quality* of approximate solutions to problem $\mathcal{I}$ generated this way can be checked against the lower bound on transmit power obtained in solving problem $\mathcal{R}$. This bound can be further motivated from a duality perspective, as in [6]; that is, the aforementioned relaxation lower bound is in fact the tightest lower bound on the optimum of problem $\mathcal{I}$ attainable via Lagrangian duality [2]. This follows from arguments in [8] (see also the single-group case in [6]), due to the fact that problem $\mathcal{I}$ is a quadratically constrained quadratic program.

4. SIMULATION RESULTS

The first step of the proposed algorithm consists of a relaxation of the original QoS beamforming problem $\mathcal{I}$ to problem $\mathcal{R}$. The original problem $\mathcal{I}$ may or may not be feasible; if it is, then so is problem $\mathcal{R}$. If $\mathcal{R}$ is infeasible, then so is $\mathcal{I}$. The converse is generally not true; i.e., if $\mathcal{R}$ is feasible, $\mathcal{I}$ need not be feasible. In order to establish feasibility of $\mathcal{I}$ in this case, the randomization - $\mathcal{MGPC}$ loop should yield at least one feasible solution. This is most often the case, as will be verified in the sequel. If the randomization - $\mathcal{MGPC}$ loop fails to return at least one feasible solution, then the (in)feasibility of $\mathcal{I}$ cannot be determined. There is, therefore, a relatively small proportion of problem instances for which (in)feasibility of $\mathcal{I}$ cannot be decided using the proposed approach.

It is evident from the above discussion that feasibility is a key aspect of problem $\mathcal{I}$ and its proposed solution via problem $\mathcal{R}$ and the randomization - $\mathcal{MGPC}$ loop. Feasibility depends on a number of factors; namely, the number of transmit antenna elements N, the number and the populations of the multicast groups, G and

G_k respectively, the channel characteristics $\mathbf{h}_i$, the channel noise variances σ_i^2, and finally the desired receive SINR constraints c_i.

Beyond feasibility, there are two key issues of interest. The first has to do with cases for which the solution to problem $\mathcal{R}$ yields an exact optimum of the original problem $\mathcal{I}$. This happens when the $N \times N$ blocks $\mathbf{X}_k$, $k \in \{1, \cdots, G\}$ turn out all being rank-one. In this case, the associated principal components solve optimally the original problem $\mathcal{I}$, i.e., in such a case $\mathcal{R}$ is not a relaxation after all.[1] The second issue has to do with the quality of the final approximate solution to problem $\mathcal{I}$ in those cases where a feasible solution can be found using the proposed two-step algorithm. As in [6], a practical figure of merit for the quality of the final approximate solution (set of beamforming vectors and power scaling factors) is the ratio of the total transmitted power corresponding to the approximate solution over $\sum_{k=1}^{G} \mathrm{trace}(\mathbf{X}_k)$ - the lower bound generated from the solution of $\mathcal{R}$.

We consider the standard i.i.d. Rayleigh fading model, i.e., the elements of the channel vectors $\mathbf{h}_i$, $\forall i \in \{1, \ldots, M\}$ are i.i.d. circularly symmetric complex Gaussian random variables of variance 1. Tables 1 and 2 summarize the results obtained using the proposed algorithm for 300 Monte-Carlo runs[2] and 1000 Gaussian randomization samples each. The simulations are repeated for a variety of choices for N, M (see column 1). The users are considered to be evenly distributed among the multicast groups, i.e., $G_k = M/G$, $\forall k \in \{1, \ldots, G\}$. For each such configuration, the problem is solved for increasing values (in dB, column 2) of the received SINR constraints (same for all users), until problem $\mathcal{R}$ becomes infeasible. The noise variance is set to $\sigma^2 = 1$ for all channels. The percentage of the 300 Monte-Carlo runs for which $\mathcal{R}$ is feasible is shown in column 3. Columns 4 and 5 report the percentage of $\mathcal{R}$ feasible solutions which yield exact solutions to problem $\mathcal{I}$ (i.e., when all $\mathbf{X}_k$'s are rank-one), and for which the ensuing randomization - $\mathcal{MGPC}$ loop yields at least one feasible solution, respectively. Finally, the last column holds the average value of the ratio of transmitted power corresponding to the final approximate solution over the lower bound obtained from the SDR solution.

The $\mathcal{R}$ feasibility percentage, and the percentage of cases where $\mathcal{R}$ is equivalent to $\mathcal{I}$, listed in columns 3 and 4, are also plotted in Figures 1 and 2, versus the requested SINR values, for most of the scenarios under consideration. It is observed that $\mathcal{R}$ is getting more difficult to solve (for increasing values of the SINR constraints) as the number G and/or the population G_k of the multicast groups increases and/or the number N of available transmit antenna elements decreases. In all configurations considered, the higher the target SINR, the less likely it is that problem $\mathcal{R}$ is feasible, which is intuitive. Interestingly though, the percentage of exact solutions to $\mathcal{I}$ generated via $\mathcal{R}$ also increases with target SINR. It seems as if rank-one solutions are more likely when operating close to the infeasibility boundary. Furthermore, if the same number of users is distributed over more multicast groups (thus, the number G_k of users per group drops) the attainable common SINR is reduced, as is perhaps intuitive. On the other hand, when the target SINR is on the relatively low side, optimum solutions are more frequently encountered in this case (e.g. see the case of 12 users distributed in 2, 3, and 4 groups for SINR of 6dB), since it is more likely for the fewer users of any group to be spatially close (the respective probability is approximately $1/G^{G_k}$). Last but not least, the randomization - $\mathcal{MGPC}$ loop yields a feasible solution with a probability higher than 90% in most cases where $\mathcal{R}$ is feasible; this solution entails transmission power that is under two times (3 dB from) the possibly unattainable lower bound, on average.

In some scenarios, $\mathcal{R}$ *consistently* yields an exact solution of $\mathcal{I}$. That is, the $\mathbf{X}_k$ blocks are all consistently rank-one. In this case, no further randomization is needed - the principal components of the extracted blocks are the optimal beamformers. More on this will be included in [5].

[1] It is interesting to find the frequency of occurrence of such an event, whose benefit is twofold: not only the problem is solved optimally, but also at smaller complexity, since the randomization step and the repeated solution of the ensuing $\mathcal{MGPC}$ problem is avoided.

[2] 3000 Monte-Carlo runs were employed in cases where $\mathcal{R}$ was feasible in less than 10% of the 300 problem instances initially considered. This was done to improve the estimation accuracy for quantities conditioned on the feasibility of $\mathcal{R}$.

5. CONCLUSIONS

Transmit beamformer design was considered in the context of co-channel multicast transmission to multiple groups of users. The problem is a generalization of downlink transmit beamforming of independent information streams to individual users ([1] and references therein); and the single-group multicast beamforming in [6]. Using [6], the general instance of the problem is easily shown to be NP-hard. A two-step approach comprising semidefinite relaxation and a randomization - multicast power control loop was proposed and shown to yield high-quality approximate solutions, plus means of testing feasibility, at manageable complexity cost.

6. REFERENCES

[1] M. Bengtsson and B. Ottersten, "Optimal and suboptimal transmit beamforming", ch. 18 in *Handbook of Antennas in Wireless Communications*, L. C. Godara, Ed., CRC Press, Aug. 2001.

[2] S. Boyd, and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004; see also http://www.stanford.edu/∼boyd/cvxbook.html.

[3] F.-R. Farrokhi, K.J.R. Liu, and L. Tassiulas, "Downlink Power Control and Base Station Assignment", *IEEE Communications Letters*, vol. 1, no. 4, pp. 102–104, July 1997.

[4] M.R. Garey, and D.S. Johnson, *Computers and Intractability. A Guide to the Theory of NP-Completeness*, W.H. Freeman and Company, 1979.

[5] E. Karipidis, N.D. Sidiropoulos, Z.-Q. Luo, "Convex Transmit Beamforming for Downlink Multicasting to Multiple Co-channel Groups", submitted to *IEEE ICASSP 2006* (invited).

[6] N.D. Sidiropoulos, T.N. Davidson, and Z.-Q. Luo, "Transmit Beamforming for Physical Layer Multicasting", *IEEE Trans. on Signal Processing*, to appear; see also *Proc. IEEE SAM 2004*.

[7] J.F. Sturm, "Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones", *Optimization Methods and Software*, vol. 11-12, pp. 625–653, 1999; see also http://fewcal.kub.nl/sturm/software/sedumi.html

[8] H. Wolkowicz, "Relaxations of Q2P", Chapter 13.4 in *Handbook of Semidefinite Programming: Theory, Algorithms, and Applications*, H. Wolkowicz, R. Saigal, L. Vandenberghe (Eds.), Kluwer Academic Publishers, 2000.

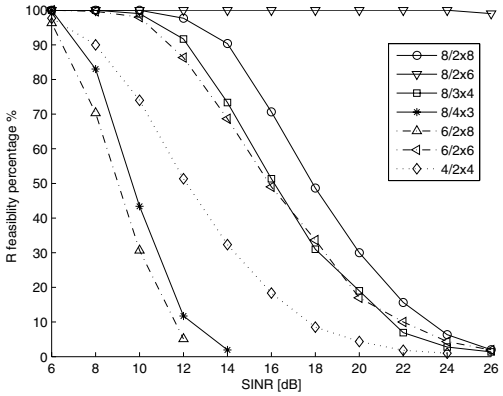


Fig. 1. $\mathcal{R}$ feasibility percentages

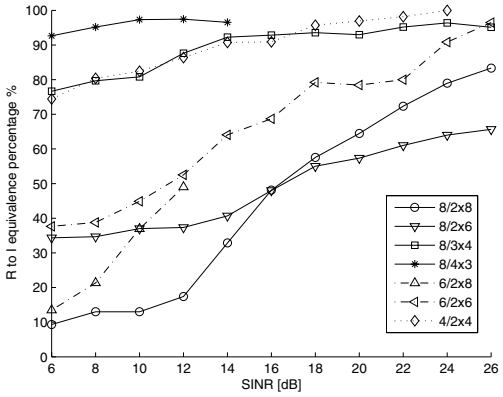


Fig. 2. $\mathcal{R}$ equivalence to $\mathcal{I}$ percentages,

Table 1. MC simulation results for QoS Beamforming (Rayleigh)

$N/G \times G_k$	SINR	$\mathcal{R}$ %	$\mathcal{R} \equiv \mathcal{I}$ %	$\mathcal{MGPC}$ %	mean
$8/2 \times 8$	6	100	9.33	99.67	1.57
$8/2 \times 6$	6	100	34.33	100	1.17
$8/3 \times 4$	6	100	76.67	100	1.04
$8/4 \times 3$	6	100	92.67	99.67	1.01
$6/2 \times 8$	6	96.33	13.49	83.74	2.74
$6/2 \times 6$	6	100	37.67	100	1.39
$6/2 \times 4$	6	100	84	99.67	1.02
$4/2 \times 8$	6	4.57	35.77	68.61	1.86
$4/2 \times 6$	6	46.67	48.57	88.57	1.64
$4/2 \times 4$	6	97.67	74.40	100	1.07
$8/2 \times 8$	8	100	13	99.33	1.85
$8/2 \times 6$	8	100	34.67	100	1.16
$8/3 \times 4$	8	100	79.67	100	1.04
$8/4 \times 3$	8	83	95.18	100	1.01
$6/2 \times 8$	8	70.33	21.33	79.62	2.05
$6/2 \times 6$	8	99.67	38.80	99.67	1.26
$6/2 \times 4$	8	100	83.33	100	1.02
$4/2 \times 6$	8	12.67	60.53	92.11	2.24
$4/2 \times 4$	8	90	80.37	100	1.05

Table 2. MC simulation results for QoS Beamforming (Rayleigh)

$N/G \times G_k$	SINR	$\mathcal{R}$ %	$\mathcal{R} \equiv \mathcal{I}$ %	$\mathcal{MGPC}$ %	mean
$8/2 \times 8$	10	100	13	99.67	1.92
$8/2 \times 6$	10	100	37	99.67	1.17
$8/3 \times 4$	10	99	80.81	99.33	1.04
$8/4 \times 3$	10	43.4	97.31	98.92	1.00
$6/2 \times 8$	10	30.67	36.96	84.78	1.64
$6/2 \times 6$	10	98	44.90	96.94	1.46
$6/2 \times 4$	10	100	82.67	100	1.02
$4/2 \times 6$	10	1.97	74.58	93.22	1.39
$4/2 \times 4$	10	74	82.43	99.10	1.04
$8/2 \times 8$	12	97.67	17.41	96.93	1.75
$8/2 \times 6$	12	100	37.33	100	1.15
$8/3 \times 4$	12	91.67	87.64	100	1.04
$8/4 \times 3$	12	11.73	97.44	99.72	1.00
$6/2 \times 8$	12	5.1	49.02	84.31	1.99
$6/2 \times 6$	12	86.33	52.51	98.07	1.37
$6/2 \times 4$	12	100	86	99	1.02
$4/2 \times 4$	12	51.33	86.36	99.35	1.14
$8/2 \times 8$	14	90.33	32.84	95.94	2.11
$8/2 \times 6$	14	100	40.67	100	1.13
$8/3 \times 4$	14	73.33	92.27	100	1.04
$8/4 \times 3$	14	1.93	96.55	100	1.10
$6/2 \times 6$	14	68.67	64.08	97.09	1.21
$6/2 \times 4$	14	100	87	100	1.01
$4/2 \times 4$	14	32.33	90.72	97.94	1.04
$8/2 \times 8$	16	70.67	48.11	95.28	1.63
$8/2 \times 6$	16	100	48	100	1.11
$8/3 \times 4$	16	51.33	92.86	100	1.03
$6/2 \times 6$	16	49	68.71	92.28	1.15
$6/2 \times 4$	16	100	88.33	99.33	1.01
$4/2 \times 4$	16	18.33	90.91	100	1.01
$8/2 \times 8$	18	48.67	57.53	94.52	1.28
$8/2 \times 6$	18	100	55	100	1.10
$8/3 \times 4$	18	31	93.55	100	1.02
$6/2 \times 6$	18	33.67	79.21	98.02	1.13
$6/2 \times 4$	18	100	87.67	99.33	1.01
$4/2 \times 4$	18	8.53	95.70	98.83	1.02
$8/2 \times 8$	20	30	64.44	97.78	1.29
$8/2 \times 6$	20	100	57.33	100	1.08
$8/3 \times 4$	20	19	92.98	98.25	1.01
$6/2 \times 6$	20	17	78.43	96.08	1.15
$6/2 \times 4$	20	100	89	100	1.01
$4/2 \times 4$	20	4.37	96.95	98.47	1.02
$8/2 \times 8$	22	15.67	72.34	95.74	1.29
$8/2 \times 6$	22	100	61	100	1.08
$8/3 \times 4$	22	6.93	95.19	99.04	1.02
$6/2 \times 6$	22	10	80	96.67	1.37
$6/2 \times 4$	22	100	91	100	1.01
$4/2 \times 4$	22	1.83	98.18	98.18	1.00
$8/2 \times 8$	24	6.33	78.95	94.74	1.39
$8/2 \times 6$	24	100	64	100	1.07
$8/3 \times 4$	24	2.76	96.39	98.80	1.02
$6/2 \times 6$	24	4.37	90.84	96.95	1.12
$6/2 \times 4$	24	100	91	98.33	1.01
$8/2 \times 8$	26	2	83.33	83.33	1.00
$8/2 \times 6$	26	99	65.66	99.63	1.07
$8/3 \times 4$	26	1.37	95.12	100	1.01
$6/2 \times 6$	26	1.9	96.49	100	1.03
$6/2 \times 4$	26	100	91.33	99	1.01
$8/2 \times 6$	28	100	65.67	98.67	1.07
$6/2 \times 4$	28	98.33	91.28	99.33	1.01
$8/2 \times 6$	30	98.67	66.55	99.32	1.07

A TWO-STAGE FASTMAP-MDS APPROACH FOR NODE LOCALIZATION IN SENSOR NETWORKS

Georgios Latsoudas, Nicholas D. Sidiropoulos[†]

Department of Electronic and Computer Engineering
Technical University of Crete
73100 Chania - Crete, Greece

ABSTRACT

Given a set of pairwise distance estimates between nodes, it is often of interest to generate a map of node locations. This is an old problem that has attracted renewed interest in the signal processing community, due to the recent emergence of wireless sensor networks and ad-hoc networks. Sensor maps are useful for estimating the spatial distribution of measured phenomena, as well as for routing purposes. Both centralized and decentralized solutions have been developed, along with ways to cope with missing data, accounting for the reliability of individual measurements, etc. We revisit the basic version of the problem, and propose a two-stage algorithm that combines algebraic initialization and gradient descent. In particular, we borrow an algebraic solution from the database literature and adapt it to the sensor network context, using a specific choice of anchor/pivot nodes. The resulting estimates are fed to a gradient descent iteration. The overall algorithm offers better performance at lower complexity than existing centralized full-connectivity solutions. Also, its performance is relatively close to the corresponding Cramér-Rao bound, especially for small values of range error variance.

1. INTRODUCTION

The problem of node localization from pairwise distance estimates has recently attracted renewed interest in the signal processing and communications literature (e.g., [1, 3, 4]), owing to the recent interest in wireless sensor networks and ad-hoc networks. Given a matrix of pairwise distances (usually estimated using received signal strength measurements and a path loss model), the localization problem asks to determine the *relative* node locations that generate these distances. In other words, one seeks a map of sensor locations with a given (approximate) distance structure. This is a classic problem originating in psychometrics [5, 6], known as *Multi-Dimensional Scaling* (MDS).

There are many MDS flavors and variants; perhaps the single most important version is *metric MDS*. The classic approach to solving MDS is based on computing the principal components of a double-centered version of the distance matrix. This works well (albeit not optimally, due to the double centering), but its complexity is cubic in the number of nodes, and thus does not scale well with network size. A popular alternative to principal component analysis (PCA) is the use of gradient descent or other numerical optimization tools that aim to optimize a *stress function*. The stress function measures the error between the given distances and those reproduced by a given configuration of points. The drawback of gradient descent and related approaches is that they require accurate initialization.

We propose a two-stage MDS algorithm that employs an algebraic initialization procedure followed by gradient descent. The algebraic initialization is based on the Fastmap [2] algorithm, borrowed from the database literature. Fastmap is a linear-complexity mapping tool, which is, however, sensitive to range measurement errors. Due to the fact that distances are invariant to coordinate frame transformations (rotation, reflection, shift), there is a need to employ three so-called *anchor nodes*, whose position is accurately known (e.g., via GPS) in order to fix a desired coordinate frame. Unfortunately, Fastmap is very sensitive to coordinate alignment, because the estimated position of every node (and thus anchor nodes as well) is only based on distances to selected *pivot* nodes - thus there is no averaging. In order to mitigate this problem, we advocate a particular choice of anchor/pivot nodes, placed at the outer edges of the network. This placement bypasses the need for alignment and thus alignment errors, thereby providing a high-quality initialization to the gradient descent. The overall algorithm affords better localization accuracy than PCA-based MDS, at substantially lower complexity cost (quadratic in the number of nodes).

The rest of this paper is structured as follows. In Section 2 we explain in detail the PCA-based MDS algorithm, and its alternative implementations. The Fastmap algorithm is briefly reviewed in Section 3. In Section 4 we describe the proposed Fastmap-MDS algorithm. Simulation results regarding the performance of the above three algorithms, and the Cramer-Rao Lower Bound for the particular localization problem, are shown in Section 5 and conclusions are drawn in section 6.

2. MULTIDIMENSIONAL SCALING

Multidimensional Scaling (MDS) [5, 6],[4] is a method used to depict the spatial structure of distance-like data using the dissimilarity measure among them. It has its origins in psychometrics and psychophysics. MDS starts by presuming that the dissimilarities of each pair of objects stem from data points in an m-dimensional space. In most cases the space in which the data is placed is 2 or 3-dimensional. The algorithm aims to find a geometric representation of the data, such that the distances between data points fit as well as possible to the given dissimilarity information.

We denote the dissimilarity measure (the estimated distances in our case), between objects i and j as d_{ij}. The set of the dissim-

[†] Contact Author. E-mail: nikos@telecom.tuc.gr, Fax: +30-28210-37542. Supported in part by ERO/ARO Contract N62558-03-C-0012.

ilarities forms the matrix $\mathbf{D}$. We also let $\widehat{d}_{ij}$ denote the Euclidean distance between two points $X_i = (x_{i1}, x_{i2}, ..., x_{im})$ and $X_j = (x_{j1}, x_{j2}, ..., x_{jm})$, i.e.

$$\widehat{d}_{ij} = \sqrt{\sum_{k=1}^{m}(x_{ik} - x_{jk})^2}, \quad (1)$$

where m is usually 2 or 3.

In classical metric MDS, we estimate the node coordinates $\mathbf{X}$ by computing the m principal components of a double-centered and element-wise squared version of the matrix $\mathbf{D}$, denoted by $\mathbf{B}$:

$$B = -\frac{1}{2}\mathbf{JPJ}, \quad (2)$$

where $\mathbf{P}$ is the matrix of squared distance measures, and J is the centering operator, ie

$$\mathbf{J} = \mathbf{I} - \mathbf{e}\mathbf{e}^T/N, \quad (3)$$

with N denoting the number of objects (sensor nodes). For an $N \times N$ matrix $\mathbf{D}$ and for m dimensions, it can be shown that

$$-\frac{1}{2}(d_{i_j}^2 - \frac{1}{N}\sum_{j=1}^{N} d_{ij}^2 - \frac{1}{N}\sum_{i=1}^{N} d_{ij}^2 + \frac{1}{N^2}\sum_{j=1}^{N}\sum_{i=1}^{N} d_{ij}^2) = \sum_{k=1}^{m} x_{ik}x_{jk}, \quad (4)$$

thus the estimated node coordinates are given by the m principal eigenvectors of the matrix $\mathbf{B}$, scaled by the square roots of the corresponding eigenvalues. With $\mathbf{U}_r$ denoting the m principal eigenvectors and $\mathbf{V}_r$ diagonal containing the corresponding eigenvalues, $\mathbf{B}_r = \mathbf{U}_r\mathbf{V}_r\mathbf{U}_r$ is an optimal least squares approximation of B, and $\mathbf{X}_r = \mathbf{U}_r\mathbf{V}_r^{1/2}$ is an approximation of the node coordinates in m-dimensional space, up to a common coordinate rotation, reflection, and shift. An alignment procedure is necessary to transform the estimated node locations to a desired frame of reference.

Direct minimization of a suitable *stress function* is an alternative to PCA-based MDS [5]. A common stress function is

$$stress^2 = \sum_{i,j}(\widehat{d}_{ij} - d_{ij})^2. \quad (5)$$

Minimization starts with an initial guess of the node positions (often random), followed by gradient descent iterations. Initialization matters a lot in this context, because the stress function is multimodal. Furthermore, the number of iterations required for convergence depends heavily on the quality of the initialization.

3. FASTMAP

The basic element of Fastmap [2] is the projection of the objects on a properly selected line. This is achieved by selecting two objects O_a, O_b, called *pivots*, and projecting all other objects on the line that passes through them. A pair of pivots is chosen for each of the m dimensions. The coordinates, (i.e. projections on the pivot line) of the objects can be found by employing the *cosine law* [2]. Thus, the first coordinate for object O_i is given by:

$$x_i = \frac{d_{ai}^2 + d_{ab}^2 - d_{bi}^2}{2d_{ab}}, \quad (6)$$

where d_{ij} is the dissimilarity measure between nodes i and j and a, b are the pivot objects. After computing these coordinates for each object O_i, we consider a hyperplane which is orthogonal to the pivot line. We then project the objects on this hyperplane, and repeat the process, this time using

$$d_{ij}^{'2} = d_{ij}^2 - (x_i - x_j)^2, \quad i, j = 1, ..., N. \quad (7)$$

A heuristic method is proposed in [2] for choosing the pivots as far as possible from one another.

In database applications there is no "natural" or preferred coordinate frame of reference, thus the final alignment step is not used, and anchors are not needed. In the context of sensor networks, however, obtaining absolute position estimates is important. Unfortunately, Fastmap is very sensitive to coordinate alignment, because the estimated position of every node (and thus anchor nodes as well) is only based on distances to the chosen pivot nodes - thus there is no averaging. In order to mitigate this problem, we advocate a particular choice of anchor/pivot nodes, placed at the outer edges of the network. In particular, we assume that the sensor nodes are spread over a square, and place the anchor nodes, which will also serve as pivots, at three vertices (see Fig. 1). This placement bypasses the need for alignment and thus alignment errors, thereby providing a high-quality initialization to the gradient descent. Anchors #1 and #2 also serve as pivots for determining the coordinates in the first dimension, while anchors #2 and #3 double as pivots for the second dimension.

4. TWO-STAGE FASTMAP-MDS APPROACH

Fastmap is a fast algebraic method that is rather sensitive to measurement errors, particularly so in the final alignment step. In our context, this sensitivity can be mitigated by proper use of anchor/pivot nodes. The resulting estimates can be used as initialization for gradient descent. Each step of gradient descent costs $\mathcal{O}(N^2)$. Assuming good-enough initialization, only a few gradient descent steps will be needed. This suggests that a substantial complexity reduction relative to PCA is possible. Interestingly, estimation accuracy can be improved as well, as we will see.

The basic steps of the two-stage algorithm are shown in Table 1. Denoting by (x_i, y_i) the estimated position of node i, the partial

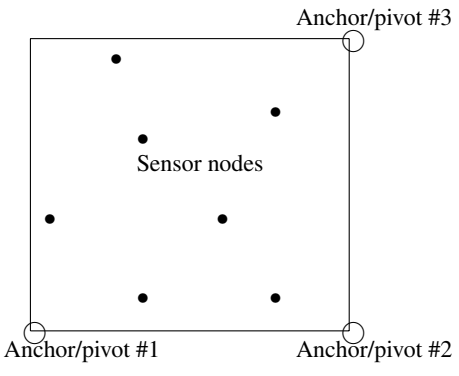


Fig. 1. Anchor-Pivot node placement for using Fastmap in sensor network localization

Table 1. The 2-D Hybrid Fastmap-MDS Algorithm

Input: **D**

1. Run Fastmap using as pivot the anchor nodes, which are placed on the three vertices of the square distribution area. Let X be the vector which contains all the estimated coordinates, which are returned by Fastmap.
2. Determine p, λ
3. For $i = 1$ to p
   begin
   - evaluate $\nabla stress$ at the point X
   - $X = X - \lambda \nabla stress$

   end
4. Output: X

derivative of the stress function in (5) is given by

$$\frac{\partial stress}{\partial x_i} = \sum_{j \neq i} \frac{(\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} - d_{ij})(x_i - x_j)}{\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}}. \tag{8}$$

with a similar expression for the partial derivative with respect to y_i. For simplicity, but also to bound complexity, a fixed number $p = 10$ of gradient descent steps is used in our simulations.

5. RESULTS

We compare the three algorithms described above, in the context of node localization in sensor networks. We consider that the network has full connectivity, that is, we have distance estimates for every pair of nodes. The distance estimates are assumed to contain an error which is proportional to the true distance between the nodes. Thus, we model the distance estimates to be

$$d_{ij} = p_{ij} + p_{ij}\mathcal{N}(0, e_r), \tag{9}$$

where p_{ij} is the true distance between nodes i and j and e_r is the measurement range error variance. Network nodes are considered to be uniformly distributed in a square with area equal to 1, i.e. the x and y coordinates of the sensor nodes are assumed uniformly distributed in $[0, 1]$. We employ the alignment procedure described in [3], in order to find the actual coordinates, and adopt root mean squared error as our estimation performance metric

$$RMSE := \frac{\sum_{i=1}^{N} \sqrt{(x_{ri} - x_{ei})^2 + (y_{ri} - y_{ei})^2}}{N}, \tag{10}$$

where x_{ei}, y_{ei} are the estimated coordinates, and x_{ri}, y_{ri} are the actual sensors coordinates. The baseline MDS algorithm is based on PCA. The complexities of the three algorithms are summarized in Table 2.

In Fig. 2 we show the RMSE performance of the three methods for a sensor network with 80 sensors, as a function of e_r. The corresponding Cramér-Rao Bound (CRB) is also plotted as a benchmark[1]. The parameter λ of the hybrid algorithm is set to 0.01 for this experiment. We observe that Fastmap exhibits poor performance, while PCA-based MDS and the proposed two-stage algorithms have better performance, as expected. Interestingly, the proposed algorithm is not only less complex, but also more accurate than PCA-MDS. This is partially attributed to the fact that PCA-MDS uses double centering, which colors the noise, whereas the proposed algorithm directly minimizes the stress function. We also observe that the Hybrid algorithm is relatively close to the CRB, especially for low range error variance.

[1]CRB derivations are omitted due to space considerations, but will be included in the journal version.

Table 2. Computational complexities

Algorithm	Complexity
Fastmap	$\mathcal{O}(mN)$
Hybrid Fastmap-SVD	$\mathcal{O}(pmN^2)$, $p << N$
MDS with SVD	$\mathcal{O}(N^3)$

In Fig. 3 we show corresponding performance results and the CRB for a network with 200 nodes. The λ parameter is set to 0.005. The estimation accuracies of both PCA-MDS and the proposed two-stage algorithm improve, as expected, relative to the previous case. Fastmap does not benefit, due to the lack of (implicit or explicit) averaging.

We now compare the three algorithms over an additive white noise measurement model, i.e., the measurements have the following form

$$d_{ij} = p_{ij} + \mathcal{N}(0, e_r), \tag{11}$$

where the variance of the measurement error is independent of the distance between the two nodes. The results are shown in Fig. 4 for the case of 80 sensor nodes, and in Fig. 5 for the case of 200 nodes. We observe again that the Hybrid algorithm exhibits better performance than the other two.

6. CONCLUSIONS

We have proposed a two-stage hybrid localization algorithm that offers a better accuracy-complexity trade-off than existing alternatives in the context of sensor networks. The new algorithm employs Fastmap, coupled with judicious selection of anchor nodes that double as pivots, to generate a computationally cheap yet sufficiently accurate initialization for gradient descent. Our simulations indicate that the overall algorithm outperforms PCA-based MDS both in terms of complexity and in terms of estimation accuracy. Future work will include pertinent modifications of this idea that are well-suited for distributed computation and missing data.

7. REFERENCES

[1] J. A. Costa, N. Patwari, A. O. Hero "Distributed Multidimensional Scaling with Adaptive Weighting for Node Localization in Sensor Networks," *ACM Trans. on Sensor Networks*, submitted.

[2] C. Faloutsos, K. Lin, "FastMap: A Fast Algorithm for Indexing, Data-Mining and Visualization of Traditional and Multimedia Datasets," in *Proc. ACM SIGMOD*, vol.24, no.2, p 163-174, 1995.

[3] X. Ji, H. Zha "Sensor Positioning in Wireless Ad-hoc Sensor Networks Using Multidimensional Scaling," in *Proc. Infocom*, pp. 2652-2661, 2004.

[4] Y. Shang, W. Ruml, Y. Zhang, M. Fromherz, "Localization from Connectivity in Sensor Networks," *IEEE Trans. on Parallel and Distr. Systems*, vol. 15, no. 11, pp. 961-974, Nov. 2004.

[5] W.S. Torgerson, "Multidimensional Scaling: I. Theory and method," *Psychometrica*, vol. 17, pp. 401-419, 1952.

[6] W.S. Torgerson, "Multidimensional Scaling of Similarity," *Psychometrica*, vol. 30, pp.379-393, 1965.

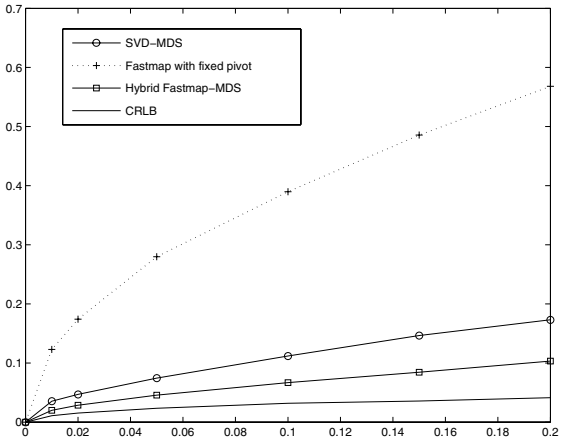


Fig. 2. RMSE performance vs measurement range error variance. N=80, all pairwise distance estimates collected. Measurement error proportional to the actual distance. 100 Monte Carlo runs.

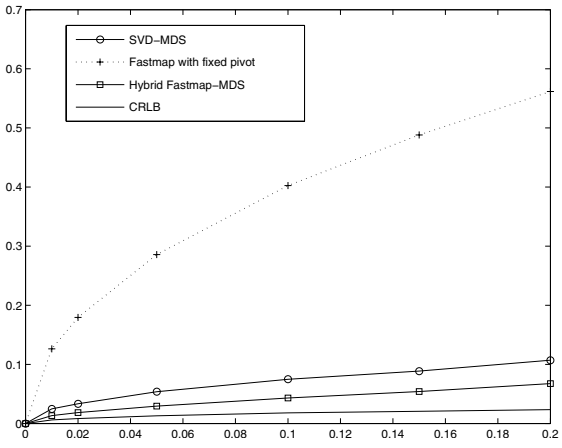


Fig. 3. RMSE performance vs measurement range error variance. N=200 sensor nodes, all pairwise distance estimates collected. Measurement error proportional to the actual distance. 100 Monte Carlo runs.

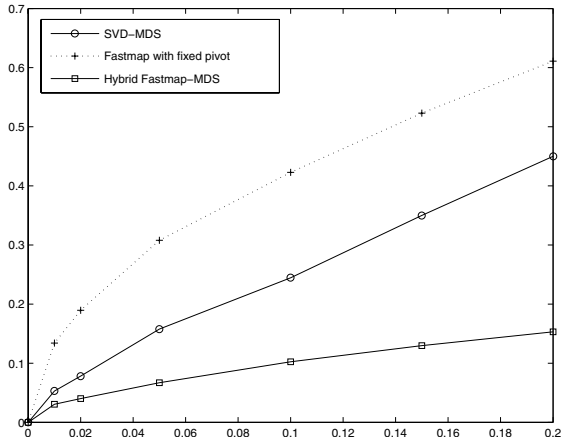


Fig. 4. RMSE performance vs measurement range error variance. N=80, additive noise measurement model, all pairwise distance estimates collected. 100 Monte Carlo runs.

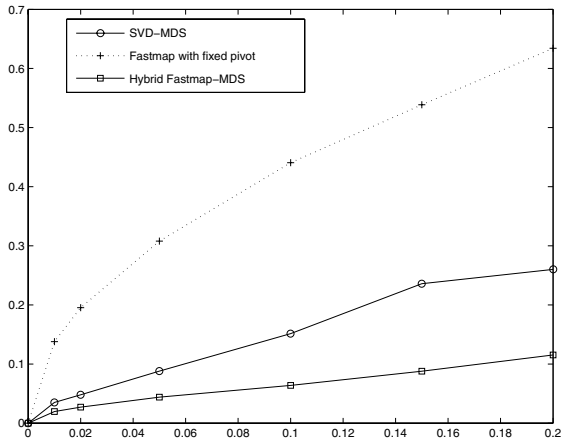


Fig. 5. RMSE performance vs measurement range error variance. N=200 sensor nodes, all pairwise distance estimates collected. Additive noise measurement model. 100 Monte Carlo runs.

BROADCASTING WITH CHANNEL STATE INFORMATION

*N. D. Sidiropoulos**

Dept. of ECE, Tech. Univ. of Crete,
73100 Chania - Crete, Greece
nikos@telecom.tuc.gr

T. N. Davidson

Dept. of ECE, McMaster University,
Hamilton, ON L8S 4K1, Canada
davidson@mcmaster.ca

ABSTRACT

We consider the problem of transmit downlink beamforming for wireless transmission in the context of certain broadcasting or multicasting applications wherein Channel State Information (CSI) is available at the transmitter, and a common message is to be transmitted to the users. Unlike the usual "blind" isotropic broadcasting scenario, the availability of CSI allows transmit optimization. We adopt a minimum transmission power criterion, subject to prescribed minimum received Signal-to-Noise Ratio (SNR) at each of the intended receivers. We also consider a related max-min SNR "fair" problem formulation subject to a transmit power constraint. The basic problem is non-convex and thus difficult to solve; however, we show that a suitable reformulation allows the application of semidefinite relaxation (SDR) techniques. SDR yields a (generally approximate) solution, but in many cases our solution is optimal, and in most cases it is within 3-4 dB from the optimal solution, which is often good enough in our intended applications. While the focus of the paper is on a wireless communication scenario, we also discuss related problems in downstream precoding for broadcasting in digital subscriber line systems.

1. INTRODUCTION

Consider a transmitter that utilizes an antenna array to broadcast (common) information to multiple radio receivers (with a single antenna) within a certain service area. The traditional approach to broadcasting is to radiate transmission power isotropically, or with a fixed directional pattern. While such an approach has the advantage that it is channel independent, it may incur a substantial performance penalty. Furthermore, in modern digital video/audio/data broadcasting and multicasting applications, it is often plausible to assume that the transmitter can acquire channel state information (CSI) for all its intended receivers. This is relatively straightforward in fixed wireless systems and Time-Division-Duplex (TDD) systems, but it can also be accomplished in more general scenarios through the use of beacon signals, periodically transmitted from the broadcasting station (and typically embedded in the transmission). The receiving radios can then feed back their CSI through a feedback channel. For the moment, we shall assume that all channels are perfectly known at the transmitter site. Most of these assumptions can be alleviated, up to a certain extent, at the expense of graceful performance degradation relative to the idealized conditions postulated above.

*Supported in part by the U.S. ARO under ERO Contract N62558-03-C-0012, and the EU under U-BROAD STREP # 506790

The key idea is this: If the transmitter has CSI for all the radios that it intends to broadcast to, then it makes sense to attempt to minimize total transmission power (and thus leakage to neighboring co-channel transmissions), subject to meeting constraints on the received Signal-to-Noise Ratio (SNR) for each individual intended receiver. Note that this is a Quality of Service (QoS) guarantee that directly translates to a guaranteed minimum information rate for each of the receivers. Also note that different receivers may have different SNR requirements, due to differing traffic requirements, and different noise and interference conditions.

Another application of the methodology developed herein can be found in downstream multicast transmission for multi-carrier and single-carrier Digital Subscriber Line (DSL) systems. In this context, (linear) *precoding* of multiple DSL loops in the same binder that wish to subscribe to a common service (e.g., news feed, video-conference, or movie multicast) can be employed to improve quality of service and/or reduce far-end crosstalk (FEXT) interference to other loops in the binder. In cases wherein the Customer-Premise Equipment (CPE) receivers are not physically co-located (as in residential service), or cannot be coordinated (as in legacy CPE systems), multiuser decoding of the downstream transmission is not feasible, while transmit precoding is viable. The most important difference between DSL and the wireless multicast scenario considered so far is that DSL channels are diagonally-dominant. That said, exploitation of the crosstalk coupling to reduce FEXT levels to other loops in the binder offers the potential for considerable gains in the management of mutual interference.

It is interesting to note that, as of today, internet multicasting (using the internet protocol's Multicast Backbone - MBone) is performed at the *network layer*, i.e., via packet-level flooding or spanning-tree access of the participant nodes and any intermediate nodes needed to access the participants. Instead, what we advocate herein is judicious *physical layer multicasting*, that is enabled by i) the availability of multiple transmitting elements; ii) exploiting opportunities for joint beamforming/precoding; and iii) the availability of CSI at the transmitting node or one of its proxies. This is a cross-layer optimization approach that exploits information that is made available at the physical layer to reduce relay retransmissions at the network layer. This provides the potential for congestion relief and considerable Quality of Service (QoS) gains.

Notation: We use lowercase boldface letters to denote column vectors, and uppercase bold letters to denote matrices. $(\cdot)^T$ denotes transpose, while $(\cdot)^H$ denotes Hermitian (conjugate) transpose. Re (Im) extracts the real (respectively, imaginary) part of its argument.

0-7803-8545-4/04/$20.00 ©2004 IEEE

2. DATA MODEL AND PROBLEM STATEMENT

We assume that each radio receiver employs a single receive antenna (and thus a single receiver front-end and downconversion chain), as is appropriate for simplicity and cost considerations in broadcasting applications. Let $\mathbf{h}_i$ denote the $N \times 1$ complex vector modeling propagation loss and phase shift from each of the N transmitting antenna elements to the receiving antenna of user $i \in \{1, \cdots, M\}$. This model assumes that the channels between the transmitter and the receivers are flat in frequency over the bandwidth of the transmitted signal, but, as we will demonstrate below, the principles of our design can be extended to the frequency-selective case in a straightforward manner.

If we let $\mathbf{w}^H$ denote the weight vector applied to the N transmitting antenna elements, then the problem of interest is to minimize the transmitted power (of a white data sequence), subject to the received signal power of user i being larger than a threshold c_i. This problem can be written as

$$\min \|\mathbf{w}\|_2^2$$

$$\textit{subject to: } |\mathbf{w}^H \mathbf{h}_i|^2 \geq c_i, \quad i \in \{1, \cdots, M\}$$

where $\mathbf{w} \in \mathbb{C}^N$. This is a quadratically constrained quadratic programming problem, but unfortunately the constraints are not convex.

2.1. Review of Pertinent Prior Art

The above problem is reminiscent of some closely-related problems. For $M = 1$, the optimum $\mathbf{w}$ is a matched filter. When the channel vectors span a ball or ellipsoid about a "nominal" channel vector (a model that implies a continuum of intended receivers), the problem can be solved *exactly* using second-order cone programming, as shown in [8]. The key observation is that one can convert the infinitely-many non-convex constraints over the ball into a *single* convex constraint, by taking advantage of rotational freedom and the Cauchy-Schwartz inequality to explicitly construct the worst-case channel vector within the said ball. Unfortunately, we are not aware of a corresponding conversion for finitely-many channel vectors (intended receivers).

Another closely-related work is that in [1] (and references therein), which considers the problem of multiuser transmit beamforming for the cellular downlink. The key difference between [1] and our formulation is that the authors of [1] consider the transmission of independent information to each of the downlink users, whereas we focus on the broadcast of common information. The mathematical formulations of these problems are not equivalent. A simple way to see this is to note that in the generic case of our formulation most of the SNR constraints will be inactive at the optimum (i.e., most of the constraints will be over-satisfied). Consider, e.g., the case of two closely-located receivers with different SNR requirements: one of the two associated constraints will be over-satisfied at the optimum. On the other hand, it is proven in [1] that, in the cellular downlink problem, the constraints are always met with equality at the optimum. The important common denominator of our work and [1] is the use of semidefinite programming tools.

3. RELAXATION

Towards solving our problem, we first recast it as follows:

$$\min_{\mathbf{w}} trace(\mathbf{w}\mathbf{w}^H)$$

$$\textit{subject to : } trace(\mathbf{w}\mathbf{w}^H \mathbf{Q}_i) \geq c_i, \quad i \in \{1, \cdots, M\},$$

where we have used the fact that $\mathbf{h}_i^H \mathbf{w}\mathbf{w}^H \mathbf{h}_i = trace(\mathbf{h}_i^H \mathbf{w}\mathbf{w}^H \mathbf{h}_i) = trace(\mathbf{w}\mathbf{w}^H \mathbf{h}_i \mathbf{h}_i^H)$, and $\mathbf{Q}_i := \mathbf{h}_i \mathbf{h}_i^H$. Now consider the following reformulation of the problem:

$$\min_{\mathbf{X} \in \mathbb{C}^{N \times N}} \quad trace(\mathbf{X})$$

$$\textit{subject to : } trace(\mathbf{X}\mathbf{Q}_i) \geq c_i, \quad i \in \{1, \cdots, M\},$$

$$\mathbf{X} \geq 0,$$

$$rank(\mathbf{X}) = 1,$$

where now $\mathbf{X}$ is an $N \times N$ complex matrix, and the inequality $\mathbf{X} \geq \mathbf{0}$ means that the matrix $\mathbf{X}$ is symmetric positive semidefinite. Note that, in the above *equivalent* formulation of our problem, the cost function is linear in $\mathbf{X}$; the trace constraints are linear inequalities in $\mathbf{X}$, and the set of symmetric positive semidefinite matrices is convex; however the rank constraint on $\mathbf{X}$ is not convex. The important observation is that the above problem is in a form suitable for semidefinite relaxation (SDR) (e.g., see [4]). That is, by dropping the rank-one constraint, one obtains the relaxed problem

$$\min_{\mathbf{X} \in \mathbb{C}^{N \times N}} \quad trace(\mathbf{X})$$

$$\textit{subject to : } trace(\mathbf{X}\mathbf{Q}_i) \geq c_i, \quad i \in \{1, \cdots, M\}, \textit{ and } \mathbf{X} \geq 0,$$

which is a semidefinite programming problem (SDP), albeit not yet in standard form. In order to put it in a standard form, we add M non-negative "slack" variables s_i, one for each trace constraint. In this way, we obtain the following formulation

$$\min_{\mathbf{X} \in \mathbb{C}^{N \times N}} \quad vec(\mathbf{I}_N)^T vec(\mathbf{X})$$

$$\textit{subject to: } vec(\mathbf{Q}_i^T)^T vec(\mathbf{X}) - s_i = c_i, \; i \in \{1, \cdots, M\}$$

$$s_i \geq 0, i \in \{1, \cdots, M\}, \textit{ and } \mathbf{X} \geq 0$$

which is now expressed in a standard form used by SDP solvers, such as SeDuMi [6].

SDP problems can be efficiently solved using interior point methods. In particular, the complexity of solving the above program is at most $O((M + N)^{6.5})$, and it is usually much less. SeDuMi [6] is a MATLAB implementation of modern interior point methods for SDP that is particularly efficient for the moderate-sized problems that are encountered in our context. Typical run times for realistic choices of N and M are about 1/10 sec, on a typical desktop computer.

4. ALGORITHM

Due to the relaxation, the matrix $\mathbf{X}_{opt}$ obtained through the SDP will not be rank-one in general. If it is, then its principal component will be the optimal solution to the original problem. If not, then the trace of $\mathbf{X}_{opt}$ is a lower bound on the power needed to satisfy the constraints. This is evident from the fact that we have

removed one of the original problem's constraints. Researchers in optimization have recently developed ways of generating good solutions to the original problem from the solution to the relaxed problem, $\mathbf{X}_{opt}$ [4, 9, 7, 5]. This process is based on *randomization:* using $\mathbf{X}_{opt}$ to generate a set of candidate weight vectors, $\{\mathbf{w}_\ell\}$, from which the "best" solution will be selected. We consider two methods for generating the $\mathbf{w}_\ell$'s, both of which have been designed so that their computational cost is negligible compared to that of computing $\mathbf{X}_{opt}$. (For consistency, the principal component is also included in the set of candidates.) In the first method (randA), we calculate the eigen-decomposition of $\mathbf{X}_{opt} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^H$ and choose $\mathbf{w}_\ell$ such that $\mathbf{w}_\ell = \mathbf{U}\mathbf{\Sigma}^{1/2}\mathbf{e}_\ell$, where $\mathbf{e}_\ell$ is uniformly distributed on the unit sphere. In the second method (randB), inspired by Tseng [7], we choose $\mathbf{w}_\ell$ such that $[\mathbf{w}_\ell]_i = \sqrt{[\mathbf{X}_{opt}]_{ii}}\, e^{j\theta_{\ell,i}}$, where the $\theta_{\ell,i}$ are independent and uniformly distributed on $[0, 2\pi)$. In both cases, $||\mathbf{w}_\ell||_2^2 = \text{trace}(\mathbf{X}_{opt})$, and hence when $\text{rank}(\mathbf{X}_{opt}) > 1$, at least one of the constraints $|\mathbf{w}_\ell^H \mathbf{h}_i|^2 > c_i$ will be violated. However, a feasible weight vector can be found by simply scaling $\mathbf{w}_\ell$ so that all the constraints are satisfied. The "best" of these randomly generated weight vectors is the one that requires the smallest scaling. The overall approach is summarized in Table 1. We point out that we have not yet been able to obtain theoretical *a priori* bounds on the extent of the sub-optimality of solutions generated in this way, but our simulation results are quite encouraging.

5. MAX-MIN FAIR BEAMFORMING

We now switch to an alternative problem that is also of interest. We consider

$$\max_{\mathbf{w}\in\mathbb{C}^N} \min \left\{|\mathbf{w}^H\mathbf{h}_i|^2\right\}_{i=1}^M$$
$$\textit{subject to: } ||\mathbf{w}||_2^2 \le P$$

It is easy to see that the constraint should be met with equality at an optimum, for otherwise $\mathbf{w}$ could be scaled up, thereby improving the objective and contradicting optimality. Thus we can focus on the equality-constrained problem. With a scaling of the optimization variable $\mathbf{w} = \sqrt{P}\tilde{\mathbf{w}}$, the equality-constrained problem can be written as

$$\max_{\tilde{\mathbf{w}}} \min \left\{P|\tilde{\mathbf{w}}^H\mathbf{h}_i|^2\right\}_{i=1}^M$$
$$\textit{subject to: } ||\tilde{\mathbf{w}}||_2^2 = 1.$$

It is clear that the solution to this problem simply scales with P; the solution scales up with $\sqrt{P}$, while the optimum value scales up with P. We can therefore restrict our attention to the problem (dropping the tilde for brevity):

$$\max_{\mathbf{w}} \min \left\{|\mathbf{w}^H\mathbf{h}_i|^2\right\}_{i=1}^M$$
$$\textit{subject to: } ||\mathbf{w}||_2^2 = 1$$

Some discussion is due at this point on the relationship between the two problem formulations: the original QoS formulation that seeks to minimize the total transmit power subject to prescribed lower bounds, c_i, on the received signal powers; and the max-min "fair" formulation seeks to maximize the received signal power of the weakest user subject to an overall transmit power constraint. Suppose that all c_i's are equal to c, and the QoS formulation yields a beamformer $\mathbf{w}_q$ and associated minimum transmit power P_q. Then we can scale the solution of the max-min fair beamformer to power P_q, and this scaled max-min fair solution, denoted $\mathbf{w}_f$, will be an optimal solution to

$$\max_{\mathbf{w}} \min \left\{|\mathbf{w}^H\mathbf{h}_i|^2\right\}_{i=1}^M$$
$$\textit{subject to: } ||\mathbf{w}||_2^2 = P_q$$

As a result, since $\mathbf{w}_q$ already attains $|\mathbf{w}_q^H\mathbf{h}_i|^2 \ge c, \forall i$, it follows that $|\mathbf{w}_f^H\mathbf{h}_i|^2 \ge c, \forall i$. Hence $\mathbf{w}_f$ also satisfies the constraints of the QoS formulation, and at the same power as $\mathbf{w}_q$. It follows that $\mathbf{w}_f$ is equivalent to $\mathbf{w}_q$. This shows that

Claim 1 *The QoS problem formulation and the max-min fair problem formulation are equivalent in the case that all the c_i's are equal.*

When the c_i's are different, however, the two problem formulations generally yield different beamformers. Claim 1 implies an indirect way of solving the max-min fair problem:

Corollary 1 *One way to solve the max-min fair problem is to solve the QoS problem with $c_i = 1, \forall i \in \{1, \cdots, M\}$, then scale the resulting solution to the desired power P.*

6. THE CASE OF FREQUENCY-SELECTIVE MULTIPATH

Although we have focused our attention so far on frequency-flat fading channels, the situation is quite similar for frequency-selective (intersymbol-interference) channels. Let $\mathbf{h}_i^{(\ell)}$ denote the ℓ-th $N \times 1$ vector tap of the baseband-equivalent discrete-time impulse response of the multipath channel between the transmitter antenna array and the (single) receive antenna of receiver-i. Assume that delay spread is limited to L non-zero vector channel taps. Define the channel matrix for the i-th receiver as

$$\mathbf{H}_i := \left[\mathbf{h}_i^{(0)}, \cdots, \mathbf{h}_i^{(L-1)}\right].$$

Beamforming the transmit array with a fixed (time-invariant) $\mathbf{w}^H$ yields a scalar equivalent channel from the viewpoint of the i-th receiver, whose scalar taps are given by

$$\left[\bar{h}_i^{(0)}, \cdots, \bar{h}_i^{(L-1)}\right]^T = \left[\mathbf{w}^H\mathbf{h}_i^{(0)}, \cdots, \mathbf{w}^H\mathbf{h}_i^{(L-1)}\right]^T,$$

or, in vector form,

$$\bar{\mathbf{h}}_i^T = \mathbf{w}^H\mathbf{H}_i.$$

Now, if a Viterbi equalizer is used for sequence estimation at the receiver, then the parameter that determines performance is [3]:

$$||\bar{\mathbf{h}}_i||_2^2 = \mathbf{w}^H\mathbf{H}_i\mathbf{H}_i^H\mathbf{w} =$$
$$trace(\mathbf{w}\mathbf{w}^H\mathbf{H}_i\mathbf{H}_i^H) = trace(\mathbf{w}\mathbf{w}^H\mathbf{Q}_i),$$

where $\mathbf{Q}_i := \mathbf{H}_i\mathbf{H}_i^H$. Therefore, both the QoS and max-min "fair" problems naturally extend to the frequency-selective case. While $\mathbf{Q}_i$ is generally of higher rank than in the flat-fading case, the principles of relaxation can be applied in an analogous manner to generate an approximation of the optimal $\mathbf{w}$.

7. INSIGHTS AFFORDED VIA DUALITY

Let us return to our original problem:

$$\min ||\mathbf{w}||_2^2$$

$$\textit{subject to: } |\mathbf{w}^H \mathbf{h}_i|^2 \geq c_i, \quad i \in \{1, \cdots, M\}$$

We can convert the problem to real-valued form; this yields a $2N \times 1$ vector of real variables, $\mathbf{x} := \left[Re\{\mathbf{w}\}^T \; Im\{\mathbf{w}\}^T\right]^T$, and the $\mathbf{Q}_i$'s are now $2N \times 2N$ symmetric matrices of rank 2: $\mathbf{Q}_i := \mathbf{g}_i \mathbf{g}_i^T + \bar{\mathbf{g}}_i \bar{\mathbf{g}}_i^T$, where $\mathbf{g}_i := \left[Re\{\mathbf{h}_i\}^T \; Im\{\mathbf{h}_i\}^T\right]^T$, and $\bar{\mathbf{g}}_i := \left[Im\{\mathbf{h}_i\}^T - Re\{\mathbf{h}_i\}^T\right]^T$. Then our original problem can be written as:

$\mathcal{P}$:

$$\min \mathbf{x}^T \mathbf{x}$$

$$\text{subject to: } \mathbf{x}^T \mathbf{Q}_i \mathbf{x} \geq c_i, \; i \in \{1, \cdots, M\}.$$

It can be shown that the (Lagrange) dual of problem $\mathcal{P}$ is a Semi-Definite Program (SDP). The dual problem is interesting, because it generates a lower bound on the minimum objective value of the original problem [2]. The dual problem is convex by virtue of its definition. This means that we can solve the dual problem and thus obtain the tightest bound obtainable via duality. This duality-derived bound can be compared to the SDR bound we used earlier. Let $\mathcal{D}(\cdot)$ ($\beta(\cdot)$) denote the dual (respectively, minimum) of a certain minimization problem, and let $\mathcal{R}(\mathcal{P})$ denote the semidefinite relaxation of $\mathcal{P}$, obtained by dropping the associated rank-one constraint. It can be shown that

Claim 2 *$\mathcal{D}(\mathcal{D}(\mathcal{P})) = \mathcal{R}(\mathcal{P})$; and $\beta(\mathcal{R}(\mathcal{P})) = \beta(\mathcal{D}(\mathcal{P}))$. That is, semidefinite relaxation yields the duality bound for $\mathcal{P}$, and the corresponding gap is equal to the duality gap.*

Claim 2 along with claim 1 directly yields the following corollary:

Corollary 2 *Let $\mathcal{F}$ denote the max-min fair problem formulation. Then $\mathcal{D}(\mathcal{D}(\mathcal{F})) = \mathcal{R}(\mathcal{F})$; and $\beta(\mathcal{R}(\mathcal{F})) = \beta(\mathcal{D}(\mathcal{F}))$. Thus, semidefinite relaxation yields the duality bound for $\mathcal{F}$, and the corresponding gap is equal to the duality gap.*

8. SIMULATION RESULTS

Simulation results are presented in Fig. 1 and Tables 2, 3, and 4.

Table 2 summarizes the results obtained using the algorithm in Table 1 with the randA option for randomization. Table 3 summarizes the results obtained using the algorithm in Table 1 and both randA and randB randomizations. In this case, the best of the two solutions (in the sense of minimizing the power boost relative to the lower bound provided by SDR) is selected in each Monte-Carlo (MC) run. The captions are otherwise self-contained. Note that, in many cases, our solutions are within 3-4 dB from the (generally conservative) lower bound on transmit power provided by SDR, and thus are guaranteed to be at most 3-4 dB away from optimal; this is often good enough from an engineering perspective. In several cases the solutions are essentially optimal. This is illustrated in Figure 1, which shows the optimized transmit beam pattern for a particular far-field multicasting scenario using a Uniform Linear antenna Array (ULA); the details of the simulation setup are included in the figure captions for ease of reference.

Table 4 summarizes our simulation results for max-min fair beamforming. Table 4 presents averages for the upper bound on minimum SNR (the optimum attained by SDP without regard to the rank-one constraint), the SDR-attained minimum SNR (after randomization), and the minimum SNR for the case of no beamforming. For the latter, we have used $\mathbf{w} = \frac{1}{\sqrt{N}} \mathbf{1}_{N \times 1}$, which fixes transmit power to 1. The number of post-SDR randomizations was set to $30NM$. (This time a function of N, M.) It is satisfying to note that the SDR solution attains a significant fraction of the (possibly unattainable) upper bound. Furthermore, SDR provides substantial gains over not beamforming at all.

We observe from Tables 2– 4, that as N and/or M increase, the quality of the solution generated by the semidefinite relaxation degrades a little. The reasons for this degradation are under investigation, but possible causes include implementation issues, such as the number of randomizations and the nature of the randomization strategy, and more fundamental issues, such as the potential for a mild degradation of the approximation quality of the method as the problem size grows. (In a related, but distinct, problem the quality of the SDR approximation degrades logarithmically in the problem size [5].)

9. CONCLUSIONS

We have taken a new look at the broadcasting/multicasting problem when channel state information is available at the transmitter. We have formulated the problem of minimizing the transmit power under multiple SNR constraints, and we have shown how its solution can be often well-approximated using semidefinite relaxation tools. We have also considered a max-min fair problem formulation. For both formulations, semidefinite relaxation yields a bound on the degree of suboptimality that is actually equal to the optimum Lagrange dual bound. This justifies, to a certain extent, the approximation introduced by relaxation. Still, it would be nice to analyze the duality gap for the problem at hand, for this would yield *a priori* bounds on the degree of suboptimality introduced by relaxation, as opposed to the *a posteriori* bound that we now have by virtue of Claim 2. For the time being, our simulation results indicate that the degree of suboptimality is often within 3-4 dB, on average, which is acceptable in our intended applications.

There are many interesting refinements and extensions to this work. These include potentially better randomization strategies, robustness issues, and extensions to multiple co-channel multicasting groups. These are the subjects of on-going work, and will be reported elsewhere.

10. REFERENCES

[1] M. Bengtsson and B. Ottersten, "Optimal and suboptimal transmit beamforming", ch. 18 in *Handbook of Antennas in Wireless Communications*, L. C. Godara, Ed., CRC Press, Aug. 2001.

[2] S. Boyd, and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004 (see also http://www.stanford.edu/~boyd/cvxbook.html).

[3] G.D. Forney, "Maximum-Likelihood Sequence Estimation of Digital Sequences in the Presence of Intersymbol Interference", *IEEE Trans. on Information Theory*, 18(3):363-378, May 1972.

[4] W.-K. Ma, T.N. Davidson, K.M. Wong, Z-Q Luo, P.-C. Ching, "Quasi-ML multiuser detection using semi-definite relaxation with application to synchronous CDMA", *IEEE Trans. on Signal Processing*, 50(4):912 -922, Apr. 2002.

[5] A. Nemirovski, C. Roos and T. Terlaky, "On maximization of quadratic form over intersection of ellipsoids with common center", *Math. Program., Ser. A*, 86:463–473, 1999.

[6] J.F. Sturm, "Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones", *Optimization Methods and Software*, 11-12:625-653, 1999. See also http://fewcal.kub.nl/sturm/software/sedumi.html

[7] P. Tseng, "Further results on approximating nonconvex quadratic optimization by semidefinite programming relaxation", *SIAM Journal on Optimization*, 14(1):268-283, July 2003.

[8] S.A. Vorobyov, A.B. Gershman, Z-Q. Luo, "Robust adaptive beamforming using worst-case performance optimization via second-order cone programming", in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing, 2002 (ICASSP2002)*, 3:2901-2904, May 13-17, 2002.

[9] S. Zhang, "Quadratic maximization and semidefinite relaxation", in *Math. Program., Ser. A*, 87:453-465, 2000.

Table 1. Broadcast Beamforming via SDR: Algorithm

- Solve the relaxed problem:

A suitable MATLAB interface for SeDuMi is as follows:

```
% H is N by M, holding the channel vectors;
% constraints is M by 1, holding the Rx power constraints
vecQs = [];
for i=1:M,
Qi = H(:,i)*H(:,i)';
vecQs = [vecQs vec(Qi.')];
end
A=[-eye(M), vecQs.'];
b=constraints;
c=[zeros(M,1); vec(eye(N))];
K.l=M; K.s=N; K.scomplex=1;
[x_opt, y_opt,info]=sedumi(A,b,c,K);
X_opt=mat(x_opt(M+1:end));
```

- Randomization:

Use randA, or randB, as described in Section 4.
It is often preferable to run both and pick the best result.

Table 2. MC simulation results: mean and standard deviation of upper bound on power boost. $\mathbf{H}$ is circularly symmetric complex i.i.d. Gaussian (Rayleigh) of variance 1. randA randomization only. # post-SDR randomizations = 300. The symbol U indicates that Rx power constraints are uniformly distributed random variables in $[0, 1]$, and redrawn for each MC run; 1 means that all Rx power constraints are fixed to 1. # MC-runs = 300.

N/M	mean (U)	std (U)	mean (1)	std (1)
4/8	1.14	0.27	1.30	0.36
4/16	1.63	0.55	1.96	0.62
8/16	2.11	0.65	2.54	0.68
8/32	3.20	0.79	3.77	0.93

Table 3. MC simulation results: mean and standard deviation of upper bound on power boost. Here, the best result from two randomization techniques (randA,randB) is chosen for each MC run. # post-SDR randomizations = 1000. # MC-runs = 1000. The remaining parameters are as in Table 2.

N/M	mean (U)	std (U)	mean (1)	std (1)
4/8	1.07	0.12	1.15	0.17
4/16	1.32	0.26	1.49	0.30
8/16	1.72	0.34	2.06	0.34
8/32	2.51	0.43	2.96	0.44

Table 4. MC simulation results for max-min fair beamforming: averages for upper bound on $\min_i SNR_i$, relaxation-attained $\min_i SNR_i$, and the $\min_i SNR_i$ for the case of no beamforming. The results are averaged over 1000 MC runs. For each MC run, $\mathbf{H}$ is re-drawn from a circularly symmetric complex i.i.d. Gaussian distribution of variance 1. The best result from two randomization techniques (randA,randB) is chosen for each MC run. # post-SDR randomizations = $30NM$. $P = 1$.

N/M	upper bound	SDR	no BMF
4/8	1.05	0.92	0.12
4/16	0.73	0.48	0.06
8/16	1.43	0.72	0.06
8/32	1.07	0.37	0.03

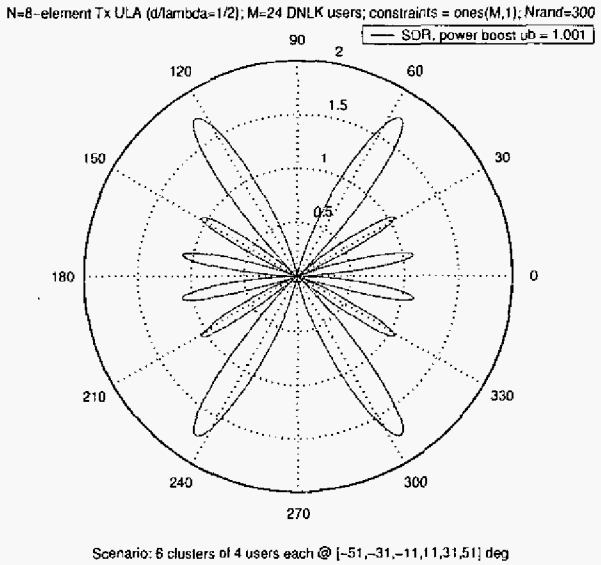


Fig. 1. Broadcast beamforming example using Algorithm in Table 1. N=8-element Tx ULA (d/λ=1/2); M=24 downlink users, in 6 clusters of 4 users each. Clusters centered at $[-51, -31, -11, 11, 31, 51]^\circ$, $\pm 2^\circ$. Symmetric lobes appear due to the inherent ULA ambiguity. All Rx power constraints set to 1. randA, # post-SDR randomizations = 300. In this case, the solution is guaranteed to be within 0.1% of the optimum.

LOW-COMPLEXITY DOWNLINK BEAMFORMING FOR MAXIMUM SUM CAPACITY

Goran Dimić

Dept. of ECE, Univ. of Minnesota,
Minneapolis MN 55455, U.S.A.
E-mail: goran@ece.umn.edu

*Nicholas D. Sidiropoulos**

Dept. of ECE, Technical Univ. of Crete,
Chania - Crete, 73100, GREECE
E-mail: nikos@telecom.tuc.gr

ABSTRACT

The problem of simultaneous multiuser downlink beamforming has recently attracted significant interest in both the Information Theory and Signal Processing communities. The idea is to employ a transmit antenna array to create multiple 'beams' directed towards the individual users, and the aim is to increase throughput, measured by sum capacity. Optimal solutions to this problem require convex optimization and so-called *Dirty Paper* (DP) precoding for known interference, which are prohibitively complex for actual online implementation at the base station. Motivated by recent results by Viswanathan *et al* and Caire and Shamai, we propose a computationally simple user selection method coupled with zero-forcing beamforming. Our results indicate that the proposed method attains a significant fraction of sum capacity, and thus offers an attractive alternative to DP-based schemes.

1. INTRODUCTION

Depending on whether or not Channel State Information (CSI) is available at the transmitter, transmit antenna arrays can be utilized in two basic ways or a combination thereof: space-time coding, and spatial multiplexing. The former can be used without CSI at the transmitter, and allows mitigation and exploitation of fading. The latter requires CSI at the transmitter, but in turn allows for much higher throughput. Until recently, transmit beamforming was mostly considered for voice services in the context of the cellular downlink. With the emergence of 3G and 4G systems, higher emphasis is being placed on packet data, which are more delay-tolerant but require much higher throughput. Hence the recent interest in transmit beamforming strategies for the cellular downlink that aim for attaining the sum capacity of the wireless channel [1, 8, 9, 4, 6, 7, 5].

*Research supported in part by the European Research Office (ERO) of the US Army under Contract N62558-03-C-0012, and in part by the Army Research Laboratory under Cooperative Agreement DADD19-01-2-0011. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of ERO and ARL of the US Army.

The scenario of interest can be modeled as a non-degraded Gaussian broadcast channel (GBC). Let N be the number of antennas at the transmitter (Base Station (BS) in a cellular context), and consider a cluster of M mobile users, each equipped with a single receive antenna. The channel between each transmit and receive antenna is constant over a certain time interval and known at the BS. The received signal is corrupted by AWGN independent across users. The BS may transmit simultaneously, using multiple transmit beams, to more than one user in the cluster.

Since the receivers cannot cooperate, successful transmission critically depends on the transmitter's ability to simultaneously send independent signals with as small interference between them as possible. Caire and Shamai [1] proposed a multiplexing technique based on coding for known interference, known as "Writing on Dirty Paper" or Costa precoding [2]. In [2], it is proven that in an AWGN channel with additional additive Gaussian interference, which is known at the transmitter in advance (non-causally), it is possible to achieve the same capacity as if there were no interference. Assuming Costa precoding and known channels at the transmitter, Vishwanath et al. [6] and Yu and Cioffi [9] have proposed algorithms that evaluate sum capacity of the GBC along with the associated optimal signal covariance matrix. However, both approaches require convex optimization in (order of) MN variables to find the optimal signal covariance matrix.

The complexity of the proposed optimization algorithms makes them unsuitable for actual implementation at the BS. A reduced-complexity suboptimal solution to sum rate maximization is proposed in [1]. It suggests the use of QR decomposition of the channel matrix combined with dirty paper (DP) coding at the transmitter. The combined approach nulls interference between data streams, and hence, it is named zero-forcing dirty-paper (ZF-DP) precoding. If $N \geq M$, ZF-DP is proven to be asymptotically optimal at both low and high SNR, but suboptimal in general; whereas zero-forcing (ZF) beamforming without DP coding is optimal in the low SNR regime and yields the same slope of throughput versus SNR in decibels as the sum capacity curve at high SNR. If $N < M$, [1] has shown that random

selection of $U \leq N$ users incurs throughput loss for both ZF-DP and ZF. Tu and Blum [5] have proposed a selection algorithm that capitalizes on multiuser diversity, thus increasing the throughput of ZF-DP precoding, and significantly narrowing the gap between ZF-DP throughput and capacity.

An important shortcoming of DP coding is that it requires vector coding and a long temporal block length to be well-approximated in practice; furthermore, with current state-of-art, such approximation entails high computational complexity [3, 8, 10]. For this reason, we advocate herein a more pragmatic approach, based on plain ZF beamforming coupled with a new user selection method. Our approach is applicable in the practically important case that the number of users exceeds the number of transmit antennas. Our simulation results indicate that, at moderate and high SNR, the proposed approach has equal slope of throughput versus SNR as the capacity curve, and it achieves a significant fraction of capacity for all SNR.

ZF beamforming without DP coding was also considered by Spencer and Haardt [4], but they did not consider user selection when $M > N$. Viswanathan et al. [7] have compared the performance of ZF versus ZF-DP, using a simpler user selection scheme that schedules the N users with the highest *individual* SINR. Under this simpler scheme, they reported that ZF is close to ZF-DP in terms of throughput. Our results further qualify [7], showing that the same is true under a more sophisticated user selection strategy that directly aims to optimize sum capacity. Furthermore, we show that with this new user selection strategy ZF comes close to attaining sum capacity.

2. ZERO-FORCING BEAMFORMING AND USER SELECTION STRATEGY

Let $h_{m,n}$ model the quasi-static, flat-fading channel between transmit antenna n and the receive antenna of user m, and denote $\mathbf{h}_m := [h_{m,1}\ h_{m,2}\ \ldots\ h_{m,N}]$. Similarly, let $\mathbf{w}_m = [w_{1,m}\, w_{2,m}\ \ldots\ w_{N,m}]^T$ ($(\cdot)^T$ denotes transpose) be the beamforming weight vector for user m. Thus the channel matrix, $\mathbf{H}$, and the beamforming weight matrix, $\mathbf{W}$, are

$$\begin{array}{c} \mathbf{H} = \left[\ \mathbf{h}_1^*\, \mathbf{h}_2^*\ \cdots\ \mathbf{h}_M^*\ \right]^* \\ \mathbf{W} = [\mathbf{w}_1\ \mathbf{w}_2\ \cdots\ \mathbf{w}_M], \end{array} \tag{1}$$

where $(.)^*$ denotes conjugate-transpose. Collecting the baseband-equivalent outputs, the received signal vector is

$$\mathbf{x} = \mathbf{HWDs} + \mathbf{n} \tag{2}$$

where $\mathbf{s}$ is the transmitted signal vector containing uncorrelated unit-power entries,

$$\mathbf{D} = \begin{bmatrix} \sqrt{p_1} & 0 & \cdots & 0 \\ 0 & \sqrt{p_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{p_M} \end{bmatrix} \tag{3}$$

accounts for power-loading and $\mathbf{n}$ is the noise vector. Note that the elements of $\mathbf{x}$ are physically distributed across the M mobile terminals. Multiuser decoding is therefore not feasible, hence each user treats the signals intended for other users as interference. Noise is assumed to be circular complex Gaussian, zero-mean, uncorrelated with variance of each complex entry $\sigma^2 = 1$.

The desired signal power received by user m is given by $|\mathbf{h}_m \mathbf{w}_m|^2 p_m$. The Signal to Interference plus Noise Ratio (SINR) of user m is

$$SINR_m = \frac{|\mathbf{h}_m \mathbf{w}_m|^2 p_m}{\sum\limits_{i \neq m} |\mathbf{h}_m \mathbf{w}_i|^2 p_i + \sigma^2}, \tag{4}$$

The problem of interest can now be formulated as

$$\boxed{\begin{array}{c} \max\limits_{\mathbf{W}} \sum\limits_{m=1}^{M} log(1 + SINR_m), \\ \text{subject to: } ||\mathbf{WD}||_F^2 \leq P, \end{array}} \tag{5}$$

where $||.||_F^2$ denotes Frobenius norm and P stands for a bound on average transmitted power.

Attaining capacity requires Gaussian signaling and long codes, yet the logarithmic SINR reward can be motivated from other, more practical perspectives as well: it can be shown that it measures the throughput of QAM-modulated systems over both AWGN and Rayleigh fading channels. The intuition is that SINR improvements eventually yield diminishing throughput returns.

ZF beamforming inverts the channel matrix at the transmitter, so that orthogonal channels between transmitter and receivers are created. It is then possible to encode users individually, as opposed to more complex long-block-vector coding needed to implement DP. Note that ZF at the transmitter does not enhance noise at the receiver. If the number of users, $M \leq N$, and $rank(\mathbf{H}) = M$, then the ZF beamforming matrix is

$$\mathbf{W} = \mathbf{H}^*(\mathbf{HH}^*)^{-1}, \tag{6}$$

which is the Moore-Penrose pseudoinverse of the channel matrix. However, if $M > N$ it is not possible to use (6) because $\mathbf{HH}^*$ is singular. In that case, one needs to select $n \leq N$ out of M users.

For $M > N$, the problem is reformulated as follows: Let $U = \{1,\ 2,\ \ldots,\ M\}$, and $S_n = \{s_u \mid s_u \in U\}$, such that $|S_n| = n$. Given $\mathbf{H} \in \mathbb{C}^{M \times N}$, select $n \leq N$, and a set of channels, $\{\mathbf{h}_{s_1},\ \ldots\ , \mathbf{h}_{s_n}\}$, which produce the row-reduced channel matrix

$$\mathbf{H}(S_n) = \left[\ \mathbf{h}_{s_1}^*\quad \mathbf{h}_{s_2}^*\quad \cdots\quad \mathbf{h}_{s_n}^*\ \right]^* \tag{7}$$

such that the sum rate is the highest achievable:

$$\max_{1\le n\le N} \max_{S_n} R_{zf}(S_n) \quad \text{subject to} \sum_{i\in S_n} \left[\mu - \frac{1}{c_i(S_n)}\right]_+ = P. \tag{8}$$

We define,

$$R_{zf}(S_n) := \sum_{i\in S_n} [log_2(\mu c_i(S_n)]_+, \tag{9}$$

where $[x]_+ = max\{0, x\}$,

$$c_i(S_n) = \{[(\mathbf{H}(S_n)\mathbf{H}(S_n)^*)^{-1}]_{i,i}\}^{-1}, \tag{10}$$

and μ is obtained by solving the water-filling equation in (8). The power-loading then yields

$$p_i = c_i(S_n)\left[\mu - \frac{1}{c_i(S_n)}\right]_+, \quad \forall i \in S_n. \tag{11}$$

The problem can be conceptually solved by exhaustive search: for each value of n, find all possible n-tuples S_n and select a pair (n, S_n) which yields maximum $R_{zf}(S_n)$. However, such an algorithm has prohibitive complexity.

We propose a reduced-complexity suboptimal algorithm, dubbed Generalized Zero Forcing (GZF), as outlined next.

1. **Initialization:**
   - Set $n = 1$.
   - Find a user, s_1, such that $s_1 = \arg\max_{u\in U} \mathbf{h}_u\mathbf{h}_u^*$.
   - Set $S_1 = \{s_1\}$ and denote the achieved rate $R_{zf}(S_1)_{max}$.
2. **While** $n < N$**:**
   - $n = n + 1$.
   - Find a user, s_n, such that
   $$s_n = \arg\max_{u\in U\setminus S_{n-1}} R_{zf}(S_{n-1}\cup\{u\}).$$
   - Set $S_n = S_{n-1}\cup\{s_n\}$ and denote the achieved rate $R_{zf}(S_n)_{max}$.
   - If $R_{zf}(S_n)_{max} \le R_{zf}(S_{n-1})_{max}$ **break** and retain solution $(n-1, S_{n-1})$.
3. **Beamforming:** $\mathbf{W} = \mathbf{H}(S_n)^*(\mathbf{H}(S_n)\mathbf{H}(S_n)^*)^{-1}$
   Power Loading: Water-filling

2.1. Implementation and Complexity

The most complex task is the evaluation of $R_{zf}(S_{n-1}\cup\{u\})$. From (9), it is split into the evaluation of the $c_i(S_{n-1}\cup\{u\})$'s followed by evaluation of μ. An efficient way to evaluate the $c_i(S_{n-1}\cup\{u\})$'s is by using the matrix inversion lemma to invert the matrix $\mathbf{A}(S_{n-1}\cup\{u\}) := \mathbf{H}(S_{n-1}\cup\{u\})\mathbf{H}(S_{n-1}\cup\{u\})^*$. Note that

$$\mathbf{A}(S_{n-1}\cup\{u\}) = \begin{bmatrix} \mathbf{A}(S_{n-1}) & \mathbf{a}_u \\ \mathbf{a}_u^* & a_{u,u} \end{bmatrix},$$

where $\mathbf{a}_u = [\mathbf{h}_{s_1}\mathbf{h}_u^*,\ \mathbf{h}_{s_2}\mathbf{h}_u^*,\ \ldots\ \mathbf{h}_{s_{n-1}}\mathbf{h}_u^*]^T$ and $a_{u,u} = \mathbf{h}_u\mathbf{h}_u^*$. Noting that $\mathbf{A}(S_{n-1})^* = \mathbf{A}(S_{n-1})$, and writing

$$\mathbf{q} = \mathbf{A}(S_{n-1})^{-1}\mathbf{a}_u, \tag{12}$$

after some algebraic manipulation we obtain

$$\begin{aligned}\mathbf{A}(S_{n-1}\cup\{u\})^{-1} = &\begin{bmatrix} \mathbf{A}(S_{n-1})^{-1} & \mathbf{0}_{n-1} \\ \mathbf{0}_{n-1}^T & 0 \end{bmatrix} \\ &+ (a_{u,u} - \mathbf{a}_u^*\mathbf{q})^{-1}\begin{bmatrix} \mathbf{q}\mathbf{q}^* & -\mathbf{q} \\ -\mathbf{q}^* & 1 \end{bmatrix},\end{aligned} \tag{13}$$

where $\mathbf{0}_{n-1}^T = [0\ 0\ \ldots\ 0]_{1\times(n-1)}$. It can be verified that each time n is increased $\mathbf{A}(S_{n-1})^{-1}$ and $a_{i,u},\ i \in S_{n-2}$, are known before the search over $u \in U \setminus S_{n-1}$ starts. Hence, evaluation of $\mathbf{A}(S_{n-1}\cup\{u\})^{-1}$ from (12) and (13) has complexity proportional to $O(n^2)$.

Given a set S_n, we have [1]

$$c_i(S_n) = |\mathbf{h}_{s_i}\mathbf{P}(S_n\setminus\{s_i\})^\perp|^2, \tag{14}$$

where $\mathbf{P}(S_n)^\perp$ denotes the projector onto the orthogonal complement of $\Omega(S_n) = span\{\mathbf{h}_{s_l} : s_l \in S_n\}$. It follows that if (8) and (11) yield $p_u = 0$, then $R_{zf}(S_{n-1}\cup\{u\}) < R_{zf}(S_{n-1})$. We discard such u. We also discard u if (8) and (11) yield $p_{s_i} = 0$ for some $s_i \in S_{n-1}$. This is done to keep complexity at bay, for otherwise combinatorial search might effectively emerge. Hence, user u is a candidate for S_n if $p_i > 0,\ \forall i \in S_{n-1}\cup\{u\}$. From the properties of water-filling, this holds if

$$\frac{n}{c_{i_{min}}(S_{n-1}\cup\{u\})} < P + \sum_{i\in S_{n-1}\cup\{u\}} \frac{1}{c_i(S_{n-1}\cup\{u\})}, \tag{15}$$

where $c_{i_{min}}(S_{n-1}\cup\{u\}) = \min_{i\in S_{n-1}\cup\{u\}} c_i(S_{n-1}\cup\{u\})$. Then, we have

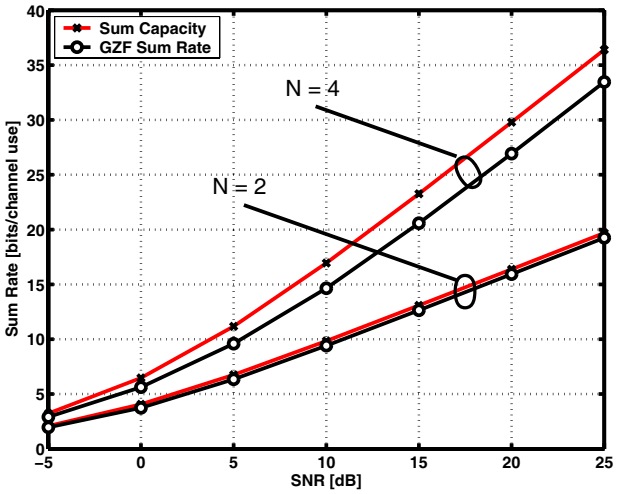
$$\mu = \frac{1}{n}\left[P + \sum_{i\in S_{n-1}\cup\{u\}} \frac{1}{c_i(S_{n-1}\cup\{u\})}\right]. \tag{16}$$

If (15) is not satisfied, we skip to the next u. The overall complexity of the algorithm is $O(N^3M)$.

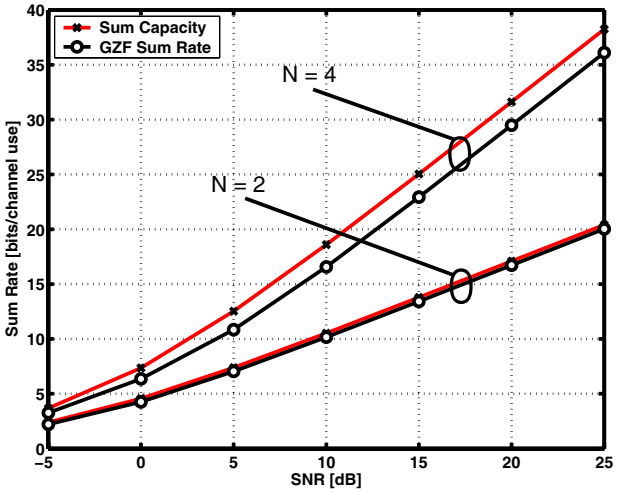
We note that the **break** in Step 2 is necessary when GZF is used, but redundant when ZF-DP is used; it is shown in [1, 5] that in the latter case, maximum sum rate can always be achieved with N active users if $P > 0$ [1]. On the other hand, when ZF alone is used, the optimum number of active users is $n_{opt} \le N$ and decreases as P decreases, so that for $P \to 0$, the ZF scheme reduces to maximum ratio combining (MRC), $n_{opt} = 1$ [1]. This also holds for the proposed GZF algorithm, which follows from the water-filling equation in (8) and the fact that $[c_1(S_1)]^{-1} = \max_{i\in U} a_{i,i}$.

3. SIMULATION RESULTS

The performance of the proposed algorithm is presented in Fig. 1. The y-axis shows sum capacity and sum rate in bits per channel use. The x-axis shows total power in dB. Noise level of every user is 1. Sum capacity and sum rates are averaged over 100 channels. Channels are complex-valued, drawn from an i.i.d. Rayleigh distribution with unit-variance for each channel entry. Note that GZF exhibits the same slope of rate increase per dB of SNR as the sum capacity curve at moderate and high SNR. Also note that given N, an increase in M narrows the gap between the sum rate, achieved using GZF, and the sum capacity. This is due to multiuser diversity - the more users that contend for transmission, the higher the probability that N of them will be almost orthogonal. This in turn reduces the advantage of DP-coding based schemes over ZF.



(a) M = 8 users



(b) M = 16 users

Fig. 1. GZF Performance

4. CONCLUSIONS

We have proposed a low-complexity algorithm for downlink transmission in the GBC for the realistic case wherein the number of users is greater than the number of transmit antennas. We have evaluated the throughput performance of the new algorithm via simulations. The results show that ZF beamforming with the proposed user selection method achieves a significant fraction of sum capacity, at a low complexity cost. The simulation results indicate that GZF achieves the same slope of throughput per dB of SNR as the capacity-achieving strategy based on the use of DP coding for known interference cancellation and convex optimization. Due to its simplicity, low complexity, and close to optimal performance, the proposed method offers an attractive alternative to earlier DP-based methods.

5. REFERENCES

[1] G. Caire and S. Shamai (Shitz), "On the Achievable Throughput of a Multi-Antenna Gaussian Broadcast Channel," in *IEEE Trans. on Info. Theory*, vol. 49, no. 7, July 2003, pp. 1691–1706

[2] M. H. M. Costa, "Writing on Dirty Paper," *IEEE Trans. on Info. Theory,* vol. IT-29, no. 3, May 1983.

[3] C.B. Peel, "On Dirty Paper Coding", *Signal Processing Magazine*, May 2003, pp. 112-113.

[4] Q. Spencer and M. Haardt, "Capacity and Downlink Transmission Algorithms for a Multi-user MIMO Channel," in *Proc. Of the 36th Asilomar Conf. On Sign. Syst. And Comp.,* Pacific Grove, CA, Nov. 2002.

[5] Z. Tu and R. S. Blum, "Multiuser Diversity for a Dirty Paper Approach," *IEEE Comm. Letters,* vol. 7, no. 8, Aug. 2003, pp. 370–372

[6] S. Vishwanath, N. Jindal and A. Goldsmith, "Duality, Achievable Rates and Sum-Rate Capacity of Gaussian MIMO Broadcast Channels," *IEEE Trans. on Info. Theory.*, vol. 49, no. 10, Oct. 2003, pp. 2658–2668

[7] H. Viswanathan, S. Venkatesan and H. Huang, "Downlink Capacity Evaluation of Cellular Networks with Known Interference Cancellation," *IEEE J. on Sel. Areas in Comm.*, vol. 21, no. 5, June 2003, pp. 802–811

[8] W. Yu and J. M. Cioffi, "Trellis Precoding for the Broadcast Channel," in *Proc. of Globecom 2001,* San Antonio, TX, November 2001.

[9] W. Yu and J. M. Cioffi, "Sum Capacity of a Gaussian Broadcast Channel," in *Proc. of IEEE Int. Symp. on Inform. Theory,* ISIT 2002, Lausanne, Switzerland, July 2002.

[10] R. Zamir, S. Shamai (Shitz), and U. Erez, "Nested Linear/Lattice Codes for Structured Multiterminal Binning," *IEEE Trans. on Inform. Theory*, vol. 48, no 6., June 2002, pp. 1250–1276