EFFICIENT IMPLEMENTATION OF A QUASI-MAXIMUM-LIKELIHOOD DETECTOR BASED ON SEMI-DEFINITE RELAXATION

Mikalai Kisialiou and Zhi-Quan Luo

University of Minnesota
Department of Electrical and Computer Engineering
Minneapolis, MN, 55455, United States,
{kisi0004, luozq}@umn.edu

ABSTRACT

Existing approaches to the Maximum-Likelihood (ML) detection problem in digital communications either suffer from exponential complexity (e.g. Sphere Decoder and its variants) or exhibit significant Bit-Error-Rate (BER) degradation (e.g. LMMSE Detector). In this paper we present an efficient implementation of a semi-definite relaxation-based detector (SDR Detector) which can achieve near-optimal BER performance with worst-case polynomial complexity. This implementation (available online) can be 100 times faster than an off-the-shelf SeDuMi-based implementation, outperforms Sphere Decoder in low Signal-to-Noise Ratio (SNR) or high dimension regimes, and matches the speed of Sphere Decoder in the high SNR regime. The core of the detector is an optimized dual-scaling interior-point method (implemented in C) for the relaxed semi-definite program. SNR-sensitive improvements are achieved by a dimension reduction strategy and a warm start technique based on a truncated version of the Sphere Decoding algorithm. Extensive numerical simulations show that the BER performance and the running time of SDR Detector compare favorably to that of other near-optimal detection strategies.

Index Terms— Maximum likelihood detection, MIMO systems, semi-definite relaxation, interior-point methods, duality.

1. INTRODUCTION

Consider a standard Rayleigh fading vector communication channel with n transmit and m receive antennas:

$$\mathbf{y} = \sqrt{\rho/n}\,\mathbf{H}\,\mathbf{s} + \mathbf{v}, \tag{1}$$

where $\mathbf{s} \in \{-1,+1\}^n$ is the vector of transmitted signals, ρ is Signal-to-Noise Ratio (SNR), $\mathbf{H} \in \mathbb{R}^{m\times n}$, $H_{ik} \sim \mathcal{N}(0,1)$ is the matrix of fading coefficients, $\mathbf{v} \in \mathbb{R}^m$, $v_i \sim \mathcal{N}(0,1)$ is i.i.d. noise, and $\mathbf{y} \in \mathbb{R}^m$ is the vector of received signals. The same channel model can be used to describe a synchronous CDMA multi-access channel with n users. ML detection is known to deliver optimal BER performance in many practical scenarios. For binary modulated signals, the ML detection problem is given by:

$$\mathbf{s}_{ML} = \arg \min_{\mathbf{s}\in\{-1,+1\}^n} \|\mathbf{y} - \sqrt{\rho/n}\,\mathbf{H}\,\mathbf{s}\|^2. \tag{2}$$

This problem is known to be NP-hard. When problem dimensions are small, exhaustive search can be applied to solve (2). However, for large m and n, the exhaustive search is impractical.

This research is supported in part by the National Science Foundation, Grant No. DMS-0610037.

In this work we focus on suboptimal strategies to solve (2) with near-optimal BER performance. Sphere Decoding algorithm [1] and its variants [2] achieve lower complexity by restricting the exhaustive search to a sphere centered at the zero-forcing estimate. In the high SNR regime, this strategy allows a fast implementation due to proximity of the ML solution to the zero-forcing estimate. However, exponential complexity [3] of Sphere Decoder makes the algorithm impractical in low SNR regime or for large problems.

An alternative approach [4] with near-optimal BER performance is based on a convex semi-definite relaxation of the ML detection problem followed by a (randomized) rounding procedure. Interior-point methods can be used to implement this strategy with polynomial worst-case complexity, $\mathcal{O}(n^{3.5})$. The running time of existing semi-definite relaxation detectors scales well with problem size and is insensitive to SNR. This insensitivity is a blessing in the low SNR regime where the ML detection problem is more difficult. However, it becomes a curse in the high SNR regime since it implies the algorithm fails to effectively exploit the low noise property of the channel, as does the Sphere Decoder. Another drawback of the current implementations is the lack efficient termination procedure. All bits are rounded simultaneously irrespective of their reliabilities.

The implementation of the quasi-ML detector presented in this paper, SDR Detector, avoids the bottlenecks of the semi-definite detector or exponential complexity of Sphere Decoder. The proposed algorithm achieves near-optimal BER performance with complexity that scales polynomially in all SNR regimes and for all problem dimensions. The core of the proposed implementation is the dual-scaling interior-point algorithm [5]. A warm start technique implemented with a truncated version of the Sphere Decoding algorithm provides an SNR-sensitive initialization. The standard randomized rounding step is replaced with a dynamic dimension reduction technique which estimates bit reliabilities at every iteration of the dual-scaling algorithm and rounds those bits whose reliabilities exceed a given threshold. The contribution of the rounded bits is then eliminated from the ML problem, leading to a reduced problem size and simplified computation in subsequent interior-point iterations.

2. QUASI-MAXIMUM-LIKELIHOOD DETECTION

A. Sphere Decoder

Sphere Decoder originates from the algorithm for computing the shortest vector in a lattice [1]. Various improvements [2] (e.g. adjustable radius search procedure) have been proposed to adapt it to the ML detection problem, demonstrating impressive running time for small systems operating in the high SNR regime. Unfortunately, (average and worst-case) complexity of the Sphere Decoding algo-

rithm is exponential [3]. A lower bound on the expected complexity of Sphere Decoder for binary-modulated signals is given by:

$$C(n) \geq 2^{n\,\eta(\rho)} - 1, \quad \eta(\rho) = 1/(4\rho + 2).$$

For small systems operating in the high SNR regime (product $n\ \eta(\rho)$ is small) the running time of Sphere Decoder implementations is dominated by initialization, memory allocations and input-output operations. For large systems or in the low SNR regime (product $n\ \eta(\rho)$ is large) the running time of any Sphere Decoding implementation grows exponentially. This behavior is a direct result of the nature of exhaustive search which is at the core of Sphere Decoder. For a polynomial time radius selection procedure the expected number of vectors found inside a sphere with at least one feasible point inside is exponential, causing exponential complexity of such implementations. Therefore, a different approach to the ML detection problem is required in order to construct an algorithm whose complexity scales well in all SNR regimes and for problems of large dimensions. Quasi-ML detection based on semi-definite relaxation provides an algorithm that offers theoretically guaranteed worst-case polynomial complexity.

B. Semi-definite relaxation

The ML detection problem (2) allows an equivalent reformulation [4]:

$$\begin{array}{rl} f_{ML} := \min & \mathrm{Trace}(\mathbf{QX}) \\ \text{s.t.} & \mathrm{diag}(\mathbf{X}) = \mathbf{1}, \\ & \mathbf{X} \succeq 0, \\ & \mathbf{X} = \mathbf{x}\mathbf{x}^T, \end{array} \tag{3}$$

where matrix $\mathbf{Q} \in \mathbb{R}^{(n+1)\times(n+1)}$ and vector $\mathbf{x} \in \mathbb{R}^{n+1}$ are:

$$\mathbf{Q} = \begin{bmatrix} (\rho/n)\,\mathbf{H}^T\mathbf{H} & -\sqrt{\rho/n}\,\mathbf{H}^T\mathbf{y} \\ -\sqrt{\rho/n}\,\mathbf{y}^T\mathbf{H} & \|\mathbf{y}\|^2 \end{bmatrix}, \qquad \mathbf{x} = \begin{bmatrix} \mathbf{s} \\ 1 \end{bmatrix}. \tag{4}$$

The semi-definite detector [4] relaxes the constraint on the rank of matrix $\mathbf{X}$ to obtain a convex Semi-Definite Program (SDP):

$$\begin{array}{rl} f_p(\mathbf{X}) := \min & \mathrm{Trace}(\mathbf{QX}) \\ \text{s.t.} & \mathrm{diag}(\mathbf{X}) = \mathbf{1}, \\ & \mathbf{X} \succeq 0. \end{array} \tag{5}$$

A subsequent randomized rounding procedure generates estimates of transmitted signals based on the optimal solution $\mathbf{X}_{opt}$ of this SDP [4]:

- Compute the spectral decomposition $\mathbf{X}_{opt} = \sum_{i=1}^{n+1} \lambda_i \mathbf{u}_i \mathbf{u}_i^T$ and set $\mathbf{v}_i = \sqrt{\lambda_i}\mathbf{u}_i$, $i = 1, \ldots, n+1$.
- Pick vector $\mathbf{v}^{max}$ that corresponds to the largest eigenvalue $\mathbf{v}^{max} = \arg\max_{1\leq i\leq n+1} \{\|\mathbf{v}_i\|\}$.
- For each entry x_i define Bernoulli distribution:

$$\begin{array}{l} \Pr\{x_i = +1\} = (1 + v_i^{max})/2, \\ \Pr\{x_i = -1\} = (1 - v_i^{max})/2. \end{array} \tag{6}$$

- Generate a fixed number (typically $5 - 10$) of i.i.d. $(n{+}1)$-dimensional vector samples $\bar{\mathbf{x}}_d$ according to the distribution (6).
- For all samples, set $\bar{\mathbf{x}}_d := -\bar{\mathbf{x}}_d$ if $(n{+}1)$-st entry of $\bar{\mathbf{x}}_d$ is equal to -1.
- Pick $\mathbf{x}_{SDR} := \arg\min_d \bar{\mathbf{x}}_d^T \mathbf{Q} \bar{\mathbf{x}}_d$ and set the best achieved objective value $f_{SDR} := \mathbf{x}_{SDR}^T \mathbf{Q} \mathbf{x}_{SDR}$.

3. SDR DETECTOR

A. Dual-scaling algorithm

The dual-scaling interior point method [5] has been developed to solve general large-scale semi-definite programs. In addition to the advantages associated with a standard interior point method (convergence proof with polynomial worst-case complexity, certificates of infeasibility when no solution exists, robustness and scalability, etc) the dual-scaling method efficiently exploits the structure and sparsity of in the dual semi-definite programs.

Existing implementations of the semi-definite relaxation detector rely on interior-point methods which solve the primal problem (5) and/or the following dual problem:

$$\begin{array}{rl} f_d(\mathbf{g}) := \max & \mathbf{g}^T\mathbf{1} \\ \text{s.t.} & \mathbf{Q} - \mathrm{Diag}(\mathbf{g}) \succeq 0. \end{array} \tag{7}$$

The algorithm computes a sequence of primal feasible (given by $\mathbf{X}$) and dual feasible (given by $\mathbf{g}, \mathbf{S}$) points on the central path, given by:

$$\begin{array}{r} \mathrm{diag}(\mathbf{X}) = \mathbf{1}, \\ \mathbf{Q} - \mathrm{Diag}(\mathbf{g}) = \mathbf{S}, \\ \mathbf{SX} = \nu\mathbf{I}. \end{array} \tag{8}$$

As $\nu \to 0$, the sequence converges to the optimal solution and the system (8) specifies Karush-Kuhn-Tucker (KKT) optimality conditions for $\mathbf{X}, \mathbf{g}, \mathbf{S}$. Each step along the central path is calculated as the solution to the following linearization of the system (8):

$$\begin{array}{c} \mathrm{diag}(\Delta\mathbf{X}) = \mathbf{0}, \\ \Delta\mathbf{S} + \mathrm{Diag}(\Delta\mathbf{g}) = \mathbf{0}, \\ \nu\mathbf{S}^{-1}\Delta\mathbf{S}\,\mathbf{S}^{-1} + \Delta\mathbf{X} = \nu\mathbf{S}^{-1} - \mathbf{X}. \end{array}$$

Solving this system of matrix equations with respect to dual variables, we obtain the following condition:

$$\nu\ \left(\mathbf{S}^{-1} \circ \mathbf{S}^{-1}\right) \Delta\mathbf{g} = \mathbf{1} - \nu\ \mathrm{diag}(\mathbf{S}^{-1}), \tag{9}$$

where $\circ$ denotes Hadamard (component-wise) product. For a given parameter ν, the preconditioned conjugate gradients method is applied to solve the linear system (9) for the dual step direction $\Delta\mathbf{g}$.

Once $\Delta\mathbf{g}$ is selected by solving (9), inexact line search of the step size τ for the updated dual variables $\mathbf{g}^+ := \mathbf{g} + \tau\Delta\mathbf{g}$ is performed by minimizing the dual potential function

$$f^d := \gamma \log\left(\bar{z} - (\mathbf{g}^+)^T\mathbf{1}\right) - \log\det\left(\mathbf{Q} - \mathrm{Diag}(\mathbf{g}^+)\right),$$

where $\bar{z} = \mathrm{Trace}(\mathbf{QX})$ is an upper bound computed at some primal feasible $\mathbf{X}$. As long as the dual vector $\mathbf{g}^+$ remains feasible, the objective weight factor γ that minimizes the dual potential function is given by $\gamma = (\bar{z} - (\mathbf{g}^+)^T\mathbf{1})/\nu$. The algorithm starts with $\tau := \min\{1,\ 0.95\tau_{max}\}$ and then backtracks until sufficient descent or termination tolerance is achieved. The maximum step size τ_{max} that ensures feasibility of $\mathbf{g}^+ \in \{\ \mathbf{g} \mid \mathbf{Q} - \mathrm{Diag}(\mathbf{g}) \succeq 0\}$, is given by the distance to the boundary of the semi-definite cone, which is equal to $\lambda_{max}^{-1}\left(\mathbf{L}^{-1}\mathrm{Diag}(\Delta\mathbf{g})\mathbf{L}^{-T}\right)$, where $\mathbf{L}$ is the lower triangular Cholesky factor of $\mathbf{S} = \mathbf{Q} - \mathrm{Diag}(\mathbf{g})$. To compute the largest eigenvalue of $\mathbf{A} \triangleq \mathbf{L}^{-1}\mathrm{Diag}(\Delta\mathbf{g})\mathbf{L}^{-T}$ we apply Lanczos procedure [6]. For a symmetric matrix $\mathbf{A}$ and a vector $\mathbf{u}_1, \|\mathbf{u}_1\| = 1$, Lanczos iteration constructs a basis $\mathbf{U}_i = [\mathbf{u}_1, \ldots, \mathbf{u}_i]$ in the Krylov subspace $\{\mathbf{u}_1, \mathbf{A}\mathbf{u}_1, \ldots, \mathbf{A}^{i-1}\mathbf{u}_1\}$, and a tridiagonal matrix $\mathbf{T}_i, i \times i$, such that

$$\begin{array}{r} \mathbf{AU}_i = \mathbf{U}_i\mathbf{T}_i + t_{i+1}\mathbf{u}_{i+1}\mathbf{1}_i^T \\ \mathbf{U}_i^T\mathbf{AU}_i = \mathbf{T}_i, \end{array}$$

Input: $R_0, \Delta R, N_r, N_v$
- R_0 is starting radius
- ΔR is radius step size
- N_r is number of radii to be searched
- N_v is number of vectors to be searched

$R := R_0$
while (number of radii searched $\leq N_r$)
  if (found a vector inside sphere with radius R)
    $\mathbf{s}_{min} := \mathbf{1}$; $f_{min} := \|\mathbf{y} - \sqrt{\rho/n}\,\mathbf{H}\,\mathbf{1}\|^2$
    while (number of vectors searched $\leq N_v$)
      for each ($\mathbf{s}$ found in the sphere)
        $f(\mathbf{s}) := \|\mathbf{y} - \sqrt{\rho/n}\,\mathbf{H}\,\mathbf{s}\|^2$
        if ($f(\mathbf{s}) < f_{min}$)
          $f_{min} := f(\mathbf{s})$; $\mathbf{s}_{min} = \mathbf{s}$
        end if
      end for
      if (no vectors left in sphere with radius R)
        return ML solution $\mathbf{s}_{min}$
      end if
    end while
    return best guess $\mathbf{s}_{min}$
  end if
  $R := R + \Delta R$
end while
return $\mathbf{s}_{min}$ is not found

Fig. 1. Truncated version of Sphere Decoder

where $\mathbf{1}_i$ is i-th basis vector. The extreme eigenvalues of $\mathbf{A}$ are well approximated by those of $\mathbf{T}_i$ with far fewer iterations i than the problem dimension n.

B. Warm start with Sphere Decoder

The Sphere Decoding algorithm with adjustable radius search serves as a fast heuristic test of low noise channel realizations. The initialization routine of SDR Detector uses a truncated version of Sphere Decoder (see Fig. 1) to curb its exponential complexity in low SNR regime or for large problems. The truncated Sphere Decoder is restricted by the following constant parameters: initial radius R_0, radius increase ΔR, upper bound N_r on the number of radius increases, and upper bound N_v on the number of times the objective function can be computed.

The maximum number of sphere expansions is selected to ensure that complexity of the truncated Sphere Decoder does not dominate complexity of the dual-scaling algorithm:

$$\mathcal{O}\left(n^{3.5}\right) \simeq \text{C(SDP)} \simeq \text{C(SD)} \simeq \mathcal{O}\left(\left(N_r\,\Delta R\right)^{n/\rho}/\rho\right).$$

Thus, the number of times the algorithm is allowed to increase the radius of the sphere is set $N_r = \mathcal{O}\left(\left(n^{3.5}\,\rho\right)^{\rho/n}/\Delta R\right)$. For the heuristic radius search procedure the expected number of vectors found within a sphere is exponential, $2^{n/\rho}$. SDR Detector heuristically sets the number of vectors allowed to be searched in a sphere to be a decreasing function of ρ and n: $N_v = \max\left\{2, \mathcal{O}\left(2^{-n/\rho}/\rho\right)\right\}$.

The smallest objective value f_{min} achieved by the truncated Sphere Decoder is used to initialize upper bound $\bar{z} := f_{min}$ in the dual-scaling algorithm. A good initial upper bound substantially improves the convergence of the dual-scaling interior-point method.

Input: β, S_{rnd}
- $\beta,\ 0 < \beta < 1,$ is reliability threshold
- S_{rnd} is set of rounded bits

/ * Find the most reliable bit * /
$S_{rnd} := \emptyset$; $w_{max} := |v_1^{max}|$; $i_{max} := 1$
for ($\forall\, i, 1 < i \leq n$)
  if ($|v_i^{max}| > w_{max}$)
    $w_{max} := |v_i^{max}|$; $i_{max} := i$
  end if
end for
/ * Round bits β-fraction away from the maximum * /
for ($\forall\, i, 1 < i \leq n$)
  if ($|v_i^{max}| > \beta\, w_{max}$)
    $x_i := \text{sign}\,(v_i^{max} v_{n+1}^{max})$ / * Round i-th bit * /
    $S_{rnd} := \{S_{rnd}, i\}$ / * Add i-th bit to the set * /
  end if
end for
/ * Reduce problem dimension * /
$n := n - |S_{rnd}|$
Recursively update matrix $\mathbf{Q}$

Fig. 2. Dimension reduction technique

C. Dimension reduction technique

Every step $\Delta\mathbf{g}$ along the central path in the dual-scaling algorithm requires an expensive linear solve of the system (9) with the cost of $\mathcal{O}(n^3)$ operations. In addition, the randomized rounding procedure (6) needs only the principal eigenvector $\mathbf{v}^{max}$ of $\mathbf{X}_{opt}$, ($\mathbf{v}^{max}$ conveys bit-reliability information). We propose a dimension reduction technique that computes bit reliabilities based on the dual variables $\mathbf{g}$ bypassing expensive computation of $\mathbf{X}$. During every interior point iteration, a fraction of the most reliable bits is rounded and their contribution is eliminated from semi-definite problems (5) and (7). This reduces the problem size and greatly simplifies the subsequent interior point iterations.

Equations (8) show that the eigenvectors of $\mathbf{X}$ and $\mathbf{S}$ are the same on the central path, and the principal eigenvector of $\mathbf{X}$ is the eigenvector of $\mathbf{S}$ corresponding to the smallest eigenvalue:

$$\mathbf{v}^{max} = \arg\min_{\mathbf{u},\|\mathbf{u}\|=1} \mathbf{u}^T\,\mathbf{S}\mathbf{u} = \arg\min_{\mathbf{u},\|\mathbf{u}\|=1} \mathbf{u}^T\,(\mathbf{Q} - \text{Diag}(\mathbf{g}))\,\mathbf{u}.$$

Given a constant parameter β and dual vector $\mathbf{g}$ on the central path, we use Lanczos procedure to compute $\mathbf{v}^{max}$ and round the β-fraction of the most reliable bits, see Fig. 2 for a description. Eliminating the contribution of the rounded bits from matrix $\mathbf{Q}$ can be accomplished with a quadratic total cost, $\mathcal{O}(n^2)$, due to the matrix structure (4):

1. Update the received vector $\mathbf{y} := \mathbf{y} - \sum_{i \in S_{rnd}} x_i \mathbf{h}_i$, where $\mathbf{h}_i$ is the i-th column of matrix $\mathbf{H}$.
2. Remove the i-th row and the i-th column of matrix $\mathbf{Q}$.
3. Update matrix $\mathbf{H}$ by removing columns with indices from S_{rnd}.
4. Recompute the last row and column of matrix $\mathbf{Q}$ as the new product: $-\sqrt{\rho/n}\,\mathbf{H}^T\mathbf{y}$.
5. Update the $(n+1, n+1)$-entry of matrix $\mathbf{Q}$ by computing the new product: $\mathbf{y}^T\mathbf{y}$.

The complexity of the dimension reduction technique, $\mathcal{O}(n^3)$, is dominated by Lanczos procedure to compute the principal eigenvector $\mathbf{v}^{max}$.

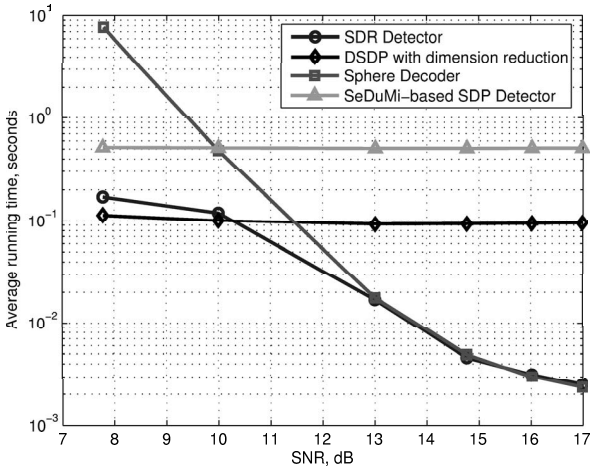

Fig. 3. Running time comparison, $n = 60$

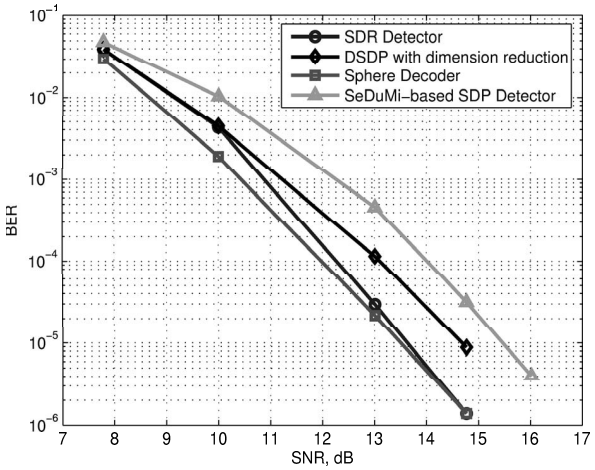

Fig. 4. Bit-error-rate comparison, $n = 60$

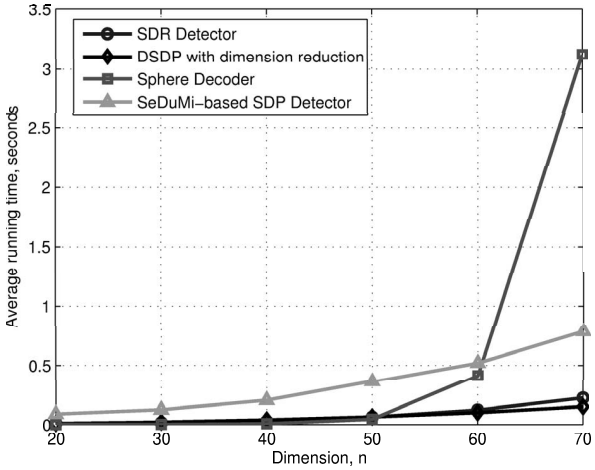

Fig. 5. Running time for large problems, $\rho = 10\ dB$

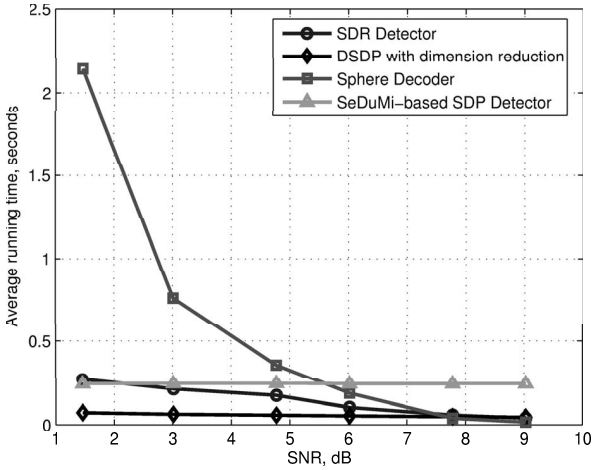

Fig. 6. Running time in low SNR regime, $n = 40$

4. SIMULATIONS

In this section we compare the running time and BER performance of SDR Detector with Sphere Decoder, SeDuMi-based semi-definite relaxation detector, and DSDP-based semi-definite relaxation detector with the dimension reduction technique. We selected an efficient implementation of the dual-scaling interior point method provided by DSDP optimization package [7]. All detectors, except for the SeDuMi-based one, are implemented in ANSI C with mex-interfaces for Matlab to eliminate language-specific differences.

Fig. 3 and 4 demonstrate the average running time and corresponding BER performance vs. SNR achieved by the selected quasi-ML detectors for problem size $n = 60$. Notice, the running time of DSDP-based (SeDuMi-based) detector is insensitive to SNR, and BER performance shows 1 dB (2-dB) SNR loss. Sphere Decoder is faster than semi-definite based detectors in high SNR regime but suffers from exponential complexity for SNR lower than 10 dB. SDR Detector matches the speed of Sphere Decoder in high SNR regime, follows the running time of polynomial detectors in low SNR regime, and enjoys near-ML BER performance.

Figs. 5 and 6 compare the average running time for large problems and in low SNR regime. The running time of polynomial complexity detectors (SDR Detector, SeDuMi and DSDP based) scales well in both regimes, remaining in the sub-second region, while the running time of Sphere Decoder exhibits exponential behavior.

5. REFERENCES

[1] U. Fincke and M. Pohst, "Improved methods for calculating vectors of short length in a lattice, including a complexity analysis," *Mathematics of Computation*, vol. 44, pp. 463 – 471, 1985.

[2] G.B. Giannakis, Z. Liu, X. Ma and S. Zhou, "Space-time coding for broadband wireless communications," *Wiley-Interscience*, 1-st ed., 2003.

[3] J. Jalden and B. Ottersten, "An exponential lower bound on the expected complexity of sphere decoding," *Proceedings of ICASSP '04*, vol. 4, pp. IV-393 – IV-396, 2004.

[4] W.K. Ma, T.N. Davidson, K.M. Wong, Z.-Q. Luo and P.C. Ching, "Quasi-maximum-likelihood multiuser detection using semi-definite relaxation," *IEEE Transactions on Signal Processing*, vol. 50, no. 4, pp. 912 – 922, 2002.

[5] S.J. Benson, Y. Ye and X. Zhang, "Solving large-scale sparse semidefinite programs for combinatorial optimization," *SIAM Journal on Optimization*, vol. 10, no. 2, pp. 443 – 461, 2000.

[6] K.C. Toh, "A note on the calculation of step-lengths in interior-point methods for semidefinite programming," *Computational Optimization and Applications*, vol. 21, pp. 301 – 310, 2002.

[7] S.J. Benson and Y. Ye, "DSDP5: Software for semidefinite programming," submitted to *ACM Transactions on Mathematical Software*, 2005.