PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA

MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH

KASDI MERBAH OUARGLA UNIVERSITY

Faculty of New Technologies of Information and Communication

Department of Electronics and Telecommunication

Thesis

Submitted in partial fulfillment of the requirements for Doctorate degree

Option: Telecommunication Systems

Presented by **LABED SMAIL**

Theme

Performance Improvement and Complexity Reduction of Massive MIMO Communication System

Jury:

President	Hamza Azzedine	Associate Professor	University of Ouargla
Director	Aounallah Naceur	Professor	University of Ouargla
Examiner	Rachedi Mohamed Yacine	Associate Professor	University of Ouargla
Examiner	Khelil Abdellatif	Professor	University of El-Oued
Examiner	Khalfa Ali	Associate Professor	University of M'Sila
Examiner	Louazene Hassiba	Associate Professor	University of Ouargla

January 2026

بسم الله الرحمن الرحيم

Acknowledgements

I would like to express my deepest gratitude and sincere appreciation to my thesis supervisor, Professor Naceur Aounallah, for his steadfast support, invaluable guidance, and unwavering dedication throughout this research journey. His expertise, insightful feedback, and constant encouragement have profoundly shaped both the direction and the quality of this work. I am truly fortunate to have benefited from his mentorship, scholarly wisdom, and exemplary professionalism.

My heartfelt thanks also go to my PhD dissertation committee members: Dr. Hamza Azzedine, Associate Professor at University of Kasdi Merbah Ouargla; Dr. Mohamed Yacine Rachedi, Associate Professor at University of Kasdi Merbah Ouargla; Pr. Abdellatif Khelil, Professor at University of Shahid Hama Lakhdar El Oued; Dr. Ali Khalfa, Associate Professor at University of Mohamed Boudiaf M'sila; and Dr. Hassiba Louazene, Associate Professor at University of Kasdi Merbah Ouargla; for their valuable comments and suggestions that helped improve the quality of this dissertation. Their time, consideration, and constructive insights are deeply appreciated.

I am equally indebted to the individuals and friends who stood by me with unwavering support, helping me navigate challenges with optimism and resilience. To each of you, I extend my sincere gratitude.

To all my family members, your love, patience, and constant inspiration have been my greatest source of strength. This achievement is as much yours as it is mine.

Finally, I would like to thank my friends and colleagues at the University of Kasdi Merbah Ouargla, as well as everyone who supported me throughout this period. Your camaraderie, encouragement, and kindness have been truly invaluable.

Abstract

This dissertation addresses the critical challenges in designing scalable, energy efficient (EE), and high performance physical layer processing for large-scale millimeterwave (mmWave) Massive MIMO systems and the emerging architectures leveraging Reconfigurable Intelligent Surfaces (RIS), which are central to the evolution of 5G Advanced and 6G wireless networks. The work is motivated by the growing demand for ultra-reliable low-latency communication (URLLC), high spectral efficiency (SE), and massive connectivity, under stringent constraints on hardware cost, power consumption, and computational complexity. The primary objective is to design and evaluate a modular framework that integrates pilot-efficient channel estimation, hardware-constrained hybrid precoding (HP), and low-complexity signal detection suitable for large-scale deployments.

Methodologically, the study combines sparsity-aware modeling, structured signal processing, and iterative inversion-free solvers within a unified and reproducible benchmarking environment based on standardized channel models and system-level simulations. The proposed sparse and learning-assisted channel estimation strategies exploit angular structure and RIS-cascaded channels to substantially reduce pilot overhead while maintaining robustness. Hybrid precoding algorithms built on orthogonal matching pursuit (OMP), alternating minimization (AltMin), and subspace iterative greedy (SIG) achieve favorable performance–complexity trade-offs under constant-modulus and limited-RF hardware constraints. Inversion-free detection schemes leveraging conjugate gradient (CG), Richardson iteration, and successive over-relaxation (SOR) attain near-MMSE accuracy with significantly lower computational and memory requirements.

Collectively, these contributions provide practical insight into bridging the gap between theoretical optimality and hardware feasibility in next-generation transceivers. The findings demonstrate that complexity-aware design enables scalable architectures capable of supporting intelligent surfaces, high-frequency operation, and dense antenna arrays. Future extensions include hardware prototyping, end-to-end learning-based transceiver designs, multi-RIS coordination, and application of the proposed framework to 6G paradigms such as terahertz communication, integrated sensing and communication (ISAC), and cell-free networks.

Keywords: millimeter-wave massive MIMO, RIS, pilot-efficient channel estimation, hybrid analog–digital precoding, inversion-free MMSE detection, sparse and learning-assisted signal processing, complexity-aware physical-layer design, 6G wireless networks.

Résumé

Cette thèse aborde le défi de la conception de traitements de couche physique à la fois évolutifs, sobres en énergie et à hautes performances pour les systèmes MIMO massifs à ondes millimétriques (mmWave), assistés par surfaces intelligentes reconfigurables (RIS), qui constituent des technologies clés dans l'évolution des réseaux 5G-Advanced et 6G. Ce travail est motivé par la croissance des besoins en communications ultra-fiables et à faible latence, en efficacité spectrale élevée et en connectivité massive, dans un contexte de fortes contraintes en termes de coût matériel, de consommation énergétique et de complexité de calcul. L'objectif principal est de concevoir et d'évaluer un cadre modulaire intégrant une estimation de canal économe en pilotes, un précodage hybride soumis aux contraintes matérielles et une détection de symboles à complexité réduite, adaptés aux déploiements à grande échelle.

Sur le plan méthodologique, l'étude combine une modélisation parcimonieuse, un traitement du signal structuré et des solveurs itératifs sans inversion explicite, au sein d'un environnement de banc d'essai unifié et reproductible fondé sur des modèles de canal normalisés et des simulations de niveau système. Les techniques proposées d'estimation de canal parcimonieuse et assistée par apprentissage exploitent la structure angulaire et la nature en cascade des canaux RIS afin de réduire significativement la charge en pilotes tout en préservant la robustesse. Les algorithmes de précodage hybride, fondés sur la poursuite orthogonale (OMP), la minimisation alternée (AltMin) et des méthodes gourmandes itératives en sous-espace (SIG), atteignent des compromis performants entre performance et complexité sous contraintes de modules constants et de nombre limité de chaînes RF. Les schémas de détection sans inversion, basés sur les méthodes du gradient conjugué (CG), de Richardson et de sur-relaxation successive (SOR), obtiennent une précision proche de la détection MMSE classique, avec des besoins en calcul et en mémoire nettement réduits.

Pris dans leur ensemble, ces apports fournissent des éléments concrets pour combler l'écart entre optimalité théorique et faisabilité matérielle dans les transcepteurs de nouvelle génération. Les résultats montrent qu'une conception attentive à la complexité permet de définir des architectures évolutives capables de prendre en charge des surfaces intelligentes, des fréquences élevées et des réseaux d'antennes denses. Parmi les prolongements possibles figurent la réalisation de prototypes matériels, la conception de transcepteurs de bout en bout fondés sur l'apprentissage, la coordination multi-RIS, ainsi que l'extension du cadre proposé à des scénarios 6G tels que les communications térahertz, les systèmes intégrant détection et communication (ISAC) et les réseaux cell-free.

Mots-clés : MIMO massif à ondes millimétriques, surfaces intelligentes reconfigu-

rables, estimation de canal économe en pilotes, précodage hybride analogique–numérique, détection MMSE sans inversion, traitement du signal parcimonieux assisté par apprentissage, conception de couche physique consciente de la complexité, réseaux sans fil 6G.

ملخص

تتناول هذه الأطروحة تحدي تطوير معالجة طبقة فيزيائية قابلة للتوسع، موفرة للطاقة، وذات أداء عالٍ لأنظمة MIMO الضخمة العاملة بالموجات المليمترية (mmWave) والمدعومة بالأسطح الذكية القابلة لإعادة التشكيل (RIS)، والتي تُعد من التقنيات المحورية في تطور شبكات الجيل الخامس المتقدم (5G-Advanced) والجيل السادس (6G). ويأتي هذا العمل استجابةً للطلب المتزايد على الاتصالات فائقة الاعتمادية ومنخفضة الكمون، والكفاءة الطيفية العالية، والاتصال واسع النطاق، وذلك في ظل قيود صارمة تتعلق بتكلفة العتاد، واستهلاك الطاقة، وتعقيد المعالجة. ويهدف هذا البحث إلى تصميم وتقييم إطار معياري يدمج بين تقدير قناة فعال من حيث الإشارات التجريبية، والتشكيل المسبق الهجين المقيد بالعتاد، والكشف منخفض التعقيد الملائم للأنظمة واسعة النطاق.

من الناحية المنهجية، تعتمد الدراسة على نمذجة قائمة على الندرة، ومعالجة إشارات منظمة، وحلول تكرارية لا تتطلب حساب معكوس المصفوفات، ضمن بيئة تقييم موحدة وقابلة لإعادة الإنتاج تعتمد على نماذج قناة معيارية ومحاكاة على مستوى النظام. وتستفيد استراتيجيات تقدير القناة المقترحة، القائمة على الندرة والتعلم، من البنية الزاوية ومن طبيعة القنوات المتسلسلة عبر RIS لتقليل الحمل التجريبي بشكل كبير مع الحفاظ على المتانة. كما تحقق خوارزميات التشكيل المسبق الهجين المبنية على Orthogonal Matching Pursuit (OMP) و Alternating Minimization (AltMin) و Subspace Iterative Greedy (SIG) توازناً فعّالاً بين الأداء والتعقيد تحت قيود العتاد ذي الطور الثابت وعدد محدود من سلاسل RF. إضافةً إلى ذلك، تُظهر مخططات الكشف غير المعتمدة على الانعكاس—باستخدام أساليب مثل Conjugate Gradient (CG)، و Richardson Iteration، و Successive Over-Relaxation (SOR)—دقة قريبة من أداء MMSE التقليدي مع متطلبات حسابية وذاكرية أقل بكثير.

بوجه عام، توفر هذه المساهمات رؤى عملية لسد الفجوة بين المثالية النظرية وقابلية التنفيذ المادي في أنظمة الإرسال والاستقبال من الجيل القادم. وتُظهر النتائج أن التصميم الواعي بالتعقيد يمكن من تطوير معماريات قابلة للتوسع قادرة على دعم الأسطح الذكية، والعمل عند الترددات العالية، وتشغيل مصفوفات هوائيات كثيفة. وتشمل آفاق العمل المستقبلي تطوير نماذج عتادية أولية، وتصميم منظومات إرسال واستقبال تعتمد على التعلم من طرف إلى طرف، والتنسيق متعدد

الأسطح RIS، إضافةً إلى توسيع الإطار المقترح ليشمل سيناريوهات الجيل السادس مثل الاتصالات بترددات التيراهيرتز، والأنظمة المدمجة للاستشعار والاتصال (ISAC)، والحلول الخلوية اللامركزية (Cell-Free Networks).

الكلمات المفتاحية: MIMO الضخم بالموجات المليمترية، الأسطح الذكية القابلة لإعادة التشكيل (RIS)، تقدير قناة منخفض الإشارات التجريبية، التشكيل المسبق الهجين التناظري–الرقمي، الكشف MMSE غير المعتمد على الانعكاس، معالجة الإشارات القائمة على الندرة والمدعومة بالتعلم، تصميم طبقة فيزيائية واعٍ بالتعقيد، شبكات لاسلكية 6G.

Contents

List of Figures XII

List of Abbreviations XX

List of Symbols XXIII

General Introduction 1

1 Background and State of the Art 5

1.1 Introduction . . . 5

1.2 Millimeter-Wave and Massive MIMO Technologies . . . 7

1.2.1 Fundamental Channel Model and Array Responses . . . 7

1.2.2 Capacity Scaling and System-Level Considerations . . . 8

1.2.3 Hybrid Analog–Digital Transceiver Architectures . . . 8

1.2.4 Symbol Detection and Precoding Algorithms . . . 8

1.2.5 Reconfigurable Intelligent Surfaces (RIS) . . . 9

1.3 Antenna Array Processing Fundamentals . . . 10

1.3.1 Array Response and Spatial Filtering . . . 10

1.3.2 Beamforming, Hybrid Precoding and Channel Estimation . . . 11

1.3.3 DOA Estimation and Advanced Algorithms . . . 11

1.3.4 Distributed, Modular, RIS and Lens Arrays . . . 12

1.3.5 Summary and Forward Links . . . 13

1.4 Channel Models and Propagation in Massive MIMO . . . 13

1.4.1 Fundamental Properties: Channel Hardening and Favorable Propagation 14

1.4.2 Composite Channel Models for Massive MIMO and RIS-Aided Systems 14

1.4.3 Symbol Detection and the Importance of Accurate CSI . . . 15

1.5 Review of Detection, Estimation and Precoding Algorithms . . . 15

1.5.1 Detection Algorithms . . . 15

1.5.2 Channel Estimation Methods . . . 16
1.5.3 Precoding Algorithms . . . 16
1.6 Limitations and Research Gaps . . . 17
1.6.1 Complexity and Hardware Bottlenecks . . . 17
1.6.2 Channel Estimation and Feedback Limitations . . . 18
1.6.3 RIS-Aided Systems and Cross-Layer Barriers . . . 18
1.6.4 Summary of Dissertation Contributions . . . 18
1.7 Conclusion . . . 19

2 Efficient Signal Detection for Uplink Massive MIMO **20**
2.1 Introduction . . . 20
2.2 Problem Formulation, Challenges, and Objectives . . . 21
2.2.1 System Model and Notation . . . 21
2.3 Iterative Detection Algorithms . . . 25
2.3.1 Conjugate Gradient (CG) . . . 25
2.3.2 Successive Over-Relaxation (SOR) . . . 27
2.3.3 Hybrid CG–SOR Detector . . . 30
2.3.4 Richardson Iteration with Structured Initialization . . . 31
2.4 Performance and Complexity Analysis . . . 35
2.4.1 Computational Complexity . . . 35
2.4.2 Convergence Behavior and Tuning . . . 36
2.4.3 Initialization Sensitivity and Ranking . . . 37
2.4.4 Simulation Results and Embedded Figures . . . 37
2.5 Implementation Considerations . . . 48
2.5.1 Finite-Precision and Numerical Stability . . . 49
2.5.2 Matrix-Free Dataflow and Memory Efficiency . . . 49
2.5.3 Control Flow and Scheduling . . . 49
2.5.4 Structured Initialization for RI . . . 50
2.5.5 Parameter Tuning Strategies . . . 50
2.5.6 Latency, Complexity, and Parallelism . . . 50
2.5.7 Target Mapping: CPU, GPU, FPGA, ASIC . . . 50
2.5.8 Robustness and System Integration . . . 51
2.5.9 Default Modes . . . 51
2.6 Summary of Contributions . . . 51
2.7 Conclusion . . . 52

3 Enhanced Channel Estimation Techniques **55**
3.1 Introduction . . . 55
3.2 Executive Summary . . . 57
3.3 Channel Estimation in Multiuser Massive MIMO . . . 58
3.4 Direction of Arrival (DOA) Estimation Methods . . . 60
3.4.1 Modified Covariance Matrix Approach . . . 60
3.4.2 Covariance Matrix Shrinkage Estimation . . . 61
3.4.3 MUSIC: Mathematical Derivation . . . 62
3.4.4 Root-MUSIC: Mathematical Derivation . . . 63
3.4.5 Pseudocode: Algorithms . . . 64
3.4.6 Remarks and Practical Guidelines . . . 66
3.5 RIS-Assisted Channel Estimation . . . 67
3.5.1 Cascaded Channel Model . . . 67
3.5.2 Pilot Signaling and CS Measurement Model . . . 67
3.5.3 Dual-Level Structured Sparsity . . . 68
3.5.4 Structured Sparse Recovery: MAP-TLSOMP and TL-SR-OMP . . . 68
3.5.5 Complexity and Metrics . . . 70
3.5.6 Performance and Design Insights . . . 70
3.6 Comparative Analysis and Performance–Complexity Visualization . . . 71
3.6.1 DOA Estimation: Modified Covariance and Shrinkage-Based Methods 71
3.6.2 RIS Channel Estimation: MAP–TLSOMP vs. TL–SR–OMP . . . 80
3.6.3 Noise Robustness and User Scalability . . . 81
3.6.4 Spectral Efficiency Comparison . . . 82
3.6.5 Complexity Visualization and Trade-offs . . . 83
3.6.6 Summary Table: Performance vs. Complexity . . . 84
3.6.7 Discussion . . . 84
3.7 Summary of Contributions . . . 84
3.7.1 Enhanced DOA Estimation via Covariance Reconstruction and Shrinkage 85
3.7.2 Structured Sparse Recovery for RIS-Assisted Channel Estimation . . . 85
3.7.3 Comparative Evaluation and Visualization . . . 86
3.8 Conclusion . . . 87

4 Hybrid Precoding for millimeter-Wave massive MIMO **88**
4.1 Introduction . . . 88
4.2 Executive Summary . . . 89

4.3 Challenges in mmWave Hybrid Beamforming . . . 92
4.4 Orthogonal Matching Pursuit (OMP) and Proposed Variants . . . 94
4.4.1 System and Signal Model . . . 94
4.4.2 Saleh–Valenzuela Channel and Array Responses . . . 95
4.4.3 Spectral Efficiency and Hybrid Objective . . . 96
4.4.4 Dictionary Construction . . . 96
4.4.5 Classical OMP for Hybrid Precoding . . . 96
4.4.6 Proposed Efficient OMP Variant: Sorted Multi-Index Augmentation . . . 97
4.4.7 Wideband and Structured Extensions . . . 99
4.5 Algorithmic Complexity Reduction Mechanisms . . . 99
4.5.1 Subspace–Energy Based Dictionary Pruning . . . 99
4.5.2 Greedy Refinement: Early Stopping and Partial Support Updates . . . 100
4.5.3 Low-Rank Channel Approximations . . . 100
4.5.4 Parallelization Across RF Chains and Subcarriers . . . 100
4.5.5 Runtime Scaling and Indicative Bounds . . . 101
4.5.6 Implementation Notes . . . 101
4.6 Performance Evaluation and Energy Efficiency . . . 101
4.6.1 Metrics and Models . . . 102
4.6.2 Simulation Setup . . . 103
4.6.3 Empirical Results from Proposed Variant . . . 103
4.6.4 Energy-Efficiency Remarks . . . 105
4.6.5 Complexity and Runtime Context . . . 106
4.7 Hardware and Implementation Considerations . . . 106
4.7.1 Phase-shifter resolution and robustness . . . 106
4.7.2 Insertion loss, RF impairments, and calibration . . . 107
4.7.3 Analog fabric selection and architecture trade-offs . . . 107
4.7.4 Real time feasibility on FPGA/ASIC . . . 107
4.7.5 Recommended configuration for the Efficient OMP variant . . . 108
4.8 Summary of Contributions . . . 108
4.9 Conclusion . . . 111

5 System-Level Evaluation and Integration **113**
5.1 Introduction . . . 113
5.2 Executive Summary . . . 113
5.3 Unified Framework for Performance–Complexity Trade-offs . . . 115

5.3.1 Signal-Theoretic Metrics: SE, Fidelity, Reliability . . . 116
5.3.2 Overhead-Aware Throughput and Energy Efficiency . . . 117
5.3.3 Algorithmic Complexity and Latency Accounting . . . 117
5.3.4 Detection Complexity and Convergence (CG/SOR/RI) . . . 118
5.3.5 Mixed-Timescale Integration and Uncertainty Propagation . . . 119
5.3.6 Scalar Utility and Pareto Frontiers . . . 119
5.3.7 Instantiation to Our Contributions and Practical Guidance . . . 120
5.4 Integration of Detection, Estimation, and Precoding Techniques . . . 120
5.4.1 Signal model and hybrid architecture . . . 120
5.4.2 Sparse channel estimation . . . 121
5.4.3 Structured hybrid precoding/combining . . . 121
5.4.4 Low–complexity detection . . . 122
5.4.5 Mixed–timescale orchestration and outcome . . . 122
5.5 Cross-Layer and Practical Implementation Considerations . . . 123
5.5.1 Feedback Compression and Codebook Design . . . 123
5.5.2 Calibration Scheduling and Hardware Impairments . . . 123
5.5.3 Scheduling and Resource Coupling . . . 124
5.6 Simulation Setup and System-Level Performance Metrics . . . 125
5.6.1 Scenarios and Channel Models . . . 125
5.6.2 Baselines and Ablation Studies . . . 125
5.6.3 System-Level Performance Metrics . . . 126
5.7 Comparative Analysis with State-of-the-Art Solutions . . . 127
5.7.1 Spectral and Energy Efficiency Benchmarks . . . 127
5.7.2 Robustness to Hardware and Channel Uncertainty . . . 127
5.7.3 Latency and Precision Mapping . . . 127
5.7.4 Control Overhead and Scheduling Complexity . . . 128
5.8 Summary of System-Level Contributions . . . 128
5.8.1 Modular and Mixed-Timescale PHY Design . . . 129
5.8.2 Complexity-Aware Precoding and Detection . . . 129
5.8.3 System-Level Evaluation and Deployment Scalability . . . 129
5.8.4 Robustness, Overhead, and Deployment Realism . . . 129
5.9 Conclusion . . . 129

Bibliography **135**

List of Figures

1.1 Clustered mmWave channel model and spatial array response geometry. . . . 7
1.2 Hybrid analog–digital BF architecture in mmWave massive MIMO systems. . 9
1.3 RIS-aided mmWave massive MIMO system model. 10
1.4 ULA, UPA, and alternative array geometries supporting spatial filtering and beam formation. 11
1.5 Hybrid analog–digital precoding for mmWave massive MIMO, featuring analog phase shift networks and digital adaptive control. 12
1.6 MUSIC spatial spectrum for multi-angle source estimation. 12
1.7 RIS-aided and lens array signal processing for environment-adaptive, energy-efficient large-scale arrays. 13
1.8 Block diagram of iterative/inversion-free MMSE detection: setup, initialization, update, stopping criterion, output. Each iteration refines $\hat{\mathbf{x}}$ based on matched-filter output and Gram matrix. 16
1.9 Conceptual diagram of CSI acquisition for mmWave/RIS-aided channels: pilot-driven (LS/MMSE), compressive sensing (OMP/ANM), and data-driven (deep learning) paths. RIS control and BS/user terminals are shown for completeness and in alignment with current literature. 17

2.1 System model of the uplink multiuser massive MIMO scenario. The diagram outlines the interaction between U single-antenna users and a base station equipped with N antennas, highlighting the role of channel state information (CSI) in enabling reliable symbol recovery. 22
2.2 Schematic diagram of the MMSE detection process in uplink massive MIMO 23
2.3 Block diagram of the proposed CG–SOR detection method. 30
2.4 Comparative BER performance of MMSE, Jacobi, SOR, CG–SOR, and CG detectors for a 128×16 uplink massive MIMO system with $k = 2$ iterations under 64-QAM modulation. 38

2.5 Comparative BER performance of MMSE, Jacobi, SOR, CG–SOR, and CG detectors for a 128×16 uplink massive MIMO system with $k = 3$ iterations under 64-QAM modulation. 38
2.6 Comparative BER performance of MMSE, Jacobi, SOR, CG–SOR, and CG detectors for a 128×16 uplink massive MIMO system with $k = 4$ iterations under 64-QAM modulation. 39
2.7 Comparative BER performance of MMSE, Jacobi, SOR, CG–SOR, and CG detectors for a 128×32 uplink massive MIMO system with $k = 3$ iterations under 64-QAM modulation. 40
2.8 Comparative BER performance of MMSE, Jacobi, SOR, CG–SOR, and CG detectors for a 128×32 uplink massive MIMO system with $k = 4$ iterations under 64-QAM modulation. 40
2.9 Influence of the relaxation parameter ω on BER performance of the SOR detector for a 128×32 uplink massive MIMO system with $k = 3$ iterations under 64-QAM modulation. Results are shown for multiple SNR values, highlighting the empirical optimum near $\omega \approx 1.2$. 41
2.10 Effect of the relaxation parameter ω on the BER performance of the SOR detector for a 128×32 uplink massive MIMO system with $k = 3$ iterations under 64-QAM modulation at SNR = 18 dB. The curve exhibits a U-shaped trend, with the lowest BER observed near $\omega \approx 1.2$, indicating the empirical optimum relaxation factor. 42
2.11 Comparison of multiplication counts required by different iterative detectors (CG, SOR, JA, and CG–SOR) in a 128×16 uplink massive MIMO system. The bars illustrate how computational load scales with the iteration index k, showing that CG maintains lower complexity while SOR, JA, and CG–SOR incur higher multiplication counts at larger k. 43
2.12 BER performance comparison of the RI detector under different initialization strategies (ZVIn_RI, ADAIn_RI, IDMIn_RI, ITMIn_RI, ISMIn_RI) for a 128×16 uplink massive MIMO system with $k = 2$ iterations and 64-QAM modulation. The curves highlight how initialization choice influences detection accuracy across varying SNR values, with structured methods offering improved convergence relative to simple zero-vector initialization. 44

2.13 BER performance comparison of the RI detector under different initialization strategies (ZVIn_RI, ADAIn_RI, IDMIn_RI, ITMIn_RI, ISMIn_RI) for a 128×16 uplink massive MIMO system with $k = 3$ iterations and 64-QAM modulation. The results demonstrate how initialization choice impacts convergence and detection accuracy across SNR values, with structured approaches outperforming simple zero-vector initialization. 44
2.14 BER performance of the RI detector with multiple initialization strategies (ZVIn_RI, ADAIn_RI, IDMIn_RI, ITMIn_RI, ISMIn_RI) for a 128×16 uplink massive MIMO system using $k = 4$ iterations under 64-QAM modulation. The results highlight how initialization choice influences detection accuracy across SNR values, with structured initialization methods achieving improved convergence compared to the zero-vector baseline. 45
2.15 BER performance of the RI detector with different initialization strategies (ZVIn_RI, ADAIn_RI, IDMIn_RI, ITMIn_RI, ISMIn_RI) for a 128×32 uplink massive MIMO system using $k = 3$ iterations under 64-QAM modulation. The results illustrate how initialization choice affects detection accuracy across SNR values, with structured approaches yielding improved convergence compared to the zero-vector baseline. 46
2.16 BER performance comparison of the RI detector with different initialization strategies (ZVIn_RI, ADAIn_RI, IDMIn_RI, ITMIn_RI, ISMIn_RI) for a 128×32 uplink massive MIMO system using $k = 4$ iterations under 64-QAM modulation. The results illustrate how initialization choice affects detection accuracy across SNR values, with structured approaches yielding improved convergence compared to the zero-vector initialization. 46
2.17 BER performance of the RI detector with different initialization strategies (ZVIn_RI, ADAIn_RI, IDMIn_RI, ITMIn_RI, ISMIn_RI) for a 128×32 uplink massive MIMO system using $k = 5$ iterations under 64-QAM modulation. The results emphasize how initialization choice impacts convergence and detection accuracy across SNR values, with structured approaches consistently outperforming the zero-vector baseline as iteration depth increases. 47

2.18 Multiplication counts of the RI detector under different initialization strategies (ZVIn_RI, IDMIn_RI, ITMIn_RI, ISMIn_RI, ADAIn_RI) for a 128×16 uplink massive MIMO system. The bars show how computational complexity scales with the iteration index k, indicating that structured initialization methods generally incur higher multiplication counts as k increases compared to the zero-vector baseline. 48

3.1 Block diagram of an OFDM communication system showing the interaction between channel estimation and signal detection. 55
3.2 Comparison of pseudospectra for non-coherent sources using the classical Min-Norm and the improved Min-Norm methods. The improved approach yields sharper and more distinct peaks across angles, thereby enhancing resolution and accuracy in DOA estimation compared to the conventional method. 72
3.3 Comparison of pseudospectra for coherent sources using the classical Min-Norm and the improved Min-Norm methods. The improved approach produces sharper and more distinct peaks across angles, thereby enhancing resolution and accuracy in DOA estimation compared to the conventional method. . . . 72
3.4 Effect of array SNR on the pseudospectrum generated by the improved Min-Norm method. Higher SNR values produce sharper and more distinct peaks across angles, thereby improving resolution and reliability in DOA estimation compared to lower SNR conditions. 73
3.5 Effect of the number of array elements on the pseudospectrum generated by the improved Min-Norm method. Increasing the array size sharpens the peaks and enhances angular resolution, thereby improving the accuracy of DOA estimation. 73
3.6 Influence of snapshot count on the pseudospectrum generated by the improved Min-Norm method. Larger numbers of snapshots yield sharper and more stable peaks across angles, enhancing resolution and robustness in DOA estimation compared to cases with fewer snapshots. 74
3.7 Root-domain representation of AoA estimation for non-coherent sources using classical and improved rooting methods. The improved approach aligns roots more distinctly along the unit circle, providing enhanced accuracy and resolution compared to the conventional root-Min-Norm technique. 74

3.8 Root-domain representation of AoA estimation for coherent sources using classical and improved rooting methods. The improved approach positions roots more distinctly along the unit circle, yielding clearer separation of source directions and enhanced estimation accuracy compared to the conventional root-Min-Norm technique. 75
3.9 RMSE performance of the improved Root-Min-Norm vs. SNR for different array sizes. Larger arrays consistently achieve lower RMSE values, highlighting the benefit of increased spatial degrees of freedom in enhancing estimation accuracy. 75
3.10 RMSE performance of the improved Root-Min-Norm vs. SNR for different snapshot counts. Larger numbers of snapshots consistently reduce RMSE across all SNR levels, demonstrating the benefit of increased temporal averaging in enhancing estimation accuracy. 76
3.11 Variation of shrinkage coefficient with the number of snapshots for different estimators at SNR = 20 dB and $N = 10$. As the sample size increases, the shrinkage parameter generally decreases, with the proposed methods showing improved stability and closer alignment to the Oracle benchmark compared to conventional approaches. 76
3.12 Comparison of MVDR-based beamformers under coherent sources at SNR = 25 dB, $N = 10$, and $K = 80$. The proposed estimators yield sharper and more distinct spectral peaks, improving source discrimination and resolution relative to conventional and existing shrinkage-based methods. 77
3.13 Comparison of MVDR-based beamformers under coherent sources at SNR = 25 dB, $N = 10$, and $K = 110$. The proposed estimators achieve sharper and more distinct spectral peaks, enabling clearer separation of source directions and improved resolution compared to conventional and existing shrinkage-based methods. 77
3.14 RMSE performance of multiple DOA estimators vs. SNR for $N = 12$ and $K = 90$. The proposed methods consistently achieve lower error across all SNR levels, with the tuned Chen variant showing the best accuracy, highlighting its robustness in high-noise conditions. 78

3.15 DOA estimation results using MVDR with tuned Chen shrinkage under varying SNR levels. Higher SNR values produce sharper and more distinct spectral peaks, improving angular resolution and source separation compared to lower SNR conditions. 78
3.16 DOA estimation results using MVDR with tuned Chen shrinkage across different array configurations. Larger arrays yield sharper and more distinct spectral peaks, improving angular resolution and source separation compared to smaller configurations. 79
3.17 RMSE performance of MVDR with tuned Chen shrinkage vs. SNR for different array sizes M. Larger arrays consistently achieve lower error across all SNR levels, demonstrating the advantage of increased spatial degrees of freedom in improving estimation accuracy. 79
3.18 DOA estimation results using MVDR with tuned Chen shrinkage for different snapshot counts. Increasing K produces sharper and more stable spectral peaks, demonstrating the benefit of additional temporal samples in improving angular resolution and estimation robustness. 80
3.19 RMSE performance of MVDR with tuned Chen shrinkage vs. SNR for varying sample sizes K. Higher snapshot counts generally reduce estimation error across SNR levels, illustrating the benefit of temporal averaging in improving robustness and accuracy. 80
3.20 NMSE vs. pilot length Q at SNR $= -10$ dB: TL–SR–OMP dominates in pilot-limited regimes. 81
3.21 NMSE vs. pilot length Q at SNR $= 10$ dB: MAP–TLSOMP yields the best NMSE across Q. 81
3.22 NMSE performance of different channel estimation algorithms vs. SNR for $Q = 90$ and $U = 20$: MAP–TLSOMP lowest NMSE; TL–SR–OMP close. . . 82
3.23 NMSE performance of different channel estimation algorithms as the number of users increases for $Q = 90$ and SNR $= 5$ dB. The TL–SR–OMP method maintains graceful scalability with growing user counts, while MAP–TLSOMP consistently achieves the highest accuracy, outperforming conventional baselines. 82
3.24 ASE performance versus SNR for different channel estimation algorithms. The structure-aware recovery approaches (MAP–TLSOMP and TL–SR–OMP) consistently achieve higher efficiency across the SNR range, approaching the Oracle benchmark and outperforming conventional baselines. 83

3.25 Computational complexity comparison of channel estimation algorithms as the number of users increases. While MAP–TLSOMP achieves high accuracy, it incurs significantly greater runtime overhead, whereas TL–SR–OMP maintains scalability with lower complexity, making it more practical for large-scale systems. 83

4.1 Fully-connected (top) vs. sub-connected (bottom) HP architectures. Fully-connected offers maximal flexibility; sub-connected reduces RF chain count and hardware cost while achieving improved EE with comparable spectral performance. 89
4.2 Hybrid BF system for multi-user mmWave transmission. Digital BB precoder outputs are mapped through RF chains and phase shifters to antennas. . . . 90
4.3 E2E HP and combining system. Includes RF and BB components, channel estimation, and CSI-driven optimization. 91
4.4 Multi-stage hybrid BF process: beam training, analog RF mapping, and digital BB coordination, highlighting how system constraints propagate through each stage. 93
4.5 SE vs. SNR for different precoding algorithms under UPA configuration. The proposed Efficient OMP variant approaches the fully-digital benchmark and outperforms all hybrid and analog baselines. 103
4.6 SE vs. SNR for different precoding algorithms under ULA configuration. Performance trends mirror those in UPA, but SE values are lower due to reduced spatial resolution. 104
4.7 SE vs. number of RF chains at fixed SNR = -5 dB. The proposed variant matches fully-digital performance when $N_{\mathrm{RF}} \geq 7$. 105

5.1 Mixed-timescale processing timeline (slow spatial statistics, fast CSI/precoding/detection). 116
5.2 Hybrid analog fabrics and component counts for power modeling: fully-connected, sub-connected, and switch-based. 116
5.3 CG–SOR detection pipeline and representative convergence/complexity curves. 118
5.4 RI with informed initialization (ISMIn/tridiagonal) and early stopping. . . . 118
5.5 Double-structured sparsity in cascaded BS–RIS–UE channels (shared rows, partially shared columns); training/control overlays for pilot-efficient estimation. 119

5.6 Hybrid analog fabrics and feedback pathways: fully-connected, sub-connected, and switch-based architectures, with codebook overlays. 123
5.7 RIS control overlays and calibration scheduling: RIS over-the-air setup. . . . 124

List of Abbreviations

4G	4th Generation
5G	5th Generation
6G	6th Generation
3GPP	3rd Generation Partnership Project
ADMM	Alternating Direction Method of Multipliers
AI	Artificial Intelligence
AltMin	Alternating Minimization
AMP	Approximate Message Passing
ANM	Atomic Norm Minimization
AoA	Angle of Arrival
AoD	Angle of Departure
ASE	Area Spectral Efficiency
ASIC	Application-Specific Integrated Circuit
AWGN	Additive White Gaussian Noise
BB	Baseband
BER	Bit Error Rate
BF	Beamforming
BLER	Block Error Rate
BS	Base Station
CG	Conjugate Gradient
CNN	Convolutional Neural Network
CS	Compressive Sensing
CSI	Channel State Information
DFT	Discrete Fourier Transform
DL	Downlink
DNN	Deep Neural Network
DOA	Direction of Arrival

DS-OMP	Double-Structured Orthogonal Matching Pursuit
eMBB	enhanced Mobile Broadband
E2E	End-to-End
EE	Energy Efficiency
ELAA	Extremely Large Aperture Array
ESPRIT	Estimation of signal parameters via rotational invariant techniques
FC	Fully Connected
FPGA	Field-Programmable Gate Array
GD	Gradient Descent
GS	Gauss–Seidel
GPU	Graphics Processing Unit
HPD	Hermitian Positive Definite
HP	Hybrid Precoding
HWI	Hardware Impairments
i.i.d.	independent and identically distributed
ISMIn	Inverse Stair Matrix Initialization
ISAC	Integrated Sensing and Communication
IoT	Internet of Things
JA	Jacobi Algorithm (detector)
KPI	Key Performance Indicator
LCMV	Linearly Constrained Minimum Variance
LLR	Log-Likelihood Ratio
LMMSE	Linear Minimum Mean-Square Error
LOS	Line-of-Sight
LS	Least Squares
LTE	Long-Term Evolution
MAC	Medium Access Control
MF	matched filter
MIMO	Multiple-Input Multiple-Output
ML	Machine Learning
MMSE	Minimum Mean-Square Error
mMTC	Massive Machine-Type Communications
MRC	Maximum Ratio Combining
MUSIC	Multiple Signal Classification

MVDR	Minimum Variance Distortionless Response
NLOS	Non-Line-of-Sight
NMSE	Normalized Mean-Square Error
NR	New Radio (5G NR)
OAS	Oracle Approximating Shrinkage
OFDM	Orthogonal Frequency-Division Multiplexing
OMP	Orthogonal Matching Pursuit
PER	Packet Error Rate
PGD	Projected Gradient Descent
PL	Path Loss
PS	Phase Shifter
QAM	Quadrature Amplitude Modulation
QoS	Quality of Service
RBLW	Rao-Blackwell Ledoit-Wolf (shrinkage estimator)
RF	Radio Frequency
RI	Richardson Iteration
RIS	Reconfigurable Intelligent Surface
SD	Steepest Descent
SE	Spectral Efficiency
SIG	Subspace Iterative Greedy
SINR	Signal-to-Interference-plus-Noise Ratio
SBL	Sparse Bayesian Learning
SNR	Signal-to-Noise Ratio
SOMP	Simultaneous Orthogonal Matching Pursuit
SOR	Successive Over-Relaxation
SVD	Singular Value Decomposition
SV	Saleh–Valenzuela (channel model)
THz	Terahertz Frequencies
ULA	Uniform Linear Array
UPA	Uniform Planar Array
URLLC	Ultra-Reliable Low-Latency Communication
VLSI	Very-Large-Scale Integration
ZF	Zero Forcing
ZVIn	Zero Vector Initialization

List of Symbols

H	Channel matrix
H_{RIS}	RIS-assisted cascaded channel matrix
H_{BR}	Base station to RIS channel
H_{RU}	RIS to user channel
K	Number of users
U	Number of users
N_t	Number of transmit antennas
N_r	Number of receive antennas
N	Number of antennas
N_{RF}	Number of RF chains
N_{cl}	Number of channel clusters
$N_{\mathrm{ray},c}$	Number of rays in cluster c
$\alpha_{c,r}$	Complex gain of ray r in cluster c
$\mathbf{a}(\theta)$	Array response vector at angle θ
$\mathbf{a}_t(\cdot)$	Transmit array response vector
$\mathbf{a}_r(\cdot)$	Receive array response vector
F	Precoding matrix
F_{RF}	Analog RF precoding matrix
F_{BB}	Digital baseband precoding matrix
$\mathbf{F}_{\mathrm{opt}}$	Optimal fully-digital precoder

N_s	Number of spatial data streams
$\mathbf{s}$	Transmitted symbol vector (dimension $N_s \times 1$ or $U \times 1$)
$\mathbb{E}\{\mathbf{s}\mathbf{s}^H\}$	Symbol covariance
$\mathbf{y}$	Received signal vector
$\mathbf{x}$	Transmitted/estimated symbol vector
$\mathbf{y}_{\text{MF}}$	Matched filter output
W	MMSE filtering matrix
$\mathbf{G}_f$	Gram / MMSE matrix at subcarrier f (e.g., $\mathbf{H}_f^H\mathbf{H}_f + \sigma^2\mathbf{I}$)
$\mathbf{z}_f$	Matched-filter vector at subcarrier f (e.g., $\mathbf{H}_f^H\mathbf{y}_f$)
$\hat{\mathbf{s}}_{\text{MMSE}}$	MMSE detector output
$\hat{\mathbf{s}}^{(k)}$	Detection estimate at iteration k
$\hat{s}_m^{(k)}$	m-th component of the estimate at iteration k
$\mathbf{n}$	AWGN noise vector
σ_n^2	Noise variance per receive antenna
σ^2	Noise variance (per-stream detection)
$\mathbf{r}^{(k)}$	Residual vector at iteration k
$\mathbf{p}^{(k)}$	CG search direction at iteration k
$\alpha^{(k)}$	CG step size at iteration k
$\beta^{(k)}$	CG conjugacy parameter at iteration k
γ_k	CG step size at iteration k (alternate notation)
$\mathcal{J}(\mathbf{s})$	Quadratic cost function in CG method
$\langle \mathbf{a}, \mathbf{b} \rangle$	Inner product (usually $\mathbf{a}^H\mathbf{b}$)
$\Re\{\cdot\}$	Real part operator
$\|\mathbf{x}\|_{\mathbf{W}}$	$\mathbf{W}$-weighted Euclidean norm $\sqrt{\mathbf{x}^H\mathbf{W}\mathbf{x}}$
κ	Condition number of $\mathbf{W}$
$\lambda_{\max}(\mathbf{W})$	Largest eigenvalue of $\mathbf{W}$

$\lambda_{\min}(\mathbf{W})$	Smallest eigenvalue of $\mathbf{W}$
$\lambda(\mathbf{W})$	Eigenvalue set of $\mathbf{W}$
$\rho(\cdot)$	Spectral radius operator
$\mathbf{r}$	Residual (generic)
$\mathbf{p}$	Search direction (generic)
ε	Stopping tolerance (residual threshold)
I	Identity matrix
D	Diagonal matrix in SOR splitting (diag($\mathbf{W}$))
L	Strictly lower triangular matrix (in splitting)
L^H	Strictly upper triangular matrix
w_{mu}	Entry (m, u) of matrix $\mathbf{W}$
$y_{\mathrm{MF},m}$	m-th element of matched filter output
ω	SOR relaxation parameter
$\omega^\star$	Optimal RI relaxation (constant minimizing spectral radius)
$\hat{\omega}_0(\psi)$	Quasi-optimal SOR relaxation as function of loading ψ
ψ	Loading factor, ratio N/U (antennas per user)
$\mathbf{B}_\omega$	SOR iteration matrix
$\mathbf{e}$	Constant term in SOR fixed-point iteration
$\mathbf{B}_\omega^{\mathrm{RI}}$	RI iteration matrix
$\mathbf{c}_\omega^{\mathrm{RI}}$	Constant vector in RI iteration
$\mathbf{e}^{(k)}$	Error vector at iteration k
ρ_k	Residual norm at iteration k
$\mathbf{w}$	Beamformer weight vector
$\mathbf{w}_{\mathrm{MVDR}}$	MVDR beamformer weight vector
R	Spatial covariance matrix
$\widehat{\mathbf{R}}$	Sample covariance matrix

$\widehat{\mathbf{R}}_{\text{FB}}$	Forward–Backward averaged covariance matrix
$\widehat{\mathbf{R}}_{\text{CR}}$	Conjugate reconstructed covariance matrix
$\widehat{\mathbf{R}}_{\text{sh}}$	Shrinkage covariance matrix
ρ_{sh}	Shrinkage coefficient (intensity)
$\mathbf{F}_{\text{sh}}$	Shrinkage target matrix (structured identity/Toeplitz)
$\mathbf{J}$	Exchange (reversal) matrix for ULA (antidiagonal ones)
T	Number of snapshots (time samples)
M	Number of array sensors / RIS receive antennas (context dependent)
$\mathbf{x}[t]$	Array observation vector at time t (snapshot)
$\mathbf{A}$	Steering matrix of sources; also used as dictionary of steering vectors
L	Dictionary size (number of atoms)
$\mathbf{R}$	Residual matrix in OMP iterations (context dependent)
$\mathcal{S}$	Support set of selected atoms (OMP)
$\mathbf{g}$	Correlation matrix in OMP
$k^\star$	Index of selected atom (OMP)
Λ_i	Selected atom indices at iteration i (OMP variant)
w_k	Correlation score for atom k
$\mathbf{b}$	Sorted index vector of correlation scores
$\mathbf{U}_r$	Left singular vectors in truncated SVD
$\mathbf{\Sigma}_r$	Singular value matrix in truncated SVD
$\mathbf{V}_r$	Right singular vectors in truncated SVD
r	Truncated rank
r_ε	Rank threshold for energy tolerance
σ_i	Singular values
ε	Energy tolerance parameter (SVD truncation)
$\mathcal{L}_{\text{proj}}$	Index set from subspace projection criterion

$\mathcal{L}_{\text{eng}}$	Index set from energy criterion
$\mathcal{L}$	Final pruned index set (dictionary pruning)
τ_{proj}	Projection threshold for pruning
τ_{eng}	Energy threshold for pruning
$\mathbf{R}_i$	Normalized residual at iteration i (OMP)
Δ_i	Marginal gain at iteration i (OMP)
$\bar{b}$	Average size of updated LS block
r_h	Effective channel rank in low-rank approximation
λ_k	Singular value in low-rank channel approximation
$\mathbf{u}_k$	Left singular vector in low-rank channel approximation
$\mathbf{v}_k$	Right singular vector in low-rank channel approximation
$\mathbf{a}_{\text{ULA}}(\phi)$	ULA steering vector (explicit ULA formula)
$\mathbf{a}_{\text{UPA}}(\phi, \theta)$	UPA steering vector (explicit UPA formula)
d	Inter-element spacing (often $d = \lambda/2$)
λ	Signal wavelength
$\mathbf{H}_f$	Channel matrix at subcarrier f (OFDM tone)
$\mathbf{F}_f$	Hybrid precoder at subcarrier f ($\mathbf{F}_{\text{RF}}\mathbf{F}_{\text{BB},f}$)
$\mathbf{W}_f$	Hybrid combiner at subcarrier f ($\mathbf{W}_{\text{RF}}\mathbf{W}_{\text{BB},f}$)
ρ	Average SNR scaling factor (used in precoding/SE expressions)
R_f	Spectral efficiency at subcarrier f (bits/s/Hz)
$\bar{R}$	Wideband average rate $\frac{1}{F}\sum_{f=1}^{F} R_f$
$\det(\cdot)$	Determinant operator
$\mathbf{G}$	BS–RIS channel matrix (alternate notation)
$\mathbf{h}_{r,u}$	RIS–user channel vector for user u
$\mathbf{H}_u$	Cascaded BS–RIS–user channel matrix for user u
$\mathbf{U}_M$	DFT dictionary matrix for BS (angular basis)

$\mathbf{U}_N$	DFT dictionary matrix for RIS (angular basis)
$\widetilde{\mathbf{H}}_u$	Sparse angular-domain cascaded channel matrix for user u
L_G	Number of dominant BS–RIS paths
$L_{r,u}$	Number of dominant RIS–user paths for user u
$\boldsymbol{\theta}_q$	RIS phase vector at slot q (unit-modulus)
$\boldsymbol{\Theta}$	RIS training matrix (phase configurations)
Q	Number of RIS training slots / RIS configurations
$\mathbf{y}_{u,q}$	Received signal for user u at slot q
$\mathbf{Y}_u$	Stacked received signals for user u across Q slots
$\mathbf{Z}_u$	Noise matrix during RIS training
$\widetilde{\mathbf{Y}}_u$	Angular-domain measurement matrix (MMV form)
$\widetilde{\boldsymbol{\Theta}}$	Transformed RIS training matrix in angular domain
$\widetilde{\mathbf{Z}}_u$	Angular-domain noise matrix
$\mathcal{I}_r$	Common row support set (BS–RIS structure)
$\mathcal{I}_\ell^{(u)}$	Column support set for row ℓ and user u
$\mathcal{I}_\ell^{\mathrm{Com}}$	Common column support subset across users
$\mathcal{I}_\ell^{\mathrm{Usr},u}$	User-specific column support subset for user u
L_c	Size of common column support
P_{total}	Total system power (component-wise model)
P_{tx}	Transmit RF power (radiated)
P_{RFch}	Power per RF chain (analog front-end)
N_{PS}	Number of phase shifters in the analog fabric
P_{PS}	Power consumed per phase shifter
N_{SW}	Number of switches (switch-based fabrics)
P_{SW}	Power consumed per switch
P_{BB}	Baseband processing power (static + dynamic)

P_{hybrid}	Total power consumption of hybrid architecture
P_{RF}	Power consumed per RF chain (alternate notation)
η_{oh}	Normalized overhead fraction (pilots, feedback, guards, RIS control)
η_{pilot}	Pilot overhead fraction
η_{feedback}	Feedback overhead fraction
η_{guard}	Guard/CP overhead fraction
η_{RIS}	RIS control overhead fraction
BLER_u	Block error rate for user u
R_u	Achievable rate for user u (bits/s/Hz)
EE	Energy efficiency (bits/s/Hz/W)
C_{corr}	Approximate cost of correlations (structured OMP)
C_{LS}	Approximate cost of least-squares updates (partial LS)
C_{ref}	Approximate cost of subspace refinement (power-method steps)
T_{pm}	Number of power-method steps for subspace refinement
$\text{Latency}_{\text{alg}}$	Algorithmic compute latency
$\text{Latency}_{\text{DMA/mem}}$	DMA / memory transfer latency
$\text{Latency}_{\text{RF switch}}$	RF switching latency
$\text{Latency}_{\text{cal}}$	Calibration window latency
$\mathcal{U}_f$	CSI uncertainty set at subcarrier f (e.g., $\{\mathbf{E} : \|\mathbf{E}\| \leq \varepsilon_f\}$)
τ_f	Threshold used for pruning / stopping at subcarrier f (uncertainty-aware)
$\mathcal{U}$	Scalar utility combining normalized SE, Latency, EE, Complexity
$\alpha, \beta, \gamma, \delta$	Non-negative weighting coefficients in utility $\mathcal{U}$
$\mathbf{W}_{\text{RF}}$	Analog combining matrix (RF combiner)
$\mathbf{W}_{\text{BB},f}$	Digital baseband combiner at subcarrier f
$\widehat{\mathbf{E}}_n$	Estimated noise subspace eigenvectors
$\widehat{\mathbf{P}}_n$	Noise projector matrix

$P_{\text{MUSIC}}(\theta)$	MUSIC pseudospectrum
$Q(z)$	Root-MUSIC polynomial
z	Complex variable in Root-MUSIC polynomial
c_ℓ	Polynomial coefficients from noise projector diagonals
$\widehat{\theta}_k$	Estimated DOA angle from Root-MUSIC
τ_{proj}	Projection threshold for dictionary pruning
τ_{eng}	Energy threshold for dictionary pruning
μ	Step size (gradient / adaptive algorithms)
$\overline{\mathcal{J}}_m$	Regularized MSE cost at iteration m (adaptive updates)
F	Number of OFDM subcarriers (context dependent)
T_{slow}	Slow-timescale update period (frames)
T_{fast}	Fast-timescale frame duration (frames)

General Introduction

The evolution of wireless communication systems from fourth-generation (4G) Long Term Evolution (LTE) to fifth-generation (5G) New Radio (NR), and toward the anticipated sixth generation (6G), has introduced unprecedented demands on SE, EE, latency, reliability, and scalability [1, 2, 3]. Future networks are expected to support URLLC, massive machine-type communications (mMTC), intelligent edge/cloud infrastructures, and artificial intelligence (AI)-driven automation across the Internet of Things (IoT) [4]. These complex and often conflicting requirements call for transformative physical-layer architectures and signal processing methodologies rather than incremental extensions of current systems [5, 6].

mmWave frequencies, typically spanning 30–300 GHz, offer large unused spectrum resources enabling multi-gigabit data rates and dense spatial reuse [7, 8]. However, communication at such high frequencies introduces major propagation challenges, including severe free-space loss, blockage sensitivity, limited diffraction, and sparse multipath environments [1, 9]. Massive multiple-input multiple-output (MIMO) systems utilizing large-scale antenna arrays address these limitations through highly directional beamforming (BF) and spatial multiplexing [5, 6]. Nevertheless, implementing fully digital massive MIMO architectures at mmWave remains constrained by power consumption, hardware cost, and calibration requirements associated with allocating a dedicated RF chain to every antenna element [10, 11].

Hybrid analog–digital architectures provide an attractive compromise between performance and hardware feasibility [12, 13]. In these systems, a reduced number of RF chains feed large antenna arrays via analog phase-shifting networks, while digital baseband (BB) processing handles multiuser transmission and adaptive shaping. Although hybrid MIMO systems significantly reduce RF hardware complexity, they give rise to new algorithmic challenges, including constrained precoder design, limited RF connectivity structures, and finite resolution phase control, all of which must be addressed under stringent latency and energy budgets [14, 15].

RIS have recently emerged as a disruptive paradigm capable of transforming the wireless propagation medium into an intelligent controllable entity [16, 17]. RIS panels composed

of reconfigurable passive or semi-passive reflecting elements can manipulate incident signals through digitally programmable phase shifts. When properly configured, RIS can enhance coverage, enable Non-Line-of-Sight (NLoS) links, improve SE, and suppress interference, all while consuming minimal energy [18, 19]. However, integrating RIS introduces additional complexity related to cascaded channel modeling, configuration overhead, calibration, and joint optimization between the base station (BS) and the RIS [20, 21].

Research Challenges. Despite significant progress in these enabling technologies, several bottlenecks continue to hinder scalable deployment of mmWave and RIS enhanced massive MIMO systems:

- **Channel State Information (CSI) Acquisition:** Sparse and dynamic propagation environments challenge classical LS/MMSE channel estimators [22]. Although compressive sensing (CS) and deep learning approaches reduce pilot overhead [23, 24], they often lack robustness and generalization under hardware impairments (HWI), model mismatch, or rapid mobility [25, 26].
- **HP:** Designing precoders under constant modulus and hardware constraints remains a difficult non-convex problem [12, 13]. Existing methods based on OMP, AltMin, and SIG improve scalability but are sensitive to CSI imperfections and computationally demanding for large antenna arrays or multiuser scenarios [27].
- **Symbol Detection:** Classical zero-forcing (ZF)/MMSE detection requires matrix inversion, which becomes computationally expensive at large array dimensions [28, 29]. Iterative inversion-free solvers such as CG, RI, and SOR offer attractive alternatives but require systematic tuning for accuracy, convergence speed, and efficient hardware implementation [15, 30].
- **Benchmarking and Integration:** Existing studies typically evaluate modules in isolation, using heterogeneous assumptions and metrics [11]. Unified, reproducible frameworks for E2E performance, latency, and EE are missing, making it difficult to quantify performance, complexity trade-offs or to map algorithmic advances into deployable system designs [13, 31].

Research Aim and Objectives. The overarching aim of this dissertation is to design and evaluate a modular, complexity-aware physical-layer framework for mmWave and RIS-

assisted massive MIMO systems. To achieve this aim, the research pursues the following objectives:

- Develop pilot efficient and robust CSI acquisition methods leveraging channel sparsity, angular structure, and data-driven processing.
- Design HP techniques based on OMP, AltMin, and SIG algorithms that explicitly account for constant modulus constraints, limited RF chains, and practical hardware architectures.
- Propose inversion-free MMSE detection schemes using iterative solvers (CG, RI, SOR) with adaptive stopping criteria and complexity-aware parameter tuning.
- Establish unified benchmarking metrics and simulation frameworks that capture Normalized Mean-Square Error (NMSE), bit error rate (BER), runtime, and support recovery accuracy under consistent channel and hardware assumptions.
- Validate the proposed methods under standardized channel models that reflect practical HWI and RIS-aided propagation, and assess their impact on E2E spectral and EE.

Significance and Contributions. This dissertation makes contributions to both theoretical development and practical implementation, summarized as follows:

- A unified low-complexity detection framework based on iterative inversion-free solvers with performance approaching that of classical MMSE detection [28, 29].
- New sparse and learning assisted channel estimation frameworks tailored to mmWave and RIS-aided systems [23, 24].
- HP structures and algorithms that integrate OMP, AltMin, and SIG approaches, optimized for hardware limited architectures [12, 13].
- A reproducible benchmarking environment incorporating accuracy metrics (NMSE, BER,SE), runtime profiling, and complexity accounting [11, 15].
- System level validation under realistic modeling assumptions using standardized channel models and representative deployment scenarios [32, 33].

Scope and Limitations. The work focuses strictly on physical-layer design, implementation, and evaluation for mmWave and RIS-enabled massive MIMO systems. While the proposed approaches are compatible with higher layer extensions and real time scheduling frameworks, network-level optimization, large-scale distributed multi-RIS orchestration, and full hardware prototyping remain outside the present scope. The numerical results rely primarily on standardized and widely used channel models and on simulation-based validation; experimental verification on hardware testbeds is left as an avenue for future research.

This dissertation is organized as follows:

- **Chapter 1: Background and State of the Art** reviews mmWave and massive MIMO technologies, antenna array processing, channel models, and existing detection, estimation, and precoding algorithms, and identifies key research gaps.
- **Chapter 2: Efficient Signal Detection for Uplink massive MIMO** develops and analyzes low-complexity inversion-free MMSE detection schemes based on CG, RI, and SOR iterations, with particular emphasis on complexity, performance trade-offs and hardware oriented tuning.
- **Chapter 3: Enhanced Channel Estimation Techniques** introduces structured, pilot efficient channel estimation methods for both conventional and RIS-aided mmWave MIMO, combining CS, sparse Bayesian inference, and data-driven enhancements.
- **Chapter 4: Hybrid Precoding for millimeter-Wave massive MIMO** proposes and evaluates HP workflows based on OMP, AltMin, and SIG, addressing hardware constraints, subspace selection, and robustness to CSI errors.
- **Chapter 5: System-Level Evaluation and Integration** integrates the proposed detection, estimation, and precoding modules into a unified physical-layer pipeline, and benchmarks E2E performance, latency, and complexity under standardized scenarios.
- **General Conclusion and Future Work** summarizes the main findings, reflects on the limitations of the current study, and outlines directions for extending the research toward experimental validation, cross-layer design, and large-scale RIS-enabled deployments.

Chapter 1

Background and State of the Art

1.1 Introduction

The evolution of wireless communication systems from 4G LTE to 5G NR and toward 6G has introduced stringent requirements across multiple dimensions: SE, EE, latency, reliability, and scalability [1, 2, 3]. These demands are driven by a new generation of applications, including URLLC, mMTC, intelligent edge/cloud services, and AI-powered automation within the IoT [2, 4]. Meeting these requirements compels a rethinking of physical-layer architectures, signal processing algorithms, and system level integration strategies.

mmWave frequencies, spanning 30–300 GHz, unlock abundant spectrum and enable high throughput, short range connectivity. However, mmWave propagation is susceptible to severe path loss (PL), reduced scattering, and high temporal variation [1, 5, 9]. When combined with massive MIMO deploying tens to hundreds of antennas mmWave systems exploit spatial multiplexing and adaptive BF to overcome these challenges and achieve unprecedented capacity. Yet, they introduce their own complexity bottlenecks and hardware limitations [6, 34, 35]. Fully digital massive MIMO at mmWave is often infeasible due to RF-chain costs, power consumption, and calibration constraints, motivating the adoption of hybrid analog–digital architectures [10, 11, 15, 36]. These hybrid systems balance performance and complexity by performing beam steering in the analog domain and flexible multiplexing in the digital BB, but require efficient algorithmic design under hardware imposed constraints.

The recent emergence of RIS further expands the wireless design space by introducing programmable electromagnetic environments [16, 17, 37]. RIS panels, composed of passive or hybrid elements, can reflect and intelligently shape incident signals to enhance coverage, improve reliability for NLOS links, and increase spectral reuse [18, 19]. However, RIS-aided

networks introduce new research challenges in channel estimation, control signaling, calibration, and distributed optimization, especially in scenarios with channel sparsity, user mobility, and feedback constraints [38, 39].

Despite these advances, several deployment bottlenecks remain. mmWave and RIS-aided channels reveal highly sparse and dynamic multipath structures, necessitating pilot efficient, robust, and scalable channel estimation methods [23, 39]. HP must judiciously balance SE and EE under constant modulus constraints and limited RF hardware, often within non-convex, hardware limited optimization spaces [12, 36]. Detection algorithms must withstand imperfect CSI, HWI, quantization, and tight latency requirements. More broadly, cross-layer coordination including overhead efficient feedback, self calibration, and real time scheduling emerges as a core challenge for large-scale, reproducible system realization [11].

Complexity theory and algorithmic scalability form a major part of the contemporary research frontier. Recent advances have yielded iterative, inversion-free alternatives to traditional matrix computations in detection, estimation, and HP such as CG, RI, and SOR-type solvers [28, 29, 36, 40]. These methods can approach optimal performance at a fraction of the computational and memory cost, allow aggressive early stopping, and are amenable to hardware parallelism and low-precision implementations [13, 15, 36]. Nonetheless, balancing performance with latency, energy, and system integration remains as critical as ever.

This dissertation addresses these multidimensional design challenges by proposing a modular, complexity-aware physical-layer pipeline for mmWave and RIS-enhanced massive MIMO systems. The developed framework integrates structured channel estimation, scalable HP, and robust inversion-free detection under realistic propagation and hardware constraints. System level evaluation is conducted using standardized channel models, state of the art baselines, and reproducible metrics. Each innovation is systematically validated and contextualized in subsequent chapters.

This chapter is organized as follows: Section 1.2 introduces millimeter-wave and massive MIMO technologies, highlighting fundamental models, hybrid architectures, and detection/precoding strategies. Section 1.3 presents antenna array processing fundamentals, including array geometries, beamforming, DOA estimation, and RIS/lens-based extensions. Section 1.4 discusses channel models and propagation properties in massive MIMO and RIS-aided systems, emphasizing channel hardening, favorable propagation, and composite modeling approaches. Section 1.5 reviews detection, estimation, and precoding algorithms, covering classical, inversion-free, and learning-based methods. Finally, Section 1.6 identifies key limitations and research gaps, motivating the contributions developed in subsequent chapters.

1.2 Millimeter-Wave and Massive MIMO Technologies

mmWave bands, ranging from 30 to 300 GHz, offer vast unused spectrum resources essential for meeting the explosive demand for wireless data throughput in 5G and beyond (6G) networks [1, 2, 34]. These frequencies enable multi-gigabits per second connectivity but require overcoming substantial physical challenges: high free-space PL increasing as the square of the carrier frequency, poor diffraction, blockage sensitivity, and a typically sparse multipath environment [7, 8].

Massive MIMO exploits the spatial degrees of freedom provided by deploying very large antenna arrays (hundreds or more elements) at the BS to achieve spatial multiplexing gains and BF gains that compensate for mmWave impairments [2, 5, 6]. Under favorable propagation conditions, the effective channel dimension supports simultaneous service to multiple users, enhancing spectrum usage and network capacity.

1.2.1 Fundamental Channel Model and Array Responses

The mmWave channel can be modeled by the sparse Saleh–Valenzuela (SV) clustered channel model as in Equation (1.10), where the BB channel matrix is a superposition of limited clusters, each consisting of few propagation rays. The array response vectors $\mathbf{a}(\theta)$ for ULAs or UPAs are critical for BF design. These vectors are analytically expressed and typically utilize an antenna spacing of $d = \lambda/2$ to prevent spatial aliasing [2].

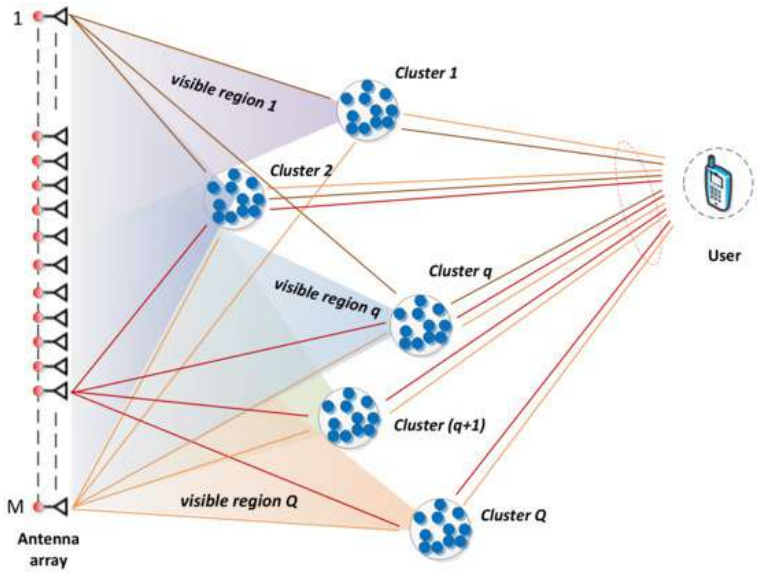

Figure 1.1: Clustered mmWave channel model and spatial array response geometry.

1.2.2 Capacity Scaling and System-Level Considerations

In massive MIMO, under Rayleigh fading and channel hardening assumptions, the achievable sum-rate capacity for K users served by N_t antennas scales approximately as:

$$C_{\text{sum}} \approx K \log_2 \left(1 + \frac{P_{\text{total}}}{K\sigma_n^2}(N_t - K)\right), \tag{1.1}$$

where C_{sum} is the achievable sum-rate capacity, K is the number of users, N_t is the number of transmit antennas, P_{total} is the total transmit power, and σ_n^2 is the noise variance. The approximation holds in the regime $N_t \gg K$ [5]. This fundamental scaling motivates the use of large antenna arrays to enhance SE and EE, balanced against channel estimation overhead and hardware costs [2].

1.2.3 Hybrid Analog–Digital Transceiver Architectures

Implementing fully digital BF at mmWave frequencies with massive antenna counts N_t is constrained by power consumption, cost, and RF hardware complexity. Hybrid architectures split digital BB processing and analog RF BF, reducing the number of costly RF chains $N_{\text{RF}} \ll N_t$ [10, 11]. The total precoder is expressed as:

$$\mathbf{F} = \mathbf{F}_{\text{RF}}\mathbf{F}_{\text{BB}}, \tag{1.2}$$

where $\mathbf{F}_{\text{RF}}$ is implemented by phase shifters constrained by constant modulus and $\mathbf{F}_{\text{BB}}$ handles digital multiplexing [13].

The hardware complexity and energy consumption reduction are captured approximately by:

$$\text{Power}_{\text{hybrid}} \approx N_{\text{RF}}P_{\text{RF}} + N_t P_{\text{PS}}, \tag{1.3}$$

where $\text{Power}_{\text{hybrid}}$ is the total power consumption of the hybrid architecture, N_{RF} is the number of RF chains, P_{RF} is the power consumed per RF chain, and P_{PS} per phase shifter [11], significantly lower than fully-digital systems with N_t RF chains.

1.2.4 Symbol Detection and Precoding Algorithms

Symbol detection algorithms must efficiently solve the task of estimating transmitted data symbols in high dimensional MIMO channels with limited complexity. Iterative methods such as CG, RI, and SOR offer inversion-free, scalable solutions that are robust to channel

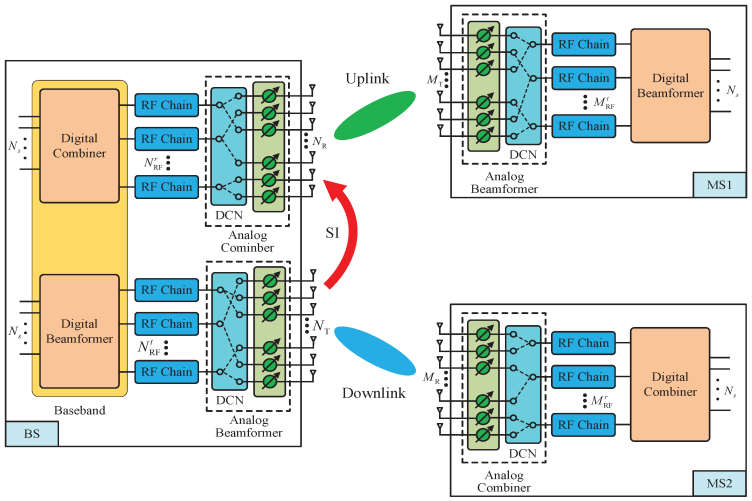

Figure 1.2: Hybrid analog–digital BF architecture in mmWave massive MIMO systems.

size and suitable for real time hardware implementation [28, 29]. These solvers are discussed in detail in Chapter 2.

HP is often realized using sparse recovery techniques like OMP and Bayesian methods, balancing the sparsity inherent in mmWave channels with hardware constraints [13, 23]. Detailed algorithmic treatments appear in Chapter 4.

1.2.5 Reconfigurable Intelligent Surfaces (RIS)

RIS technology integrates low-power, passive (or semi-passive) metasurfaces into the environment to shape wireless propagation dynamically [16, 17]. The E2E composite channel matrix from BS to user including RIS effects is modeled as:

$$\mathbf{H}_{\text{RIS}} = \mathbf{H}_{\text{RU}} \mathbf{\Phi} \mathbf{H}_{\text{BR}}, \tag{1.4}$$

where $\mathbf{H}_{\text{BR}}$ models the BS-to-RIS channel, $\mathbf{H}_{\text{RU}}$ models the RIS-to-user channel, and $\mathbf{\Phi} = \text{diag}(e^{j\phi_1}, \ldots, e^{j\phi_M})$ encodes the scattering phase shifts of the M RIS elements [20, 21, 37]. This cascaded structure leads to highly sparse, low-rank effective channels, enabling programmable spatial environments. Channel estimation and joint precoding optimization in RIS systems pose new challenges and are actively researched in Chapter 3 and Chapter 4.

This section provides the fundamental physical-layer and system level context that underpins subsequent chapters on channel estimation (Chapter 3), symbol detection and iterative solvers (Chapter 2), and HP algorithm design (Chapter 4), as well as system simulations and complexity analysis in Chapter 5.

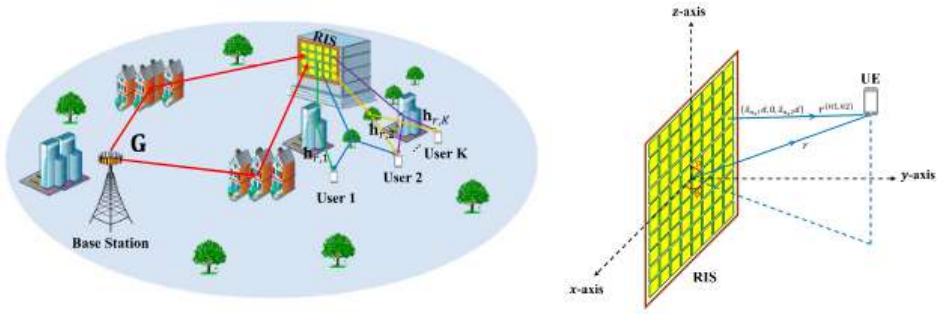

Figure 1.3: RIS-aided mmWave massive MIMO system model.

1.3 Antenna Array Processing Fundamentals

Antenna array processing forms the theoretical backbone for spatial filtering, adaptive BF, and direction of arrival (DOA) estimation in contemporary wireless systems [41, 42, 43]. Canonical array geometries, including the uniform linear array (ULA), uniform planar array (UPA), and more recently sparse, thinned, and modular configurations, support the trade-off between spatial resolution and hardware complexity [6, 44]. Typically, inter-element spacing is set to $\lambda/2$ to maximize beam directivity and suppress grating lobes, though distributed and conformal arrangements are also gaining importance in 5G/6G and extremely large aperture array (ELAA) contexts [44, 45, 46].

1.3.1 Array Response and Spatial Filtering

The array response vector of an N-element ULA at direction θ is:

$$\mathbf{a}(\theta) = \frac{1}{\sqrt{N}} \begin{bmatrix} 1 & e^{j\frac{2\pi}{\lambda}d\sin\theta} & \dots & e^{j(N-1)\frac{2\pi}{\lambda}d\sin\theta} \end{bmatrix}^T, \tag{1.5}$$

where d is inter-element spacing and λ is the carrier wavelength.

The beamformer output is

$$y = \mathbf{w}^H \mathbf{x}, \tag{1.6}$$

where $\mathbf{w}$ is the weight vector and $\mathbf{x}$ the received array signal. Classical designs include delay and sum, linearly constrained minimum variance(LCMV), and the minimum variance distortionless response (MVDR) beamformer, which maximizes signal to noise ratio (SNR) subject to response constraints [41, 42]:

$$\mathbf{w}_{\text{MVDR}} = \frac{\mathbf{R}^{-1}\mathbf{a}(\theta_0)}{\mathbf{a}^H(\theta_0)\mathbf{R}^{-1}\mathbf{a}(\theta_0)}, \tag{1.7}$$

where $\mathbf{R}$ is the spatial covariance matrix.

Different array geometries linear, circular, planar, conformal are illustrated in recent literature, e.g., [45].

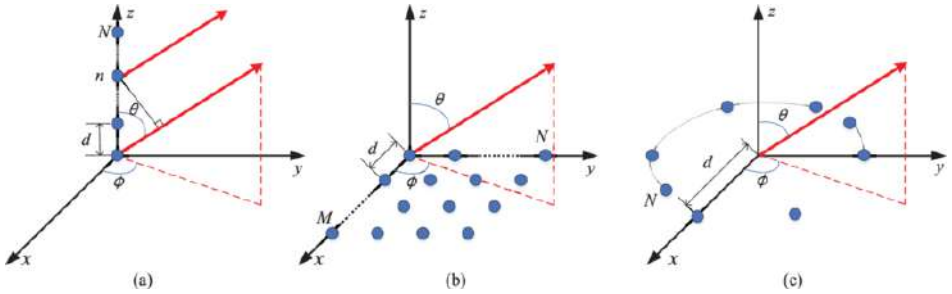

Figure 1.4: ULA, UPA, and alternative array geometries supporting spatial filtering and beam formation.

1.3.2 Beamforming, Hybrid Precoding and Channel Estimation

In mmWave and massive MIMO, analog RF BF using networks of phase shifters or switches constrained to constant modulus serves as a front-end for digital BB precoding implemented per subcarrier [10, 12, 47]. Hybrid analog–digital architectures, and increasingly deep learning-aided approaches, achieve close to optimal performance with realistic complexity and hardware constraints [48]. Current surveys, such as [43, 45], comprehensively illustrate analog, digital, and hybrid BF block diagrams together with performance trade-offs and practical feeding/coupler networks (Wilkinson, Butler matrix, Branchline).

Iterative and machine learning (ML)-enabled algorithms now address symbol detection, channel estimation, and HP for spatially sparse, high-dimensional arrays, as in [48, 49]. Modular processing, as addressed in [50], further enables distributed calibration and scalable system implementation for ELAA and next-generation 6G scenarios.

1.3.3 DOA Estimation and Advanced Algorithms

DOA estimation is vital for array gain and spatial selectivity. Classical and advanced algorithms include multiple signal classification (MUSIC), estimation of signal parameters via rotational invariance techniques (ESPRIT), Bayesian, and compressive-sensing-based estimators, enabling super-resolution multipath analysis and user/device localization [23, 41, 44, 51]. Performance plots for these methods (e.g., MUSIC spatial spectrum) are widely available [45].

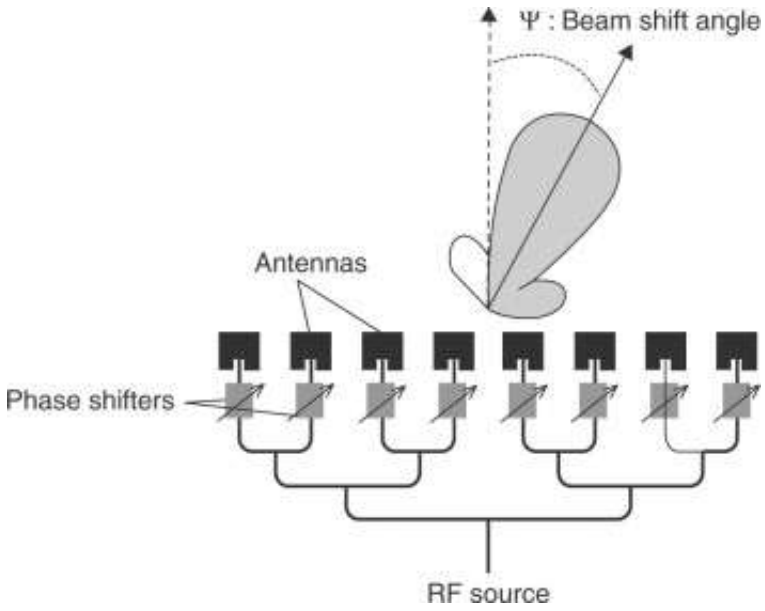

Figure 1.5: Hybrid analog–digital precoding for mmWave massive MIMO, featuring analog phase shift networks and digital adaptive control.

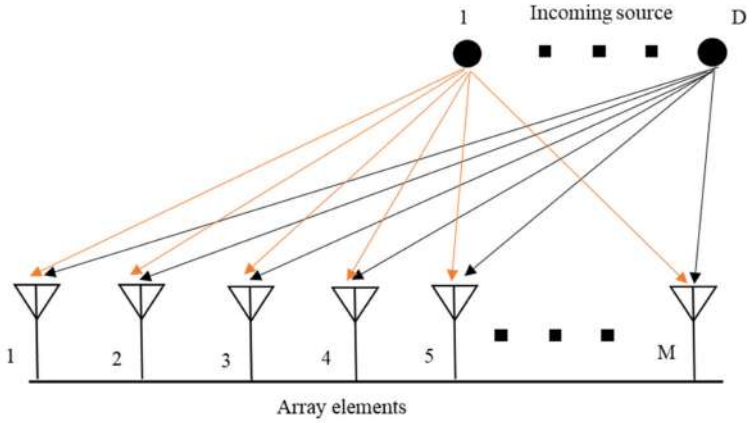

Figure 1.6: MUSIC spatial spectrum for multi-angle source estimation.

1.3.4 Distributed, Modular, RIS and Lens Arrays

Contemporary trends extend array processing via distributed, cell-free, or ELAA configurations [50, 52], addressing interconnect and power constraints with system level distributed BB, RIS and lens arrays now provide programmable reflection or focused beamspace channelization boosting coverage, EE and environmental adaptation [16, 46, 53]. Recent reviews [45, 54] offer annotated diagrams for RIS and lens-aided spatial focusing, including massive MIMO, mmWave/THz, and holographic architectures.

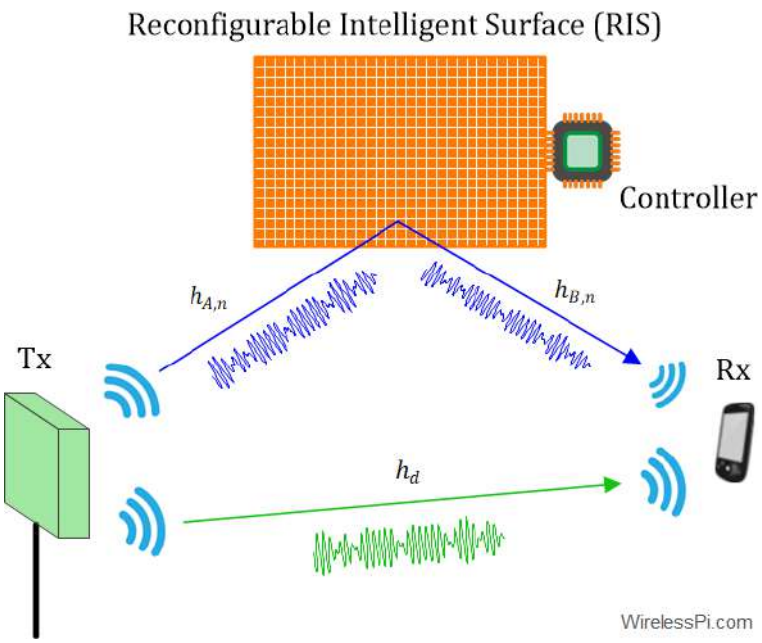

Figure 1.7: RIS-aided and lens array signal processing for environment-adaptive, energy-efficient large-scale arrays.

1.3.5 Summary and Forward Links

This section has surveyed core and emerging array processing principles, algorithms, and architectures spanning mathematical models, BF, hybrid design, advanced DOA estimation, and system level innovations such as RIS and lens integration. These fundamentals enable and motivate advanced channel estimation (Chapter 3), scalable detection/precoding (Chapter 2), and spatially aware, programmable wireless networks (Chapter 4, Chapter 5).

1.4 Channel Models and Propagation in Massive MIMO

Channel modeling lies at the heart of realistic analysis and robust design for massive MIMO and RIS-aided systems, especially as wireless networks transition to the mmWave domain [32, 33]. Understanding the key propagation properties, as well as precise mathematical models, is crucial for accurate CSI acquisition, reliable symbol detection, and overall system performance.

1.4.1 Fundamental Properties: Channel Hardening and Favorable Propagation

Two fundamental phenomena define massive MIMO channels: *channel hardening* and *favorable propagation*. *Channel hardening* refers to the phenomenon where, as the number of antennas N increases, the effective channel for each user becomes nearly deterministic that is, the variance of the channel gain decreases and the channel behaves almost like an Additive White Gaussian Noise (AWGN) channel [5, 6, 55]:

$$\frac{1}{N}|\mathbf{h}_k|^2 \to \mathbb{E}[|h|^2], \quad N \to \infty \tag{1.8}$$

where $\mathbf{h}_k$ is the channel vector for user k, N is the number of antennas, and $\mathbb{E}[|h|^2]$ is the expected channel power. As N grows large, the channel norm becomes nearly deterministic.

This hardening effect simplifies resource allocation, power control, and signal detection.

Favorable propagation refers to the nearly orthogonal channels between users in large antenna systems, which enables massive spatial multiplexing and simple linear detection/precoding schemes to approach their theoretical optimality [5, 6]:

$$\frac{1}{N}\mathbf{h}_k^{\mathrm{H}}\mathbf{h}_j \to 0, \quad k \neq j, ; N \to \infty \tag{1.9}$$

where $\mathbf{h}_k$ and $\mathbf{h}_j$ are the channel vectors for users k and j. This property implies that user channels become asymptotically orthogonal.

This orthogonality underpins the scalability of massive MIMO for dense user and device deployments.

1.4.2 Composite Channel Models for Massive MIMO and RIS-Aided Systems

A widely used composite channel model for mmWave and massive MIMO scenarios is the clustered SV model:

$$\mathbf{H} = \sqrt{\frac{N_t N_r}{N_{\mathrm{cl}} N_{\mathrm{ray}}}} \sum_{c=1}^{N_{\mathrm{cl}}} \sum_{r=1}^{N_{\mathrm{ray},c}} \alpha_{c,r}, \mathbf{a}r(\theta c, r^{\mathrm{AoA}}), \mathbf{a}t^{H}(\theta c, r^{\mathrm{AoD}}) \tag{1.10}$$

Here, N_t and N_r denote the numbers of transmit and receive antennas, N_{cl} is the number of scattering clusters, $N_{\mathrm{ray},c}$ is the number of propagation rays within cluster c, $\alpha_{c,r}$ is the complex gain of ray (c, r), $\theta c, r^{\mathrm{AoA}}$ and $\theta c, r^{\mathrm{AoD}}$ are the AoA and AoD of ray (c, r), and $\mathbf{a}_t(\cdot)$ and $\mathbf{a}_r(\cdot)$ denote the transmit and receive array response vectors [33].

1.4.3 Symbol Detection and the Importance of Accurate CSI

For reliable symbol detection in massive MIMO and RIS-assisted systems, robust and timely CSI is pivotal. Inaccurate or outdated CSI degrades the performance of both conventional (e.g., linear MMSE, ZF, MRC) and modern (iterative, AI-based) detector algorithms, potentially causing error floors or dramatic increases in BER under mobility or model mismatch [28, 29].

Parametric, compressed sensing, and geometry-based modeling approaches have therefore arisen to support efficient and robust CSI acquisition, especially vital in systems with large antenna counts or RIS integration, where pilot and feedback overhead are major bottlenecks [33]. Advanced channel models directly enable scalable estimation and detection frameworks, forming the focus of later chapters.

1.5 Review of Detection, Estimation and Precoding Algorithms

Modern massive MIMO and mmWave systems demand robust, scalable algorithms for detection, channel estimation, and precoding, especially under stringent hardware and deployment constraints. Advances in both theory and implementation have enabled high-dimensional signal recovery with reduced complexity and improved reliability, particularly relevant in hybrid and RIS-aided architectures [23, 24, 28, 29, 48].

1.5.1 Detection Algorithms

Detection in massive MIMO benefits from channel hardening and favorable propagation but is challenged by high matrix inversions and HWI linear approaches such as MMSE and ZF are widely adopted for their simplicity, with MMSE detection yielding the solution:

$$\hat{\mathbf{x}}_{\text{MMSE}} = (\mathbf{H}^H\mathbf{H} + \sigma_n^2\mathbf{I})^{-1}\mathbf{H}^H\mathbf{y} \tag{1.11}$$

where $\hat{\mathbf{x}}_{\text{MMSE}}$ is the estimated transmitted symbol vector, $\mathbf{H}$ is the channel matrix, $\mathbf{y}$ is the received signal vector, σ_n^2 is the noise variance, and $\mathbf{I}$ is the identity matrix.

However, in large-scale systems, direct matrix inversion is impractical. Inversion-free and iterative solvers such as CG, RI, and SOR enable scalable detection, support early stopping, and allow for mixed-precision hardware mapping while retaining near-optimal perfor-

mance [28, 29, 30]. These techniques are also instrumental in robust, low-latency detectors for hardware constrained or RIS-enhanced settings.

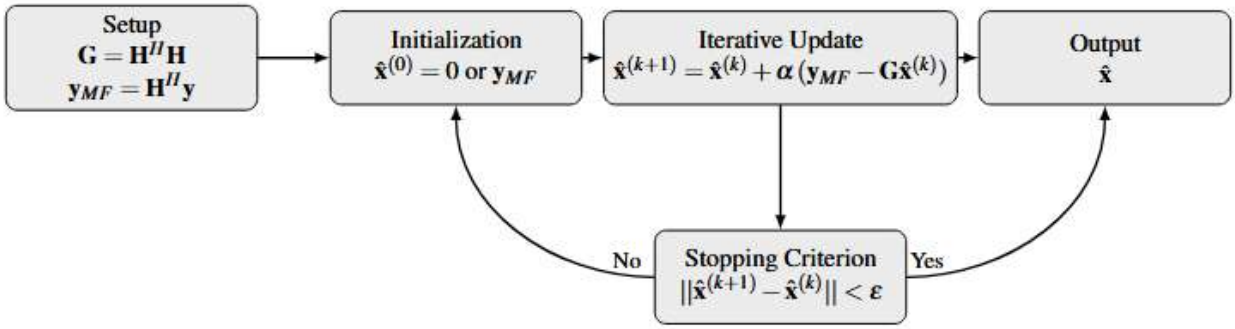

Figure 1.8: Block diagram of iterative/inversion-free MMSE detection: setup, initialization, update, stopping criterion, output. Each iteration refines $\hat{\mathbf{x}}$ based on matched-filter output and Gram matrix.

1.5.2 Channel Estimation Methods

CSI is the backbone of detection and precoding in massive MIMO. Classical least squares (LS) techniques are attractive for their simplicity but are susceptible to pilot contamination, noise, and scalability limitations [22]. Covariance-aided minimum mean square error (MMSE) estimators offer improved accuracy, especially when channel statistics are available, but at increased computational cost [30].

Recent advances exploit the intrinsic sparsity of massive MIMO and mmWave/RIS-aided channels. CS, parametric, and tensor-based strategies, such as atomic norm minimization (ANM) and structured Bayesian sparse recovery, enable pilot efficient and robust estimation, even with mobility and hardware constraints [23, 24, 56, 57]. Deep learning models (DNNs, CNNs), either as E2E estimators or as priors for hybrid approaches, further enhance estimation in non-ideal, dynamic environments [30, 48].

1.5.3 Precoding Algorithms

Precoding transforms multiuser interference into spatial gain. Fully-digital linear methods (MMSE/ZF) yield state-of-the-art performance but are hampered by high RF chain and hardware costs in mmWave. Analog-only solutions (beam steering) are less flexible. Hybrid digital-analog architectures combine the benefits, using a limited number of RF chains with analog beamformers [10, 12, 58].

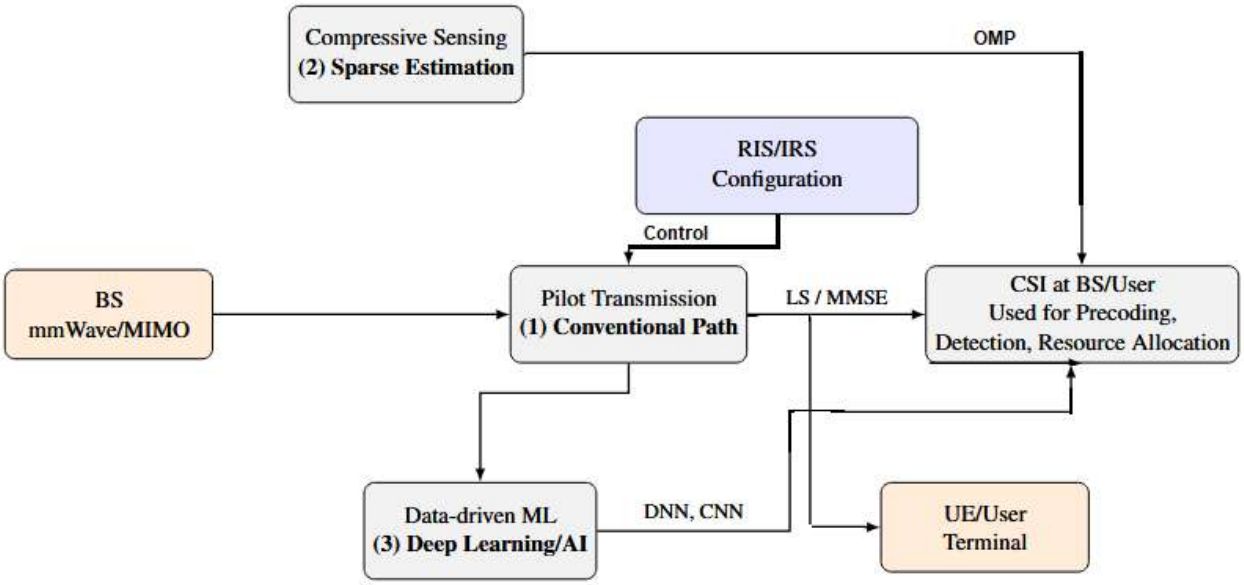

Figure 1.9: Conceptual diagram of CSI acquisition for mmWave/RIS-aided channels: pilot-driven (LS/MMSE), compressive sensing (OMP/ANM), and data-driven (deep learning) paths. RIS control and BS/user terminals are shown for completeness and in alignment with current literature.

Greedy and AltMin approaches provide performance-complexity balance. OMP–based schemes, structured for hardware and EE, are central to modern mmWave and RIS systems. Subspace pruning and iterative LS refinements (partial LS) yield complexity controlled solutions for real time systems [13, 31, 58]. Recent contributions leverage deep learning, with CNN-based or meta-learned codebooks, outperforming classical designs in dynamic or hardware impairment scenarios [48, 59, 60, 61].

1.6 Limitations and Research Gaps

Recent advances in mmWave massive MIMO, HP, and RIS-aided architectures have substantially improved wireless system capacity, adaptability and efficiency [1, 27, 34, 48]. Nonetheless, several critical challenges must be addressed to realize truly scalable, robust, and EE deployments.

1.6.1 Complexity and Hardware Bottlenecks

Fully digital and hybrid systems continue to face high computational costs and hardware barriers, especially for real time implementations with large scale antenna arrays [28, 29].

Although algorithms such as CG, RI, and SOR enable inversion-free processing, comprehensive analysis linking complexity, accuracy, and power consumption under HWI is still lacking. HP strategies based on AltMin, OMP, or deep learning often involve iterative, non-convex searches whose practical optimization remains challenging, especially under strict RF chain limits and quantization effects [12, 13, 27]. Furthermore, the absence of unified benchmarking standards complicates cross system comparisons and reproducibility.

1.6.2 Channel Estimation and Feedback Limitations

Accurate CSI is essential for high performance detection and precoding, but practical systems struggle with sparse, time varying multipath and limited feedback [23, 24, 62]. Classical estimation methods (LS, MMSE) assume stationary, rich scattering, while modern deployments show dynamic and non-ideal conditions. Compressive sensing and tensor-based methods reduce pilot overhead but rely on sparsity or prior knowledge that may be absent in rapidly changing scenarios [63]. Data-driven and deep learning approaches for CSI inference, though promising, can suffer from poor generalization, especially when training data fails to reflect the diversity and non-idealities of hardware or real wireless channels [25, 26].

1.6.3 RIS-Aided Systems and Cross-Layer Barriers

RIS deployments bring novel opportunities as well as challenges in physical modeling and system integration. Most models neglect critical aspects including mutual element coupling, calibration drift, and limited scalability [20, 64]. Distributed control, phase synchronization, and real time feedback remain underexplored, with most experimental work limited to small-scale environments.

System level integration across detection, estimation, and precoding is rarely performed in a unified framework. Algorithms are usually developed in isolation, and few studies provide E2E evaluation addressing latency, runtime and energy metrics [11, 27, 48]. Reproducibility across platforms and realistic deployment scenarios is not yet standard, hindering technology transfer and fair performance comparison.

1.6.4 Summary of Dissertation Contributions

This dissertation addresses the above limitations by proposing and validating a reproducible, complexity-aware framework for mmWave and RIS-enhanced massive MIMO. Key contributions include:

- Fast, inversion-free MMSE detection using CG, RI, and SOR, with early stopping and hardware-aware parameter adaptation.
- Pilot efficient, structured CSI acquisition via sparse, compressive, and data-driven estimators under RIS constraints.
- HP architectures leveraging OMP, AltMin, and SIG with subspace pruning and modular analog-digital mapping.
- Unified E2E benchmarking using standardized models, reproducible metrics, and system level integration with hardware constraints.

These advances are systematically evaluated throughout the dissertation, demonstrating their impact on scalable, next-generation wireless networks.

1.7 Conclusion

This chapter has established the foundational technologies, models, and algorithmic paradigms that underpin the dissertation's contributions. Through a structured review of mmWave and massive MIMO architectures, antenna array processing fundamentals, realistic channel modeling, and state-of-the-art detection, estimation, and precoding algorithms, both the central opportunities and persistent limitations of next-generation wireless systems have been contextualized.

Key insights include the promise of hybrid analog–digital transceivers for scalable mmWave deployment, the transformative potential of RIS in shaping programmable propagation environments, and the emergence of inversion-free solvers and sparse recovery techniques for complexity-aware signal processing. This overview highlighted critical bottlenecks in channel estimation, hardware constrained precoding, and system-level reproducibility, particularly under dynamic, sparse, and feedback limited conditions.

By identifying these limitations and research gaps, the chapter sets the stage for the modular, reproducible, and hardware-aware frameworks to be developed in subsequent chapters. These frameworks encompass structured CSI acquisition, HP workflows, inversion-free detection algorithms, unified benchmarking methods, and comprehensive system level integration. Collectively, these contributions enable scalable and complexity-aware design for mmWave and RIS-enhanced massive MIMO systems. The forthcoming chapters present, validate, and evaluate these innovations in detail, culminating in a cohesive framework for next-generation wireless networks.

Chapter 2

Efficient Signal Detection for Uplink Massive MIMO

2.1 Introduction

Massive MIMO technology has emerged as a cornerstone of next-generation wireless systems, offering significant gains in SE, reliability, and energy savings. In uplink transmission, however, the computational burden associated with optimal signal detection, particularly matrix inversion in MMSE filters, poses a major challenge for scalability.

As the number of antennas at the BS increases, conventional detection algorithms become impractical due to their high complexity and memory requirements. This chapter addresses the need for low-complexity detection schemes that maintain near-optimal performance while reducing computational overhead.

We focus on iterative methods such as CG, SOR and RI with optimized initialization strategies. These algorithms are evaluated in terms of BER, convergence speed, and arithmetic complexity.

This chapter is organized as follows: Section 2.2 formulates the uplink massive MIMO detection problem, outlines computational challenges, and defines the research objectives. Section 2.3 introduces inversion-free iterative detection algorithms, including Conjugate Gradient (CG), Successive Over-Relaxation (SOR), and Richardson Iteration (RI) with structured initialization. Section 2.4 provides comparative performance evaluations and complexity analyses of the proposed methods. Section 2.5 discusses implementation considerations, including MATLAB simulation insights and hardware feasibility. Section 2.6 summarizes the key contributions of this chapter. Finally, Section 2.7 concludes the chapter with remarks on

findings and implications for scalable uplink detection in massive MIMO systems.

Executive Summary

This chapter presents a suite of efficient iterative detection algorithms for uplink massive MIMO systems. The proposed methods aim to approximate MMSE performance while significantly reducing computational complexity.

Key contributions include:

- A hybrid CG-SOR detector that combines Krylov subspace acceleration with relaxation-based convergence, where CG generates an accurate initial solution refined by SOR in final iterations.
- A RI framework enhanced by five initialization strategies to accelerate convergence, including the novel ISMIn and ADAIn techniques.
- A comparative analysis of BER performance and multiplication count across all methods.
- Implementation insights for MATLAB-based simulation and hardware feasibility.

These contributions are supported by two publications:

- Journal article: *Efficient Iterative Detection Based on Conjugate Gradient and Successive Over-Relaxation Methods for Uplink Massive MIMO Systems* [29].
- Conference paper: *Low-complexity and Efficient Massive MIMO Signal Detection Based on RI Method with Various Initialization Techniques* [28].

2.2 Problem Formulation, Challenges, and Objectives

2.2.1 System Model and Notation

We consider the uplink of a multiuser massive MIMO system in which a BS equipped with N antennas serves U single antenna users, typically with $N \gg U$. Let $\mathbf{s} \in \mathbb{C}^{U\times 1}$ denote the transmitted symbol vector, $\mathbf{y} \in \mathbb{C}^{N\times 1}$ the received signal at the BS, and $\mathbf{H} \in \mathbb{C}^{N\times U}$ the uplink channel matrix. Under a flat-fading, independent and identically distributed (i.i.d). Rayleigh model and AWGN with variance σ_n^2 per receive antenna, the received signal is modeled as:

$$\mathbf{y} = \mathbf{H}\mathbf{s} + \mathbf{n}, \qquad \mathbf{n} \sim \mathcal{CN}(\mathbf{0}, \sigma_n^2 \mathbf{I}_N). \tag{2.1}$$

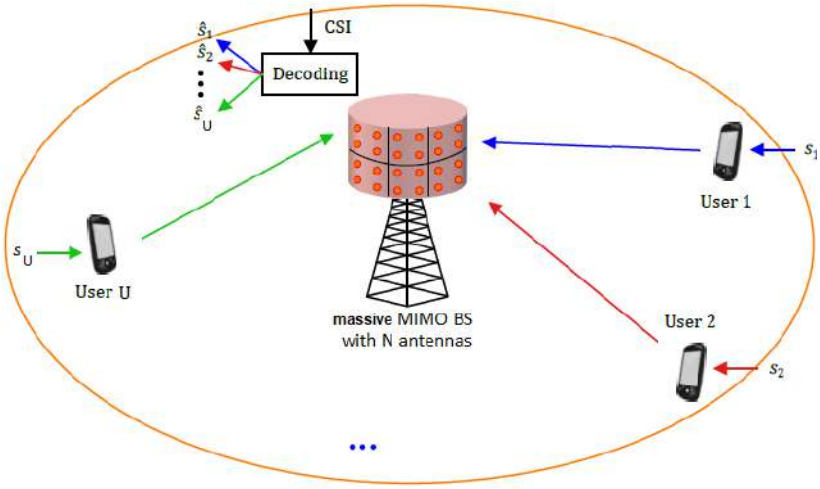

Figure 2.1: System model of the uplink multiuser massive MIMO scenario. The diagram outlines the interaction between U single-antenna users and a base station equipped with N antennas, highlighting the role of channel state information (CSI) in enabling reliable symbol recovery.

As shown in Figure 2.1, the schematic highlights the multiuser uplink scenario where the BS simultaneously receives signals from U users through its N antennas. This visualization clarifies how the transmitted symbols are mapped through the channel matrix $\mathbf{H}$ and subsequently decoded using CSI. It emphasizes the central role of the BS antenna array in separating user streams and sets the stage for the mathematical formulation in (2.1).

Define the MF output and the MMSE filtering matrix as:

$$\mathbf{y}_{\text{MF}} \triangleq \mathbf{H}^H \mathbf{y}, \qquad \mathbf{W} \triangleq \mathbf{H}^H \mathbf{H} + \sigma^2 \mathbf{I}_U, \tag{2.2}$$

where σ^2 denotes the per-stream detection noise parameter. The linear MMSE estimate is given by:

$$\hat{\mathbf{s}}_{\text{MMSE}} = \mathbf{W}^{-1} \mathbf{y}_{\text{MF}} = \left(\mathbf{H}^H \mathbf{H} + \sigma^2 \mathbf{I}_U \right)^{-1} \mathbf{H}^H \mathbf{y}. \tag{2.3}$$

where $\hat{\mathbf{s}}_{\text{MMSE}} \in \mathbb{C}^{U \times 1}$ denotes the MMSE detector output.

While (2.3) achieves excellent detection performance, explicitly inverting or factorizing $\mathbf{W}$ incurs cubic-time complexity in U, which impedes scalability in large multiuser deployments [65, 66].

Figure 2.2 illustrates the computational pipeline of the MMSE detector. The diagram makes explicit the sequence of operations matched filtering, conjugate multiplication, matrix inversion, and equalization that transform the received signal into symbol estimates. This block view underscores why matrix inversion is the computational bottleneck, motivating

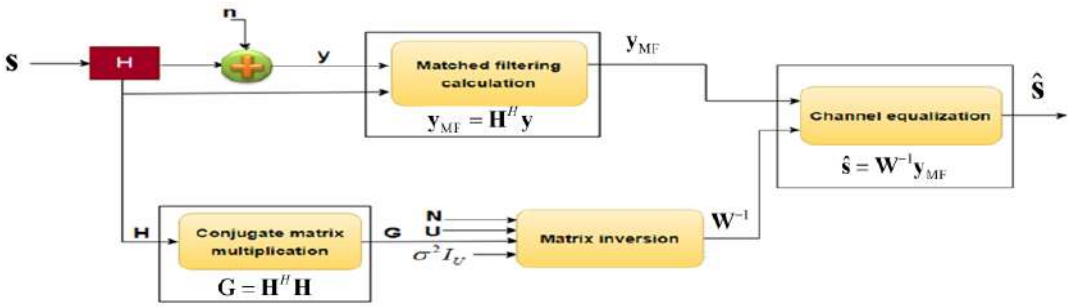

Figure 2.2: Schematic diagram of the MMSE detection process in uplink massive MIMO

the iterative algorithms introduced later in this chapter to achieve near-MMSE performance with reduced complexity.

Computational Challenges

The core bottleneck lies in solving the Hermitian positive definite (HPD) system $\mathbf{W}\hat{\mathbf{s}} = \mathbf{y}_{\text{MF}}$. In large-scale scenarios, direct inversion or dense factorization leads to:

- **Arithmetic complexity:** Inversion scales as $O(U^3)$, while matrix–vector products scale as $O(U^2)$ [65].
- **Memory traffic and locality:** Dense matrix operations stress memory bandwidth and cache locality [66].
- **Latency and energy:** Real time constraints and power budgets are challenged by cubic-time kernels [65].

These constraints motivate inversion-free, iterative solvers that rely on matrix–vector products with $\mathbf{W}$ and simple vector updates.

Iterative Formulations and Design Knobs

We target iterative updates of the form:

$$\hat{\mathbf{s}}^{(k+1)} = \mathcal{F}\Big(\hat{\mathbf{s}}^{(k)};\, \mathbf{y}_{\text{MF}}, \mathbf{W}, \boldsymbol{\theta}\Big), \tag{2.4}$$

where $\hat{\mathbf{s}}^{(k)} \in \mathbb{C}^{U\times 1}$ is the detection estimate at iteration k, $\mathcal{F}(\cdot)$ denotes the iterative update mapping, $\mathbf{y}_{\text{MF}}$ and $\mathbf{W}$ are as defined above, and $\boldsymbol{\theta}$ collects algorithmic parameters such as step sizes and relaxation factors. Three complementary families are explored:

- **Conjugate Gradient (CG):** A Krylov-subspace solver with per-iteration cost dominated by one $\mathbf{W}\mathbf{p}^{(k)}$ product. CG is widely used for large-scale MIMO detection due to its favorable convergence properties [65, 66].
- **SOR:** Based on the splitting $\mathbf{W} = \mathbf{D} + \mathbf{L} + \mathbf{L}^H$, SOR performs in-place forward sweeps parameterized by a relaxation factor $\omega \in (0, 2)$, offering regular memory access and fixed-point friendliness [66].
- **Richardson Iteration (RI):** An update of the form $\hat{\mathbf{s}}^{(k+1)} = \hat{\mathbf{s}}^{(k)} + \omega(\mathbf{y}_{\text{MF}} - \mathbf{W}\hat{\mathbf{s}}^{(k)})$, with scalar relaxation. Convergence depends on both ω and the initialization $\hat{\mathbf{s}}^{(0)}$, motivating structured initial guesses [65].

Research Objectives

This chapter pursues the following objectives:

1. **Near-MMSE performance with quadratic cost:** Replace explicit inversion with iterative solvers whose per-iteration complexity is $O(U^2)$, achieving near-MMSE BER with a small iteration budget.
2. **Load-aware, low-tuning parameterization:** Use quasi-optimal relaxation selections (e.g., $\omega = \hat{\omega}_0(\psi)$ as a function of loading $\psi = N/U$) to avoid per-realization tuning while preserving robustness.
3. **High quality initialization:** Employ one stabilized CG step to seed the proposed cascade, and develop structured initializations for RI (emphasizing inverse-stair) to accelerate convergence.
4. **E2E efficiency:** Provide closed-form operation counts, memory considerations, and hardware friendly formulations (streaming, in-place updates, minimal divisions).
5. **Reproducible benchmarking:** Validate via BER–SNR curves and complexity analyses on representative settings (e.g., $N = 128$, $U \in \{16, 32\}$, 64-QAM), comparing against CG, SOR, RI variants, and MMSE baselines.

Approach Overview

To meet these objectives, we adopt a cascade strategy that initializes with a single CG step to obtain a numerically stable $\hat{\mathbf{s}}^{(1)}$, then refines with SOR using a loading-aware relaxation

ω. In parallel, we investigate RI equipped with structured initializations and quantify when inverse-stair initialization offers the best convergence performance trade-off.

Subsequent sections detail algorithmic updates, parameter choices, complexity accounting, and performance results under consistent simulation protocols.

2.3 Iterative Detection Algorithms

This section presents three inversion-free iterative detectors for uplink massive MIMO: CG, SOR, and RI with structured initialization. Each method solves the linear MMSE system

$$\mathbf{W}\hat{\mathbf{s}} = \mathbf{y}_{\mathrm{MF}}, \qquad \mathbf{W} = \mathbf{H}^H\mathbf{H} + \sigma^2\mathbf{I}_U \succ 0, \tag{2.5}$$

using only matrix–vector products and simple vector operations, thereby avoiding explicit matrix inversion and reducing complexity relative to direct MMSE.

2.3.1 Conjugate Gradient (CG)

The CG method is a Krylov-subspace iterative solver tailored for HPD systems. In the context of uplink massive MIMO detection, it solves the linear system (2.5) efficiently.

We define the MF output and MMSE filtering matrix as:

$$\mathbf{y}_{\mathrm{MF}} \triangleq \mathbf{H}^H\mathbf{y}, \tag{2.6}$$

$$\mathbf{W} \triangleq \mathbf{H}^H\mathbf{H} + \sigma^2\mathbf{I}_U. \tag{2.7}$$

The CG method solves (2.5) without explicit inversion by iteratively minimizing the strictly convex quadratic cost function:

$$\mathcal{J}(\mathbf{s}) \triangleq \tfrac{1}{2}\mathbf{s}^H\mathbf{W}\mathbf{s} - \Re\{\mathbf{s}^H\mathbf{y}_{\mathrm{MF}}\}. \tag{2.8}$$

where $\mathcal{J}(\mathbf{s})$ is a strictly convex quadratic functional of the candidate vector $\mathbf{s} \in \mathbb{C}^{U\times 1}$, $\Re\{\cdot\}$ denotes the real part operator.

Recent Advances: Recent studies have extended CG-based detection to address practical constraints in massive MIMO systems. Berthe et al. [67] proposed an adaptive preconditioned gradient descent (GD) detector that improves soft-output performance while maintaining low complexity. Fang et al. [68] introduced a finite-precision CG framework that mitigates arithmetic bottlenecks and enables precision-aware convergence tuning. Neama

and Al-Rubaye [69] developed a hybrid ADMM-CG detector that leverages tridiagonal initialization and alternating updates to reduce iteration count and improve robustness under high modulation orders. Karahan and Kalaycıoğlu [70] unfolded the CG algorithm into a deep learning architecture with Tikhonov regularization, enabling trainable step sizes and dynamic adaptation to channel conditions.

CG Updates: Using the standard inner product $\langle \mathbf{a}, \mathbf{b} \rangle = \mathbf{a}^H \mathbf{b}$, the CG iterations proceed as follows:

$$\mathbf{r}^{(k)} = \mathbf{y}_{\text{MF}} - \mathbf{W}\hat{\mathbf{s}}^{(k)}, \tag{2.9}$$

$$\alpha^{(k)} = \frac{\langle \mathbf{r}^{(k)}, \mathbf{r}^{(k)} \rangle}{\langle \mathbf{p}^{(k)}, \mathbf{W}\mathbf{p}^{(k)} \rangle}, \tag{2.10}$$

$$\hat{\mathbf{s}}^{(k+1)} = \hat{\mathbf{s}}^{(k)} + \alpha^{(k)} \mathbf{p}^{(k)}, \tag{2.11}$$

$$\mathbf{r}^{(k+1)} = \mathbf{r}^{(k)} - \alpha^{(k)} \mathbf{W}\mathbf{p}^{(k)}, \tag{2.12}$$

$$\beta^{(k)} = \frac{\langle \mathbf{r}^{(k+1)}, \mathbf{r}^{(k+1)} \rangle}{\langle \mathbf{r}^{(k)}, \mathbf{r}^{(k)} \rangle}, \tag{2.13}$$

$$\mathbf{p}^{(k+1)} = \mathbf{r}^{(k+1)} + \beta^{(k)} \mathbf{p}^{(k)}. \tag{2.14}$$

Here $\mathbf{r}^{(k)}$ is the residual vector at iteration k, $\mathbf{p}^{(k)}$ is the CG search direction, $\alpha^{(k)} \in \mathbb{R}$ is the step size, and $\beta^{(k)} \in \mathbb{R}$ is the direction update coefficient.

The directions $\{\mathbf{p}^{(k)}\}$ are $\mathbf{W}$-conjugate, meaning $\mathbf{p}^{(i)H} \mathbf{W} \mathbf{p}^{(j)} = 0$ for $i \neq j$.

Convergence and Rate: Since $\mathbf{W} \succ 0$, the cost function (2.8) is strictly convex with a unique minimizer $\hat{\mathbf{s}}^\star = \mathbf{W}^{-1}\mathbf{y}_{\text{MF}}$. CG enforces optimality along each search direction and maintains conjugacy, guaranteeing convergence in at most U iterations in exact arithmetic. The error in the $\mathbf{W}$-norm satisfies:

$$\|\hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^\star\|_{\mathbf{W}} \leq 2 \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^k \|\hat{\mathbf{s}}^{(0)} - \hat{\mathbf{s}}^\star\|_{\mathbf{W}}, \qquad \kappa = \frac{\lambda_{\max}(\mathbf{W})}{\lambda_{\min}(\mathbf{W})}, \tag{2.15}$$

The norm $\|\mathbf{x}\|_{\mathbf{W}} \triangleq \sqrt{\mathbf{x}^H \mathbf{W} \mathbf{x}}$ is the $\mathbf{W}$-weighted Euclidean norm, and κ is the condition number of $\mathbf{W}$. Which shows linear convergence governed by the condition number κ.

Complexity and Implementation: Each iteration requires one matrix–vector product $\mathbf{W}\mathbf{p}^{(k)}$ and several vector operations, yielding $O(U^2)$ complexity per iteration for dense $\mathbf{W}$. Efficient implementation avoids forming $\mathbf{W}$ explicitly by computing $\mathbf{W}\mathbf{s}$ as $\mathbf{H}^H(\mathbf{H}\mathbf{s}) + \sigma^2 \mathbf{s}$.

Algorithm 1 CG Method for Uplink Massive MIMO Detection

Require: Channel matrix $\mathbf{H}$, received vector $\mathbf{y}$, iteration index k, noise variance σ^2

Ensure: Estimated transmitted signal $\hat{\mathbf{s}}$

Initialization:
  Compute matched-filter output $\mathbf{y}_{\text{MF}}$ using (2.6)
  Compute MMSE weight matrix $\mathbf{W}$ using (2.7)
  Set $\hat{\mathbf{s}}^{(0)} = \mathbf{0}$, $\mathbf{r}^{(0)} = \mathbf{y}_{\text{MF}}$, $\mathbf{p}^{(0)} = \mathbf{r}^{(0)}$, $k = 0$
while $\mathbf{r}^{(k)} \neq \mathbf{0}$ **do**
    Compute step-size $\alpha^{(k)}$ using (2.10)
    Update estimate $\hat{\mathbf{s}}^{(k+1)}$ using (2.11)
    Update residual $\mathbf{r}^{(k+1)}$ using (2.12)
    Compute CG scalar $\beta^{(k)}$ using (2.13)
    Update search direction $\mathbf{p}^{(k+1)}$ using (2.14)
    Increment $k \leftarrow k + 1$
end while
return $\hat{\mathbf{s}}$

Notes on Initialization and Stopping: A practical initialization sets $\hat{\mathbf{s}}^{(0)} = \mathbf{0}$ and $\mathbf{r}^{(0)} = \mathbf{y}_{\text{MF}}$ as in (2.9). Termination is based on a residual norm threshold $\|\mathbf{r}^{(k)}\|_2 \leq \varepsilon$ or a maximum iteration count.

2.3.2 Successive Over-Relaxation (SOR)

The SOR method is a Gauss–Seidel(GS) variant augmented with a relaxation parameter $\omega \in (0, 2)$ to accelerate convergence on HPD systems. In uplink massive MIMO detection, SOR solves the linear MMSE system (2.5) by exploiting matrix splitting and forward substitution, yielding in-place updates and regular memory access that are well-suited to low-latency, energy-efficient hardware. Empirical and analytical evidence shows that SOR-based detectors can approach MMSE performance with a small number of sweeps when the relaxation is chosen judiciously and simple preconditioners are used [71, 72, 73, 74, 75, 76].

Matrix splitting: Decompose the HPD matrix as

$$\mathbf{W} = \mathbf{D} + \mathbf{L} + \mathbf{L}^H, \tag{2.16}$$

Here $\mathbf{D} = \mathrm{diag}(\mathbf{W})$ is the diagonal of $\mathbf{W}$, $\mathbf{L}$ is its strictly lower triangular part, and $\mathbf{L}^H$ is the strictly upper triangular part. The SOR iteration (matrix form) is

$$\hat{\mathbf{s}}^{(k)} = \Big(\mathbf{L} + \tfrac{1}{\omega}\mathbf{D}\Big)^{-1}\Big(\mathbf{y}_{\mathrm{MF}} + \big(\tfrac{1}{\omega} - 1\big)\mathbf{D}\hat{\mathbf{s}}^{(k-1)} - \mathbf{L}^H\hat{\mathbf{s}}^{(k-1)}\Big), \tag{2.17}$$

which corresponds to the elementwise forward-sweep update

$$\hat{s}_m^{(k)} = (1-\omega)\hat{s}_m^{(k-1)} + \frac{\omega}{w_{mm}}\left(y_{\mathrm{MF},m} - \sum_{u<m} w_{mu}\hat{s}_u^{(k)} - \sum_{u>m} w_{mu}\hat{s}_u^{(k-1)}\right), \quad m = 1,\ldots,U. \tag{2.18}$$

In (2.18), $\hat{s}_m^{(k)}$ denotes the mth component of the estimate at iteration k, w_{mu} is the (m,u)th entry of $\mathbf{W}$, $y_{\mathrm{MF},m}$ is the mth element of $\mathbf{y}_{\mathrm{MF}}$, and $\omega \in (0,2)$ is the SOR relaxation factor.

Fixed-point form and iteration matrix: Define

$$\mathbf{B}_\omega \triangleq \Big(\mathbf{L} + \tfrac{1}{\omega}\mathbf{D}\Big)^{-1}\Big(\big(\tfrac{1}{\omega} - 1\big)\mathbf{D} - \mathbf{L}^H\Big), \qquad \mathbf{e} \triangleq \Big(\mathbf{L} + \tfrac{1}{\omega}\mathbf{D}\Big)^{-1}\mathbf{y}_{\mathrm{MF}}, \tag{2.19}$$

so that the iteration reads

$$\hat{\mathbf{s}}^{(k)} = \mathbf{B}_\omega\,\hat{\mathbf{s}}^{(k-1)} + \mathbf{e}. \tag{2.20}$$

Here $\mathbf{B}_\omega$ is the SOR iteration matrix and $\mathbf{e}$ is the corresponding constant term in the affine fixed-point iteration.

Convergence and relaxation tuning (sketch): For HPD $\mathbf{W}$, GS ($\omega = 1$) converges, i.e., $\rho(\mathbf{B}_1) < 1$. Classical SOR theory guarantees $\rho(\mathbf{B}_\omega) < 1$ for all $0 < \omega < 2$, hence (2.20) converges to the unique solution of (2.5). In massive MIMO practice, load-aware prescriptions select ω from the antenna to user ratio $\psi = N/U$, avoiding per-realization searches. Such rules have been shown to be near-optimal across channel/load regimes [72, 73, 75]. A representative mapping is

$$\hat{\omega}_0(\psi) \approx \frac{2}{1 + \sqrt{(1-\psi^{-1})^2 + \psi^{-2}}}, \tag{2.21}$$

In (2.21), $\psi \triangleq N/U$ denotes the BS loading factor, i.e., the ratio between the number of BS antennas N and the number of users U.

which captures how the optimal relaxation increases with loading and moderates under correlation [75].

Recent advances and hybridization :

- Hardware efficient SOR designs demonstrate high throughput and EE with fixed-point datapaths and predictable memory access [72, 74].

- Load-aware/channel-aware relaxation policies and lightweight diagonal/block preconditioners improve robustness under antenna correlation and high order modulations [71, 75].

- Hybrid detectors e.g., warm starting SOR with one or two Krylov/gradient steps (CG/PGD) or alternating with Jacobi (JA)/GS reduce iteration counts and latency while maintaining near-MMSE performance [73, 76].

Complexity and implementation: For dense $\mathbf{W}$, one forward sweep via (2.18) requires $4U^2+4U$ multiplications (dominant term $O(U^2)$). No matrix inversion is needed, and $\mathbf{W}$ need not be explicitly stored if entries (or an operator) are accessible on-the-fly. SOR naturally maps to fixed-point and parallel architectures; diagonal scaling or simple preconditioning can reduce iteration counts when $\mathbf{D}$ is poorly conditioned [74, 75].

Algorithm 2 SOR Method for Uplink Massive MIMO Detection

Input: Channel matrix $\mathbf{H}$, received vector $\mathbf{y}$, relaxation parameter ω, noise variance σ^2, maximum iteration index i

Output: Estimated transmitted signal $\hat{\mathbf{s}}$

Initialization:

  Compute matched filter output $\mathbf{y}_{\mathrm{MF}} = \mathbf{H}^H\mathbf{y}$ (see Eq. (2.6))

  Compute filtering matrix $\mathbf{W} = \mathbf{H}^H\mathbf{H} + \sigma^2\mathbf{I}$ (see Eq. (2.7))

  Set diagonal matrix $\mathbf{D} = \mathrm{diag}(\mathbf{W})$

  Set strictly upper triangular part $\mathbf{L}^H = -\,\mathrm{triu}(\mathbf{W})$

  Set strictly lower triangular part $\mathbf{L} = -\,\mathrm{tril}(\mathbf{W})$

  Choose relaxation parameter $0 < \omega < 2$ (see Eq. (2.21))

  Initialize estimate $\hat{\mathbf{s}}^{(0)} = \mathbf{0}$

for $k = 0$ to i **do**

  Construct the SOR iteration using Eq. (2.17) (equivalently Eq. (2.18))

end for

Return: Final estimate $\hat{\mathbf{s}}$

Practical tips :

- Choose $\omega \in [1.0, 1.6]$ as a robust default; $\omega = 1$ recovers GS.

- Use (2.21) when per-realization tuning is undesirable; calibrate the iteration budget i to latency constraints and allow early stopping via a residual test.

- Warm start SOR with one or two CG (or preconditioned gradient) steps to reduce sweeps at target BER [73, 76].

- Apply diagonal scaling/preconditioning if $\mathbf{D}$ is poorly conditioned; this improves stability and reduces iteration counts [74, 75].

2.3.3 Hybrid CG–SOR Detector

To combine the rapid convergence of CG with the hardware-friendliness of SOR, we propose a hybrid detection method that uses a CG warm start to obtain a stable initial estimate, followed by fine-tuning via inexpensive SOR sweeps. This approach avoids matrix inversion and achieves near-MMSE performance with reduced iteration count and predictable complexity.

Figure 2.3 illustrates the overall processing flow of the CG–SOR detector, showing how the matched filter output is processed through CG initialization and SOR refinement stages.

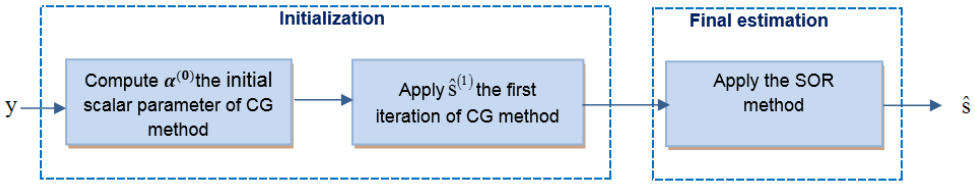

Figure 2.3: Block diagram of the proposed CG–SOR detection method.

The CG–SOR detector operates in two stages: a single CG iteration to accelerate initial convergence, followed by multiple SOR sweeps for low-cost residual correction. This structure is particularly well suited to fixed-point and streaming architectures.

Algorithmic Summary

Algorithm 3 summarizes the procedural steps of the hybrid CG–SOR detector. It consolidates matched filtering, CG-based initialization, and SOR-based refinement into a unified iterative process.

Algorithm 3 Hybrid CG–SOR Detection for Uplink massive MIMO

Input: Channel matrix $\mathbf{H}$, received vector $\mathbf{y}$, noise variance σ^2, CG iterations k_{cg}, SOR sweeps k_{sor}, relaxation factor ω, tolerance ε

Output: Estimated symbol vector $\hat{\mathbf{s}}$

Compute matched filter output: $\mathbf{y}_{\text{MF}} = \mathbf{H}^H\mathbf{y}$
Define matrix-free operator: $\mathbf{W}\mathbf{x} = \mathbf{H}^H(\mathbf{H}\mathbf{x}) + \sigma^2\mathbf{x}$
Stage 1: CG Warm-Start
Initialize: $\hat{\mathbf{s}}^{(0)} = \mathbf{0}$, $\mathbf{r}^{(0)} = \mathbf{y}_{\text{MF}}$, $\mathbf{p}^{(0)} = \mathbf{r}^{(0)}$
for $i = 0$ to $k_{\text{cg}} - 1$ **do**
    Compute $\alpha^{(i)}$, update $\hat{\mathbf{s}}^{(i+1)}$, $\mathbf{r}^{(i+1)}$
    Compute $\beta^{(i)}$, update $\mathbf{p}^{(i+1)}$
end for
Stage 2: SOR Refinement
Set $\hat{\mathbf{s}}_{\text{SOR}}^{(0)} = \hat{\mathbf{s}}^{(k_{\text{cg}})}$
for $j = 1$ to k_{sor} **do**
    for $m = 1$ to U **do**
        Apply SOR update using Eq. (2.18)
    end for
    if $\|\mathbf{y}_{\text{MF}} - \mathbf{W}\hat{\mathbf{s}}^{(j)}\|_2 \leq \varepsilon$ **then**
        break
    end if
end for
return $\hat{\mathbf{s}}^{(j)}$

Here k_{cg} and k_{sor} denote the number of CG iterations and SOR sweeps, respectively, ω is the SOR relaxation parameter, and $\varepsilon > 0$ is the stopping tolerance on the residual norm.

The hybrid CG–SOR detector effectively trades a small number of CG iterations for fast, in place SOR refinement. It delivers near-MMSE performance under a computational footprint compatible with real time hardware implementation, and is particularly well suited to low-power, latency constrained platforms such as FPGAs and ASICs.

2.3.4 Richardson Iteration with Structured Initialization

RI is a classical stationary method for solving HPD linear systems. It iteratively updates the estimate along the residual scaled by a relaxation parameter and avoids matrix inversion, an

important advantage for uplink massive MIMO MMSE detection under tight latency/energy budgets. In massive MIMO, RI has been shown to reduce the arithmetic complexity of linear detection from cubic to quadratic order while retaining near-MMSE performance when paired with appropriate relaxation and initialization policies [28, 77, 78, 79]. Recent advances also address correlated channels, fixed-point deployment, and structured warm starts to further improve convergence and robustness [68, 80, 81].

Algorithmic formulation: Consider the linear MMSE system

$$\mathbf{W}\,\hat{\mathbf{s}} = \mathbf{y}_{\mathrm{MF}}, \qquad \mathbf{W} = \mathbf{H}^H\mathbf{H} + \sigma^2\mathbf{I}_U, \qquad \mathbf{y}_{\mathrm{MF}} = \mathbf{H}^H\mathbf{y}, \tag{2.22}$$

with HPD $\mathbf{W}$. Define the residual

$$\mathbf{r}^{(k)} \triangleq \mathbf{y}_{\mathrm{MF}} - \mathbf{W}\hat{\mathbf{s}}^{(k)}. \tag{2.23}$$

where $\mathbf{r}^{(k)}$ is the residual vector at iteration k.

The RI update at iteration k is

$$\hat{\mathbf{s}}^{(k+1)} = \hat{\mathbf{s}}^{(k)} + \omega\left(\mathbf{y}_{\mathrm{MF}} - \mathbf{W}\hat{\mathbf{s}}^{(k)}\right) = (\mathbf{I} - \omega\mathbf{W})\hat{\mathbf{s}}^{(k)} + \omega\,\mathbf{y}_{\mathrm{MF}}, \tag{2.24}$$

i.e., the fixed-point iteration

$$\hat{\mathbf{s}}^{(k+1)} = \mathbf{B}_\omega^{\mathrm{RI}}\,\hat{\mathbf{s}}^{(k)} + \mathbf{c}_\omega^{\mathrm{RI}}, \qquad \mathbf{B}_\omega^{\mathrm{RI}} \triangleq \mathbf{I} - \omega\mathbf{W}, \quad \mathbf{c}_\omega^{\mathrm{RI}} \triangleq \omega\,\mathbf{y}_{\mathrm{MF}}. \tag{2.25}$$

Here $\mathbf{B}_\omega^{\mathrm{RI}}$ is the RI iteration matrix for a given relaxation parameter ω, and $\mathbf{c}_\omega^{\mathrm{RI}}$ is the associated constant vector.

Convergence conditions and relaxation selection: Let $\lambda(\mathbf{W}) \subset [\lambda_{\min}, \lambda_{\max}]$ with $0 < \lambda_{\min} \le \lambda_{\max}$. The error $\mathbf{e}^{(k)} = \hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{\star}$ evolves as

$$\mathbf{e}^{(k+1)} = (\mathbf{I} - \omega\mathbf{W})\,\mathbf{e}^{(k)}, \qquad \rho(\mathbf{I} - \omega\mathbf{W}) < 1 \Longleftrightarrow 0 < \omega < \frac{2}{\lambda_{\max}}. \tag{2.26}$$

The notation $\rho(\cdot)$ denotes the spectral radius, i.e., the largest magnitude of the eigenvalues of its matrix argument.

The optimal constant relaxation minimizing the spectral radius is

$$\omega^\star = \frac{2}{\lambda_{\min} + \lambda_{\max}}, \qquad \rho(\mathbf{I} - \omega^\star\mathbf{W}) = \frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}} = \frac{\kappa - 1}{\kappa + 1}, \quad \kappa = \frac{\lambda_{\max}}{\lambda_{\min}}, \tag{2.27}$$

Here $\lambda_{\min}$ and $\lambda_{\max}$ denote the minimum and maximum eigenvalues of $\mathbf{W}$, respectively.

ensuring linear convergence. When eigenvalue bounds are unavailable in real time, realization agnostic, load-aware heuristics e.g., $\hat{\omega}_0 \approx 1/(N+U)$ deliver robust convergence across loads and channel models [28, 78, 80]:

$$\hat{\omega}_0(N,U) \approx \frac{1}{N+U}. \tag{2.28}$$

In (2.28), N is the number of BS antennas and U is the number of users, so $\hat{\omega}_0(N,U)$ depends only on the system dimensions.

Structured initialization strategies: RI is sensitive to the initial guess $\hat{\mathbf{s}}^{(0)}$; informed initialization sharply reduces iteration counts:

- **ZVIn**: $\hat{\mathbf{s}}^{(0)} = \mathbf{0}$ — minimal cost, baseline convergence.
- **DI (Diagonal-Inverse)**: $\hat{\mathbf{s}}^{(0)} = \mathbf{D}^{-1}\mathbf{y}_{\mathrm{MF}}$, $\mathbf{D} = \mathrm{diag}(\mathbf{W})$ — per-stream scaling [79].
- **ITMIn (Inverse Tridiagonal)**: construct $\mathbf{W}_{\mathrm{tri}}$ and set $\hat{\mathbf{s}}^{(0)} = \mathbf{W}_{\mathrm{tri}}^{-1}\mathbf{y}_{\mathrm{MF}}$ — linear-time solve, robust under moderate correlation.
- **ISMIn (Inverse Stair)**: use stair-structured $\mathbf{S} \approx \mathbf{W}$ and set $\hat{\mathbf{s}}^{(0)} = \mathbf{S}^{-1}\mathbf{y}_{\mathrm{MF}}$ — strong empirical performance in massive MIMO [28, 82].
- **ADAIn (Adaptive Init)**: $\hat{\mathbf{s}}^{(0)} = \hat{\omega}_0\,\mathbf{y}_{\mathrm{MF}}$ with $\hat{\omega}_0$ from (2.28) — negligible overhead [78].
- **CG-based Warm Start (CG1)**: one stabilized CG step from $\hat{\mathbf{s}}^{(0)} = \mathbf{0}$,

$$\alpha^{(0)} = \frac{\|\mathbf{y}_{\mathrm{MF}}\|_2^2}{\mathbf{y}_{\mathrm{MF}}^H \mathbf{W}\mathbf{y}_{\mathrm{MF}}}, \qquad \hat{\mathbf{s}}^{(0)} = \alpha^{(0)}\mathbf{y}_{\mathrm{MF}}, \tag{2.29}$$

  which often provides a strong initial residual reduction.

Comparative studies indicate ISMIn and DI provide excellent convergence–complexity trade-offs, while ADAIn offers the lightest initialization; CG1 further reduces the iteration budget for tight latency constraints [28, 68, 80].

Recent Advances and Applications: RI has gained renewed attention in massive MIMO research due to its simplicity, scalability, and compatibility with hardware constraints. Kang et al. [79] demonstrated that RI-based detectors can reduce MMSE complexity from $O(U^3)$ to $O(U^2)$, enabling real time deployment in large-scale systems. Labed and Aounallah [28] evaluated multiple structured initializations and showed that inverse-stair and CG-based

warm starts significantly improve convergence speed. Chen et al. [80] proposed load-aware relaxation tuning for correlated channels, while Park and Kim [81] developed fixed-point RI architectures optimized for very large-scale integration (VLSI). Fang et al. [68] further analyzed finite-precision effects and recommended precision-aware warm starts to stabilize early iterations. These contributions confirm RI's role as a low-complexity, inversion-free alternative for uplink detection under diverse channel and hardware conditions.

Complexity and performance trade-offs: Each RI step requires one $\mathbf{W}$ vector product and $O(U)$ axpy operations; for dense $\mathbf{W}$ this is $O(U^2)$ per iteration. Avoid materializing $\mathbf{W}$ by implementing $\mathbf{Wx}$ as $\mathbf{H}^H(\mathbf{Hx}) + \sigma^2\mathbf{x}$. With a strong initialization, RI can reach near-MMSE performance in relatively few iterations; total cost scales as $O(kU^2)$ for k iterations plus a one time initializer [28, 79]. Fixed-point aware scaling of ω and residuals improves numerical robustness; recent finite-precision analyses provide guidance on quantization and restart policies [68, 81].

Algorithm 4 RI for Uplink Massive MIMO Detection

Input: Channel matrix $\mathbf{H}$, received vector $\mathbf{y}$, relaxation parameter ω, maximum iterations $k_{\max}$, noise variance σ^2
Output: Estimated symbol vector $\hat{\mathbf{s}}$
Initialization:
  Compute matched filter output $\mathbf{y}_{\text{MF}} = \mathbf{H}^H\mathbf{y}$ (see Eq. (2.6))
  Define operator $\mathbf{x} \mapsto \mathbf{H}^H(\mathbf{Hx}) + \sigma^2\mathbf{x}$ (implicit $\mathbf{W}$)
  Choose structured initial guess $\hat{\mathbf{s}}^{(0)}$ (e.g., ZVIn, DI, ITMIn, ISMIn, ADAIn, or CG1)
  Set iteration index $k = 0$
while $k < k_{\max}$ **do**
  Compute residual $\mathbf{r}^{(k)} = \mathbf{y}_{\text{MF}} - \mathbf{W}\hat{\mathbf{s}}^{(k)}$ (Eq. (2.23))
  if $\|\mathbf{r}^{(k)}\|_2 \leq \varepsilon$ **then**
    break
  end if
  Update estimate $\hat{\mathbf{s}}^{(k+1)} = \hat{\mathbf{s}}^{(k)} + \omega\,\mathbf{r}^{(k)}$ (Eq. (2.24))
  Increment iteration index $k \leftarrow k + 1$
end while
Return: Final estimate $\hat{\mathbf{s}}^{(k)}$

2.4 Performance and Complexity Analysis

This section presents a comprehensive analysis of computational complexity, convergence behavior, and empirical detection performance for inversion-free iterative detectors CG, SOR, and RI together with hybrid schemes (CG–SOR) and RI detectors with structured initialization. The discussion integrates closed form cost models with simulation evidence under realistic uplink massive MIMO settings ($N \times U \in \{128\times16,\ 128\times32\}$, 64-QAM, AWGN, flat Rayleigh fading), highlighting latency–energy trade-offs and hardware suitability.

2.4.1 Computational Complexity

Let U denote the number of users and N the number of BS antennas. For dense $\mathbf{H}$, the dominant per-iteration cost is governed by matrix–vector multiplications and a small number of vector inner products/axpy operations. To avoid forming $\mathbf{W}$ explicitly, all methods should compute $\mathbf{Wx}$ implicitly as $\mathbf{H}^H(\mathbf{Hx}) + \sigma^2\mathbf{x}$.

Table 2.1: Computational complexity (dominant multiplications)

Method	**Complexity (Multiplications)**
MMSE (direct solve/factorization)	$O(U^3)$
CG (per k iterations)	$kU^2 + 6kU$
SOR (per k sweeps)	$4kU^2 + 4kU$
CG–SOR (1 CG + k SOR)	$(4k+1)U^2 + (4k+3)U$
RI (ZVIn)	$4kU^2 + 2U(k+1)$
RI (ISMIn)	$4kU^2 + U(2k+4) - 3$
RI (ADAIn)	$4kU^2 + U(2k+1) + 1$

In this table, U denotes the number of users, k is the number of iterations or sweeps, and ZVIn, ISMIn, and ADAIn denote Zero-Vector Initialization, Inverse Stair Matrix Initialization, and Adaptive Initialization, respectively.

Key takeaways:

- All iterative detectors reduce the per-detection cost from $O(U^3)$ (direct MMSE) to $O(kU^2)$, enabling scalability to large user dimensions.

- CG has slightly higher per-iteration overhead (inner products/divisions) but typically requires fewer iterations on well-conditioned systems or with preconditioning.

- SOR and RI exhibit highly regular, reduction-free iterations with predictable $O(U^2)$ cost per sweep/step; their iteration counts depend on relaxation and (for RI) initialization.

- Hybrid CG–SOR amortizes a light CG warm start over a handful of SOR sweeps, while ISMIn–RI combines a strong initializer with the lightest RI iteration.

2.4.2 Convergence Behavior and Tuning

Convergence is driven by the conditioning of $\mathbf{W} = \mathbf{H}^H\mathbf{H}+\sigma^2\mathbf{I}_U$, the loading factor $\psi = N/U$, and channel correlation.

- **CG.** Linear convergence in the $\|\cdot\|_{\mathbf{W}}$-norm with rate bounded by

$$\left\|\hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{\star}\right\|_{\mathbf{W}} \leq 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{k}\left\|\hat{\mathbf{s}}^{(0)} - \hat{\mathbf{s}}^{\star}\right\|_{\mathbf{W}}, \qquad \kappa = \frac{\lambda_{\max}}{\lambda_{\min}}, \tag{2.30}$$

and finite termination in at most U steps in exact arithmetic. Diagonal/IC/learned preconditioning reduces κ and iteration budgets.

- **SOR.** Converges for $0 < \omega < 2$. Over-relaxation ($\omega > 1$) accelerates GS. A realization-agnostic, load-aware rule

$$\hat{\omega}_0(\psi) \approx \frac{2}{1+\sqrt{(1-\psi^{-1})^2+\psi^{-2}}}, \qquad \psi = \frac{N}{U}, \tag{2.31}$$

yields near-optimal convergence across channel realizations without per-instance tuning. Diagonal scaling mitigates variance in w_{mm} and supports robust fixed-point operation.

- **RI.** Converges for $0 < \omega < 2/\lambda_{\max}$; the optimal constant

$$\omega^{\star} = \frac{2}{\lambda_{\min}+\lambda_{\max}}, \quad \rho(\mathbf{I}-\omega^{\star}\mathbf{W}) = \frac{\kappa-1}{\kappa+1}, \tag{2.32}$$

minimizes the spectral radius. When spectral bounds are unavailable, a practical rule

$$\hat{\omega}_0(N,U) \approx \frac{1}{N+U} \tag{2.33}$$

delivers stable convergence, especially when paired with structured initialization.

Hybridization (e.g., CG–SOR, CG-initialized RI) leverages rapid initial descent and low-cost refinement to improve latency at fixed accuracy.

2.4.3 Initialization Sensitivity and Ranking

Initialization critically impacts SOR and especially RI. Consistent with our studies, the following ranking (best to weakest) holds:

- Inverse Stair Matrix (ISMIn);
- Diagonal-Inverse (DI), CG warm start (CG1);
- ADAIn, ITMIn;
- Zero Vector (ZVIn).

ISMIn provides a strong prior that markedly reduces the RI iteration budget with modest overhead, often achieving near-MMSE in few steps. CG1 is an attractive alternative when a single Krylov step is available in the pipeline.

2.4.4 Simulation Results and Embedded Figures

Setup and objectives: We assess inversion-free detectors under uplink massive MIMO with $N = 128$, $U \in 16, 32$, 64-QAM, AWGN, and flat Rayleigh fading. The goals are to: (i) quantify BER vs. SNR of CG–SOR against standalone CG, SOR, and JA at small iteration budgets; (ii) analyze SOR's relaxation parameter ω; (iii) contextualize results with operation-count comparisons; and (iv) evaluate RI with structured initializations.

CG–SOR at light loading ($U = 16$): Figures 2.4–2.6 report BER vs. SNR for 128×16 with $k = 2, 3, 4$ SOR sweeps applied after one CG warm-start step. As k increases from 2 to 4, CG–SOR rapidly approaches MMSE and consistently outperforms standalone JA, SOR, and CG at the same iteration budget, confirming that a single CG step provides a strong initial descent while SOR sweeps deliver low-cost refinement.

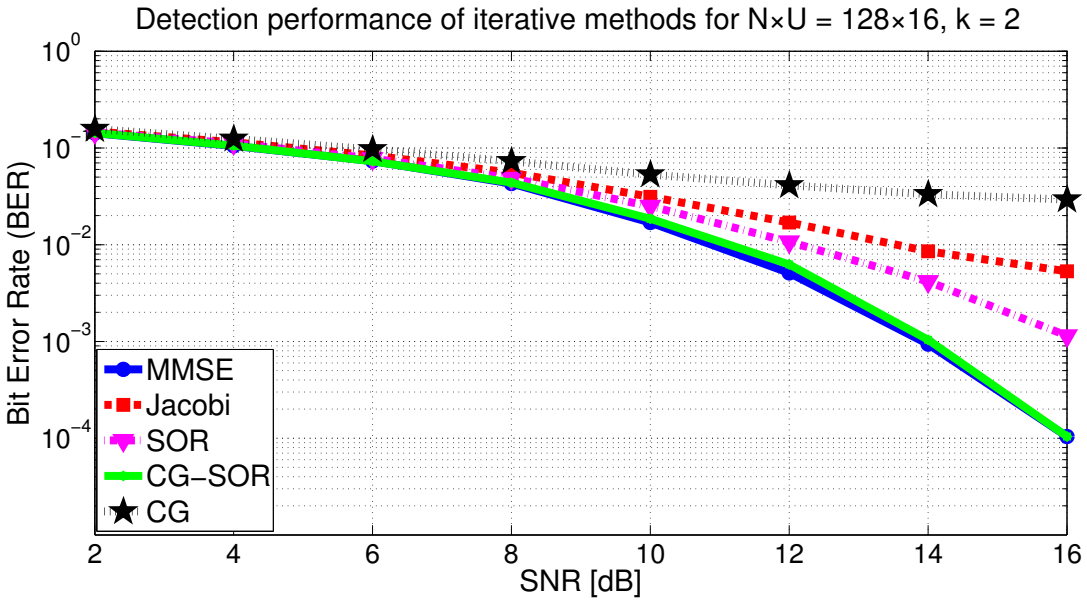

Figure 2.4: Comparative BER performance of MMSE, Jacobi, SOR, CG–SOR, and CG detectors for a 128×16 uplink massive MIMO system with $k = 2$ iterations under 64-QAM modulation.

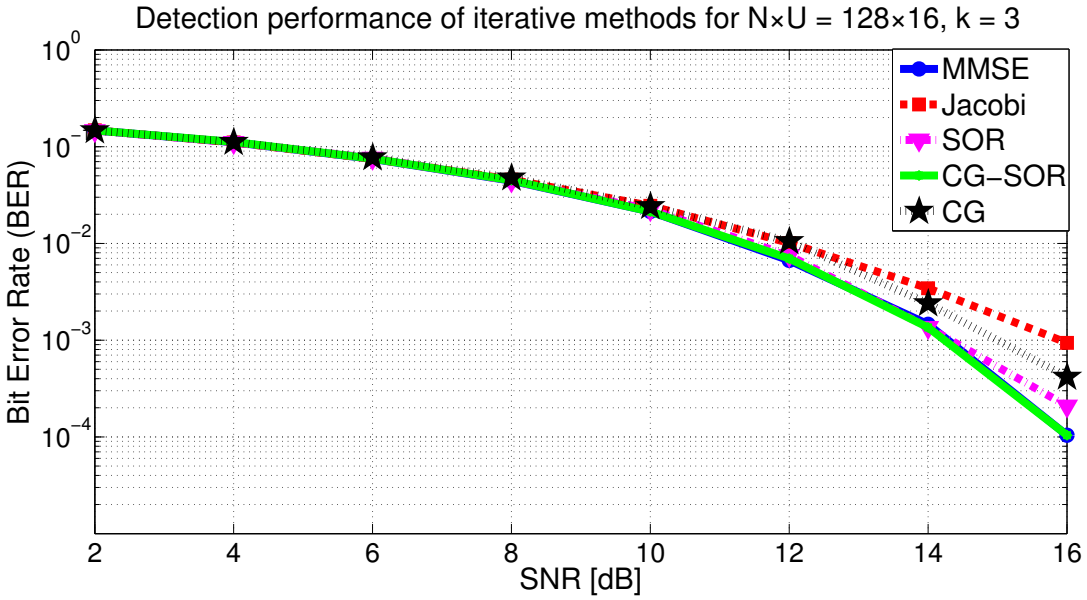

Figure 2.5: Comparative BER performance of MMSE, Jacobi, SOR, CG–SOR, and CG detectors for a 128×16 uplink massive MIMO system with $k = 3$ iterations under 64-QAM modulation.

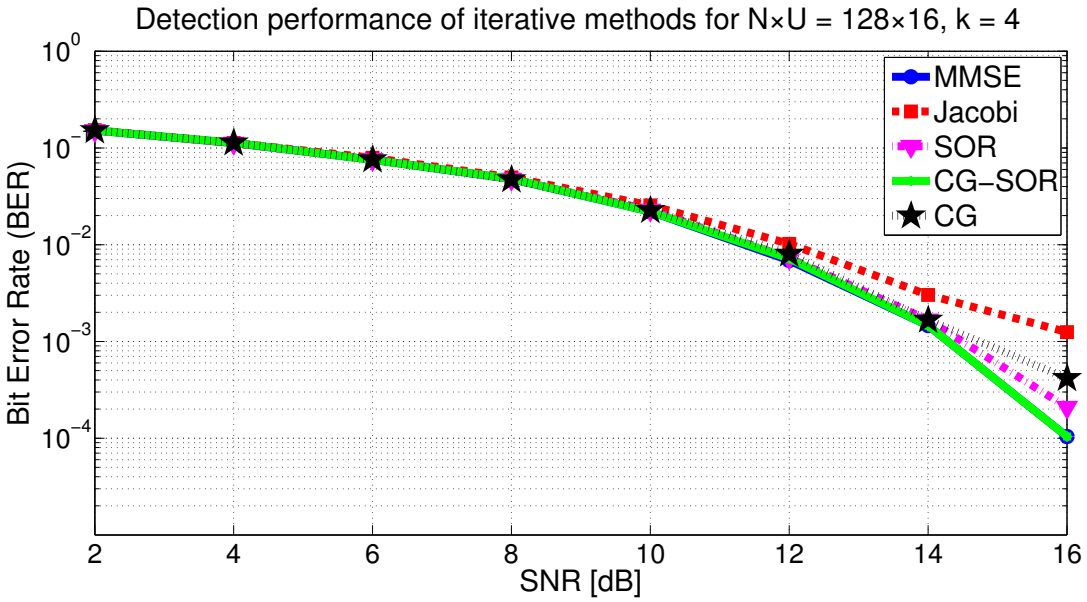

Figure 2.6: Comparative BER performance of MMSE, Jacobi, SOR, CG–SOR, and CG detectors for a 128×16 uplink massive MIMO system with $k = 4$ iterations under 64-QAM modulation.

CG–SOR at higher loading ($U = 32$): Figures 2.7–2.8 show that at 128×32, a single CG step followed by $k = 3$–4 SOR sweeps still attains near-MMSE BER across SNRs. The hybrid thus maintains its performance–cost advantage under higher loading.

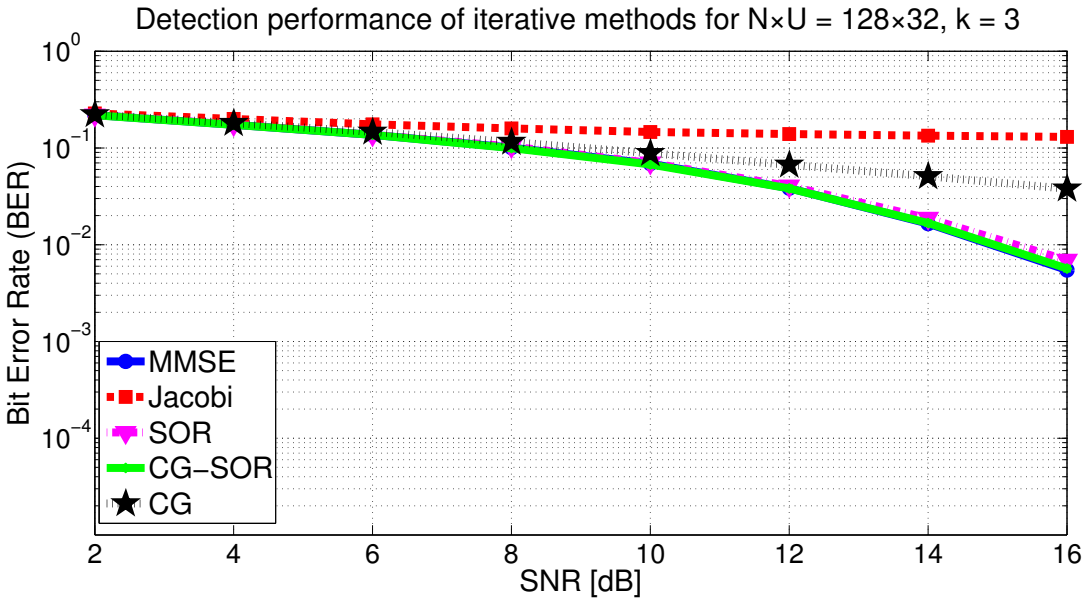

Figure 2.7: Comparative BER performance of MMSE, Jacobi, SOR, CG–SOR, and CG detectors for a 128×32 uplink massive MIMO system with $k = 3$ iterations under 64-QAM modulation.

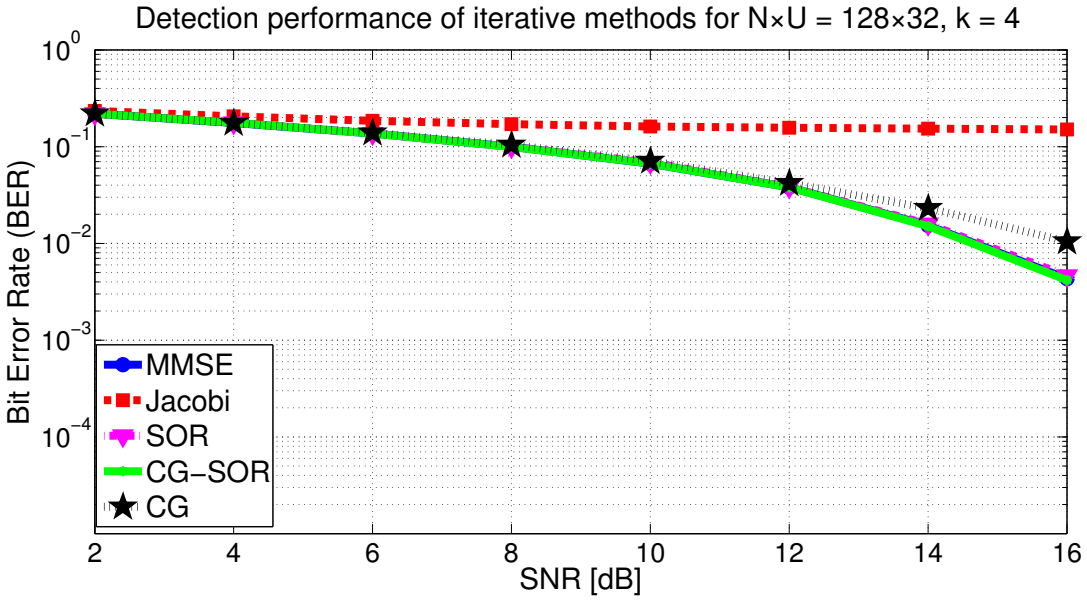

Figure 2.8: Comparative BER performance of MMSE, Jacobi, SOR, CG–SOR, and CG detectors for a 128×32 uplink massive MIMO system with $k = 4$ iterations under 64-QAM modulation.

SOR relaxation analysis and tuning: Figures 2.9–2.10 plot BER vs. ω at 128×32 with $k = 3$. The U-shaped curves indicate an empirical optimum near $\omega \approx 1.2$ at medium–high SNR, consistent with load-aware tuning for $\psi = N/U = 4$. This explains why the SOR stage in CG–SOR is particularly effective when ω is set near this region.

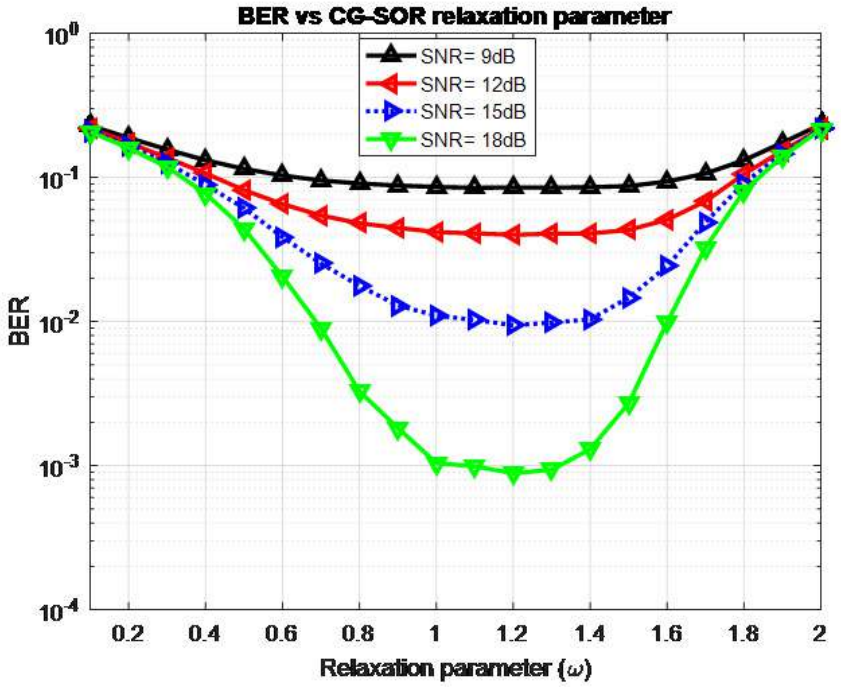

Figure 2.9: Influence of the relaxation parameter ω on BER performance of the SOR detector for a 128×32 uplink massive MIMO system with $k = 3$ iterations under 64-QAM modulation. Results are shown for multiple SNR values, highlighting the empirical optimum near $\omega \approx 1.2$.

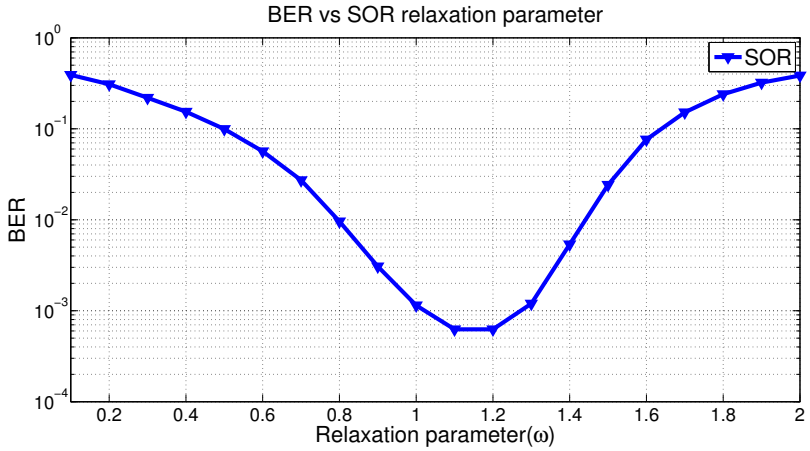

Figure 2.10: Effect of the relaxation parameter ω on the BER performance of the SOR detector for a 128×32 uplink massive MIMO system with $k = 3$ iterations under 64-QAM modulation at SNR $= 18$ dB. The curve exhibits a U-shaped trend, with the lowest BER observed near $\omega \approx 1.2$, indicating the empirical optimum relaxation factor.

Operation-count context for CG–SOR: Figure 2.11 compares multiplication counts for iterative methods at 128×16. While CG has the lowest per-iteration constant, CG–SOR achieves a superior accuracy–latency balance for small k because the CG warm start reduces the number of SOR sweeps needed to reach near-MMSE performance.

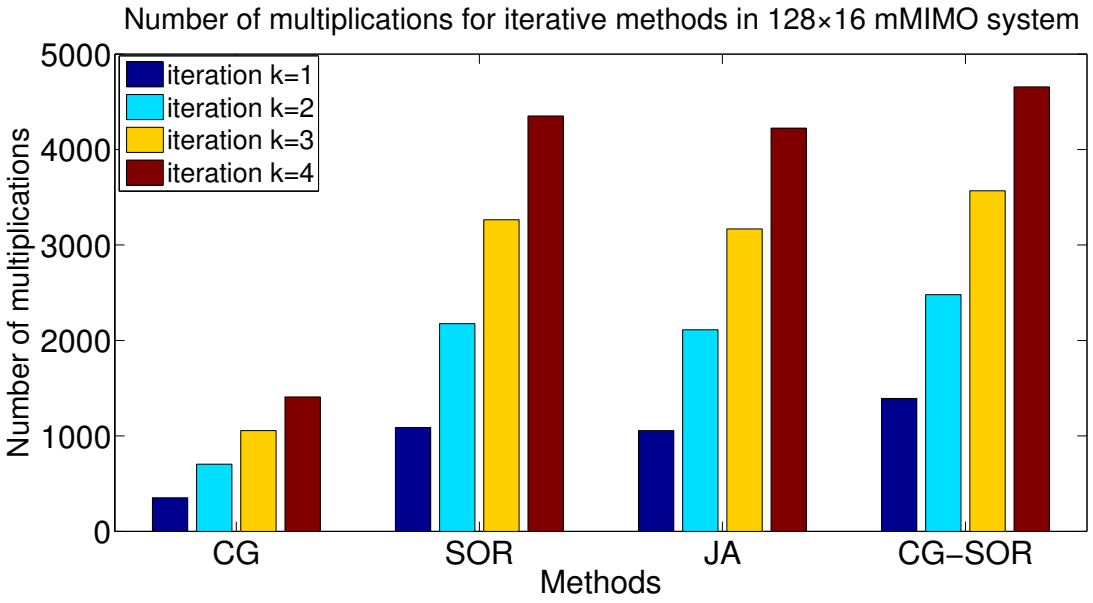

Figure 2.11: Comparison of multiplication counts required by different iterative detectors (CG, SOR, JA, and CG–SOR) in a 128×16 uplink massive MIMO system. The bars illustrate how computational load scales with the iteration index k, showing that CG maintains lower complexity while SOR, JA, and CG–SOR incur higher multiplication counts at larger k.

RI with structured initialization at $U = 16$: Figures 2.12–2.14 compare RI using ZVIn, ADAIn, DI, ITMIn, and ISMIn for $k = 2$–4. ISMIn–RI consistently yields the best BER across SNRs and approaches MMSE with fewer iterations, showing that stronger initializations are "worth it" because they reduce the dominant $4kU^2$ iteration cost.

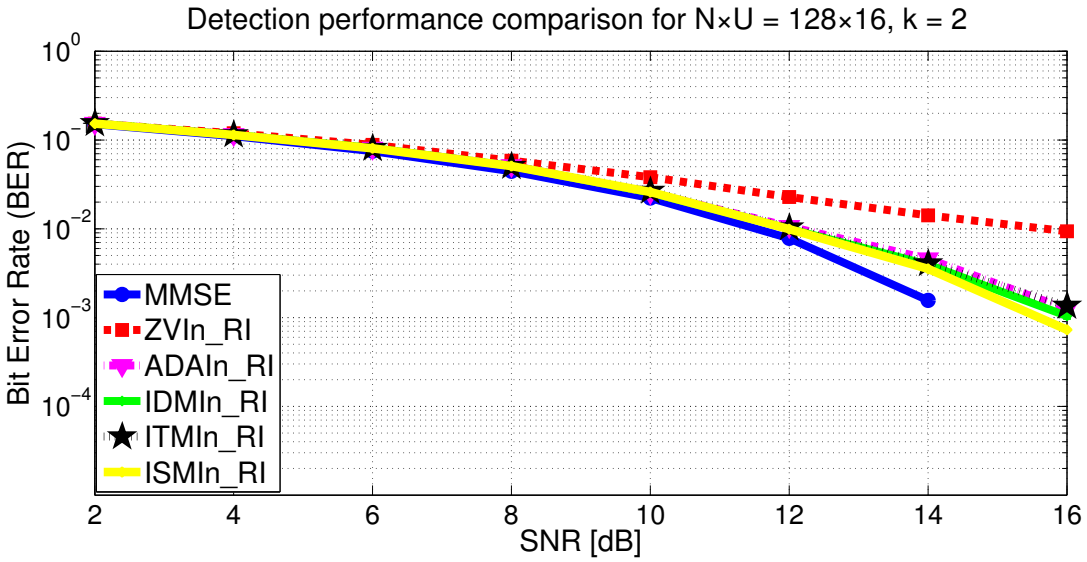

Figure 2.12: BER performance comparison of the RI detector under different initialization strategies (ZVIn_RI, ADAIn_RI, IDMIn_RI, ITMIn_RI, ISMIn_RI) for a 128×16 uplink massive MIMO system with $k = 2$ iterations and 64-QAM modulation. The curves highlight how initialization choice influences detection accuracy across varying SNR values, with structured methods offering improved convergence relative to simple zero-vector initialization.

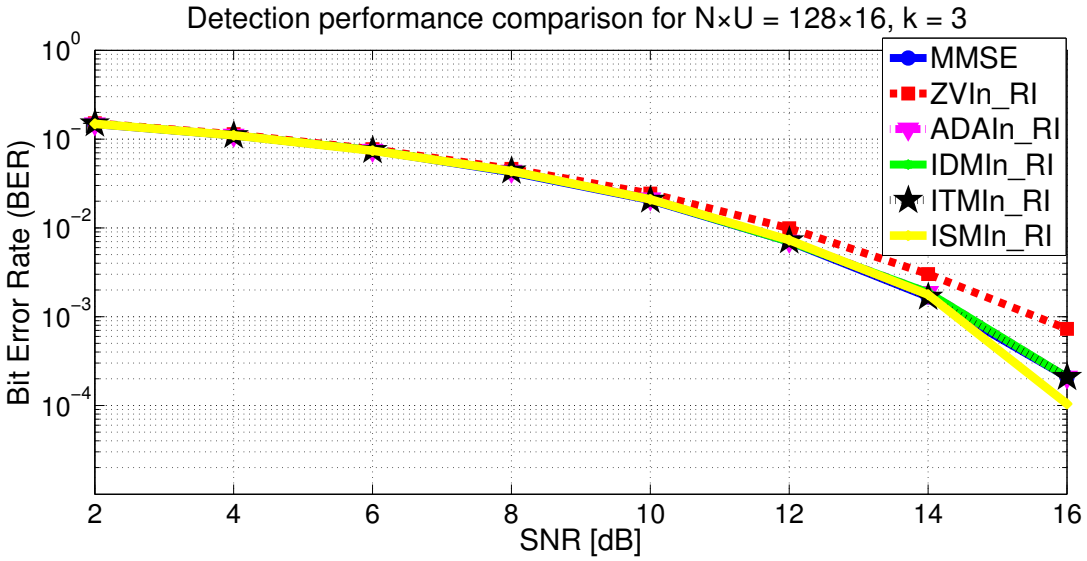

Figure 2.13: BER performance comparison of the RI detector under different initialization strategies (ZVIn_RI, ADAIn_RI, IDMIn_RI, ITMIn_RI, ISMIn_RI) for a 128×16 uplink massive MIMO system with $k = 3$ iterations and 64-QAM modulation. The results demonstrate how initialization choice impacts convergence and detection accuracy across SNR values, with structured approaches outperforming simple zero-vector initialization.

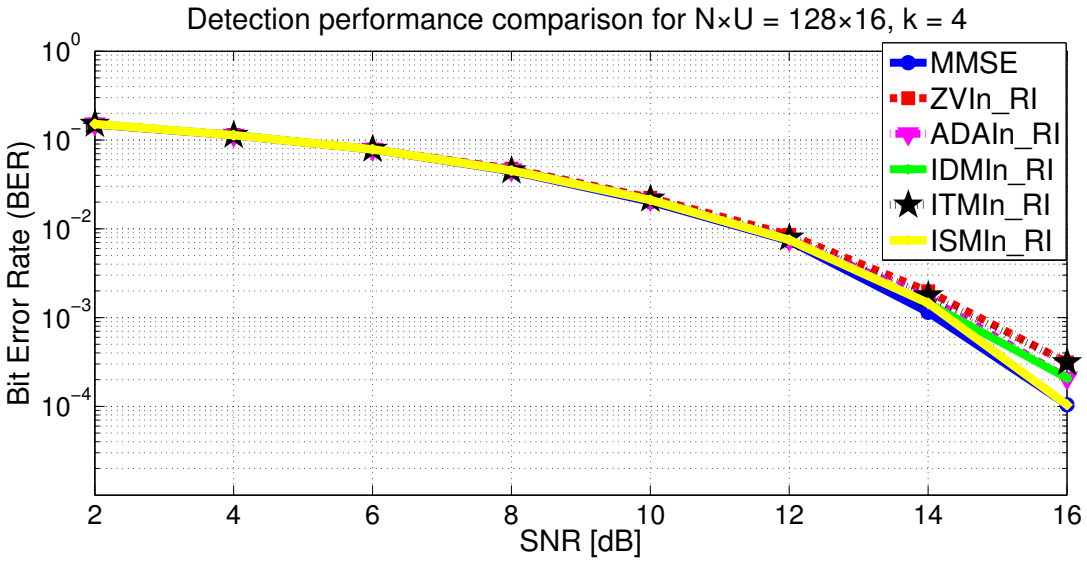

Figure 2.14: BER performance of the RI detector with multiple initialization strategies (ZVIn_RI, ADAIn_RI, IDMIn_RI, ITMIn_RI, ISMIn_RI) for a 128×16 uplink massive MIMO system using $k = 4$ iterations under 64-QAM modulation. The results highlight how initialization choice influences detection accuracy across SNR values, with structured initialization methods achieving improved convergence compared to the zero-vector baseline.

RI with structured initialization at $U = 32$: Figures 2.15–2.17 show the same ranking at 128×32: ISMIn–RI is best, followed by DI/ADAIn/ITMIn, with ZVIn generally the weakest. The ordering is stable across SNRs and iteration budgets, indicating initializer selection is the primary lever for RI's accuracy–latency trade-off.

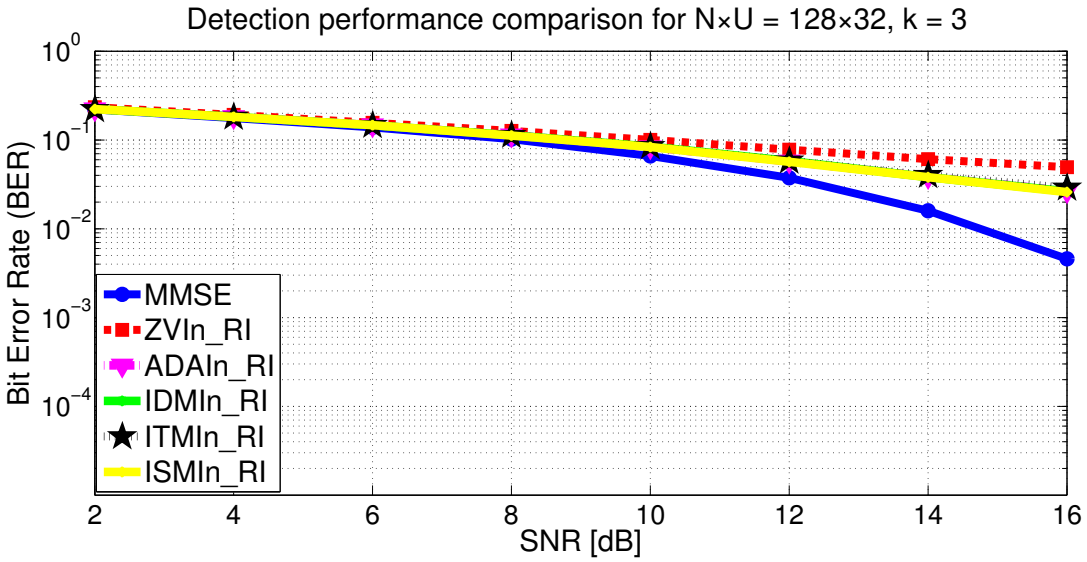

Figure 2.15: BER performance of the RI detector with different initialization strategies (ZVIn_RI, ADAIn_RI, IDMIn_RI, ITMIn_RI, ISMIn_RI) for a 128×32 uplink massive MIMO system using $k = 3$ iterations under 64-QAM modulation. The results illustrate how initialization choice affects detection accuracy across SNR values, with structured approaches yielding improved convergence compared to the zero-vector baseline.

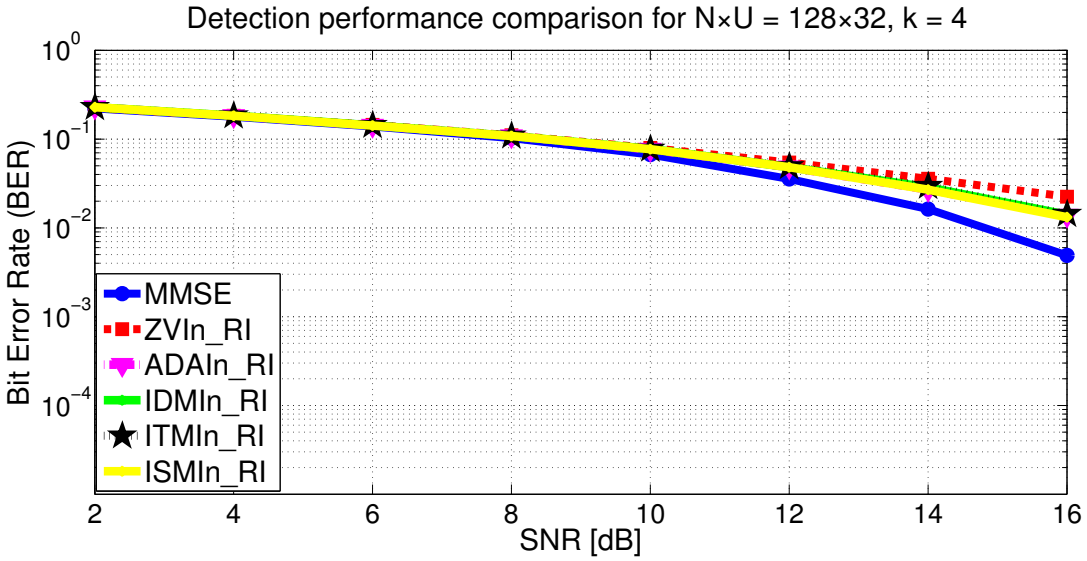

Figure 2.16: BER performance comparison of the RI detector with different initialization strategies (ZVIn_RI, ADAIn_RI, IDMIn_RI, ITMIn_RI, ISMIn_RI) for a 128×32 uplink massive MIMO system using $k = 4$ iterations under 64-QAM modulation. The results illustrate how initialization choice affects detection accuracy across SNR values, with structured approaches yielding improved convergence compared to the zero-vector initialization.

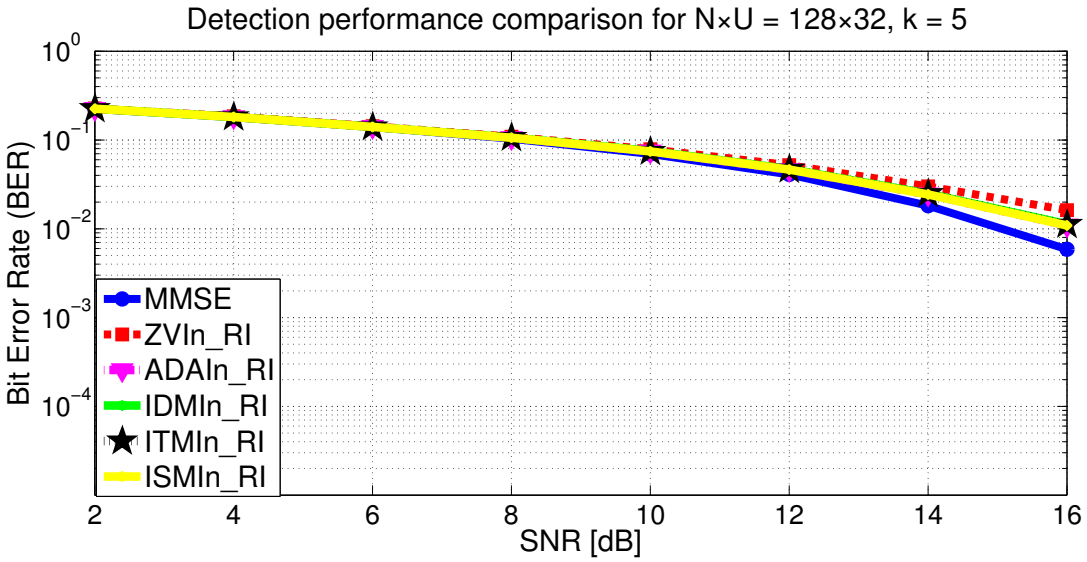

Figure 2.17: BER performance of the RI detector with different initialization strategies (ZVIn_RI, ADAIn_RI, IDMIn_RI, ITMIn_RI, ISMIn_RI) for a 128×32 uplink massive MIMO system using $k = 5$ iterations under 64-QAM modulation. The results emphasize how initialization choice impacts convergence and detection accuracy across SNR values, with structured approaches consistently outperforming the zero-vector baseline as iteration depth increases.

Operation-count context for RI: Figure 2.18 indicates that initialization overheads are minor relative to the iteration cost, so stronger initializations (especially ISMIn) improve overall efficiency by reducing the iterations needed to reach a target BER.

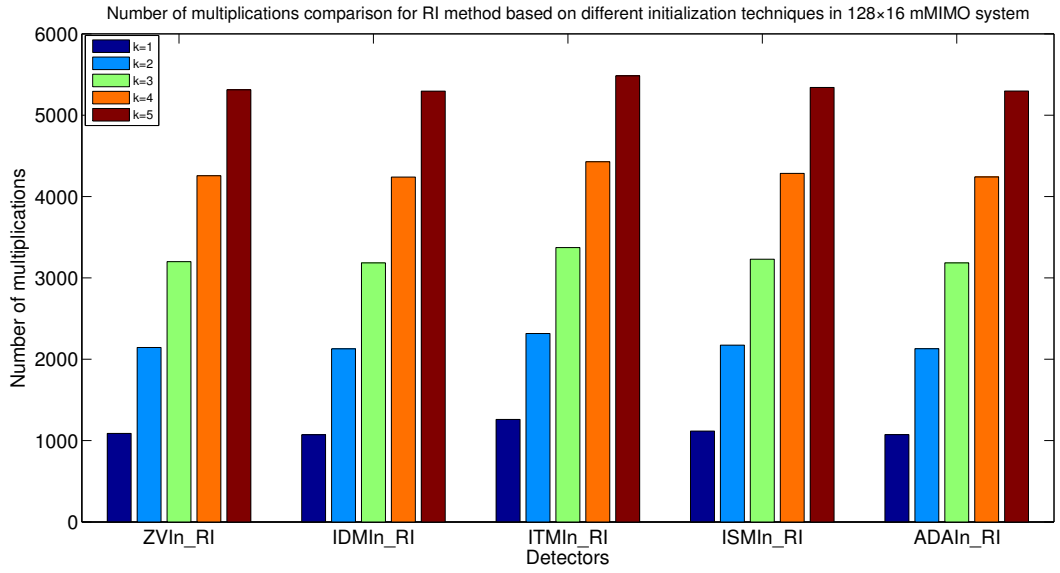

Figure 2.18: Multiplication counts of the RI detector under different initialization strategies (ZVIn_RI, IDMIn_RI, ITMIn_RI, ISMIn_RI, ADAIn_RI) for a 128×16 uplink massive MIMO system. The bars show how computational complexity scales with the iteration index k, indicating that structured initialization methods generally incur higher multiplication counts as k increases compared to the zero-vector baseline.

Synthesis and implications: (i) One CG step followed by $k = 3$–4 SOR sweeps attains near-MMSE BER at both loading factors and outperforms standalone JA, SOR, and CG at comparable operation counts; tuning ω near its empirical optimum (≈ 1.2 for $\psi = 4$) is key. (ii) For RI, initializer choice dominates at fixed k: ISMIn consistently delivers the best BER with the fewest iterations, while DI and ADAIn are strong alternatives; since iteration cost outweighs initialization cost, stronger initializations are generally preferable. (iii) Operation-count comparisons corroborate that hybrids (CG–SOR) and structured-initialized RI yield the best accuracy–latency balance at small iteration budgets.

2.5 Implementation Considerations

This section consolidates best practices for deploying inversion-free uplink massive MIMO detectors CG, SOR, GS, RI with structured initialization, and hybrid CG–SOR across software and hardware targets. It emphasizes finite-precision robustness, matrix-free dataflow, scheduling, initialization, parameter tuning, parallel mapping, and E2E system considerations for CPU/GPU/FPGA/ASIC platforms.

2.5.1 Finite-Precision and Numerical Stability

Fixed vs. floating point. FP32 provides robust convergence across SNR and channel conditions and simplifies prototyping; mixed-precision variants lower power by reserving higher precision for inner products and residual updates. For FPGA/ASIC, fixed-point datapaths with widened accumulators (for inner products, step-size computations) and unbiased/stochastic rounding mitigate saturation and rounding drift in CG/GS/SOR loops [68, 72, 83, 84, 85].

Scaling and preconditioning. Column normalization of H, diagonal scaling with $D = \mathrm{diag}(W)$, and simple JA preconditioning improve conditioning and accelerate SOR/GS/RI under antenna correlation and finite-precision constraints [72, 85].

Residual maintenance. Periodically refresh $r = y_{\mathrm{MF}} - W\hat{s}$ (e.g., every 4–8 iterations or at stage boundaries, CG→SOR) to control error accumulation in finite precision; this is especially useful for CG with soft-output extensions and matrix-free implementations [68, 83].

2.5.2 Matrix-Free Dataflow and Memory Efficiency

Avoid explicit W. Realize $Wx = H^H(Hx) + \sigma^2 x$ via two GEMVs, reducing storage from $\mathcal{O}(U^2)$ to $\mathcal{O}(NU)$, improving cache locality, and unifying compute kernels across CG/GS/-SOR/RI [86, 87].

Streaming and tiling. Tile H to fit caches/SRAM; stream rows for $z = Hx$, then a transposed view for $u = H^H z$, and add $\sigma^2 x$ in-place. Double-buffer tiles to overlap compute and memory transfers on GPUs/FPGAs [86, 87].

Accumulator headroom. Provision 6–8 extra bits in fixed-point MAC chains and calibrate scale factors per coherence interval to minimize overflow while preserving dynamic range for soft-output metrics [68, 83].

2.5.3 Control Flow and Scheduling

CG warm-start. One–two CG iterations (CG1/CG2) provide strong initial estimates with low overhead; fuse inner products, step-size, AXPY, and residual refresh; then switch to SOR/GS for terminal refinement to exploit rapid local convergence [72, 83].

SOR/GS sweeps. Use in-place forward (and optional backward) sweeps over $m = 1, \ldots, 2U$ to leverage immediately updated components; expose relaxation ω at runtime or select from a load-factor LUT ($\psi = N/U$) using quasi-optimal rules [72].

RI updates. Fuse matrix-free $W\hat{s}$ with the AXPY-style update $\hat{s}^{(k+1)} = \hat{s}^{(k)} + \omega\big(y_{\mathrm{MF}} -$

$W\hat{s}^{(k)}\big)$, and enable early stopping based on absolute/relative residual norms with hysteresis to avoid premature termination under noise-variance mismatch [78, 87].

2.5.4 Structured Initialization for RI

Inverse-stair (ISMIn). Construct a stair-structured S from W with alternating zeroed off-diagonals and compute $S^{-1} = D_s^{-1}(2D_s - S)D_s^{-1}$ with $D_s = \mathrm{diag}(S)$. Cache S and D_s^{-1} across the coherence interval. ISMIn achieves superior BER/iteration versus diagonal and tridiagonal approximations at similar setup cost [82, 85, 88].

Diagonal (DI) and ADAIn. DI uses $\hat{s}^{(0)} = D^{-1}y_{\mathrm{MF}}$ with $D = \mathrm{diag}(W)$ for a robust, low-cost start. ADAIn uses $\hat{s}^{(0)} = \hat{\omega}_0\, y_{\mathrm{MF}}$ with $\hat{\omega}_0 \approx 1/(N+U)$ to minimize setup when iteration budgets are tight [78].

2.5.5 Parameter Tuning Strategies

SOR relaxation ω. Use ψ-aware presets (e.g., $\omega \approx 1.1$–1.3 for $\psi \approx 4$–8) and a short per-frame calibration over a few candidates using residual norms; this mirrors theory for SPD systems and empirical optima (e.g., $\omega \approx 1.2$ at 128×32) [72].

Iteration budgets and early stopping. Favor "few high quality" iterations: CG1$\rightarrow$SOR$_{3\text{–}4}$ or ISMIn–RI$_{3\text{–}5}$. Terminate when $\|r^{(k)}\|_2 \le \tau_{\mathrm{abs}}$ or $\|r^{(k)}\|_2/\|r^{(0)}\|_2 \le \tau_{\mathrm{rel}}$; log realized iteration counts to inform higher-layer schedulers [83, 87].

2.5.6 Latency, Complexity, and Parallelism

Per-iteration cost. Matrix-free methods cost two GEMVs per iteration ($\mathcal{O}(NU)$); hybrid flows reduce the required iterations to reach near-MMSE BER [72, 83, 86].

Latency hiding. Overlap Hx and $H^H z$ GEMVs with AXPY/post-ops via command queues/streams; precompute and reuse $y_{\mathrm{MF}} = H^H y$ across the coherence time; batch users/-subcarriers to amortize kernel launch overhead on GPUs [86].

Soft-output extensions. For LLRs, reuse CG/GS residuals and diagonal approximations to form per-layer post-equalization signal to noise ratio (SNR) estimates with minimal extra cost [83, 85].

2.5.7 Target Mapping: CPU, GPU, FPGA, ASIC

Multicore CPU. Parallelize GEMVs over N (Hx) and over U/N ($H^H z$) with thread-local reductions; apply NUMA-aware placement; use BLAS with fused AXPY to cut memory

traffic [87].

GPU. Use batched GEMV, tile reuse in shared memory, persistent kernels for iterative loops, and stream-level concurrency across subcarriers/users; keep H resident on device [86].

FPGA/ASIC. Implement systolic MAC arrays streaming tiles from on-chip SRAM; use fixed-point with wide accumulators and saturating adders; expose ω, k, precision, and thresholds via CSRs; employ approximate relaxations and finite-precision–validated CG steps [68, 83, 84].

2.5.8 Robustness and System Integration

Channel correlation and HWI. Diagonal loading (slightly larger $\sigma^2 I$), stronger initialization (ISMIn or CG2), and one–two extra terminal sweeps stabilize convergence under spatial correlation and finite-precision distortions [68, 85].

Noise-variance mismatch. Re-estimate σ^2 from pilots; when uncertain, prefer robust flows (CG→SOR or ISMIn–RI) and slightly tighter early-stop thresholds to avoid under-iterating at low SNR [87, 78].

Distributed/BB scaling. Partition H and y across nodes (antenna sharding) with local GEMVs and global reductions of partial sums; maintain deterministic iteration counts for predictable latency in disaggregated RANs [87].

2.5.9 Default Modes

Throughput-optimized. CG1 $\rightarrow$ $\mathrm{SOR}_{3\text{–}4}$, ψ-aware ω, matrix-free products, diagonal scaling, early stop; FP16/FP32 on GPU; fixed-point on FPGA/ASIC with wide accumulators [83, 86].

Lightweight. $\mathrm{ISMIn–RI}_{3\text{–}5}$, $\omega \approx 1/(N+U)$ or short calibration; DI fallback if ISMIn setup is constrained; residual refresh every 3–4 iterations [82, 78].

High-load/correlated channels. CG2 $\rightarrow$ $\mathrm{SOR}_{3\text{–}5}$ or $\mathrm{ISMIn–RI}_{4\text{–}6}$ with diagonal loading; brief ω calibration; enable stochastic rounding on hardware [68, 83, 85].

2.6 Summary of Contributions

This thesis delivers a validated framework for low-complexity, inversion-free uplink massive MIMO detection, achieving near-MMSE performance through algorithmic innovation and comprehensive evaluation.

C1. Hybrid CG–SOR Detector

Hybrid detector combines one CG warm-up iteration with SOR sweeps, eliminating matrix inversion. Achieves near-MMSE BER at $\mathcal{O}(NU)$ complexity, outperforming standalone CG, GS, SOR across Monte Carlo evaluations.

C2. Load-Aware Relaxation Parameters

Relaxation parameters tuned to system loading ($\psi = N/U$) with residual-norm early-stopping ensure bounded iterations and predictable latency under spatial correlation.

C3. Structured RI Initialization

ISMIn initialization outperforms DI, ADAIn, and tridiagonal methods, delivering optimal BER-per-iteration trade-off under high loading and 64-QAM.

C4. Matrix-Free Dataflow Analysis

Streaming GEMV formulation $\mathbf{Ws} = \mathbf{H}^H(\mathbf{Hs}) + \sigma^2\mathbf{s}$ maximizes memory efficiency and arithmetic intensity across iterative methods.

C5. System-Level Performance Validation

Theoretical complexity models confirmed through extensive simulations across SNR, loading (16–32 users), and modulations, demonstrating superior convergence vs. standard iterative detectors.

Core Impact: C1–C3 establish novel algorithms (hybrid CG–SOR, load-adaptive tuning, ISMIn initialization); C4–C5 provide rigorous analysis confirming practical deployment feasibility for 5G/6G uplink processing.

2.7 Conclusion

This chapter has established a principled, inversion-free methodology for uplink massive MIMO detection that bridges algorithmic innovation with hardware conscious implementation, achieving near-MMSE reliability at substantially reduced computational cost and with

predictable, low-latency execution across contemporary platforms. By casting MMSE detection as matrix-free iterative solvers CG, GS, SOR, and RI, the framework consistently converged in a small number of iterations over representative SNRs, loading factors, and modulation orders, while preserving numerical stability under finite-precision constraints.

Real-time capability rests on three core pillars:

1. A cascaded CG warm start followed by SOR refinement, coupling stable global progress with fast local convergence without explicit matrix inversion.

2. Load-aware parameterization: relaxation factors guided by the loading factor $\psi = N/U$, augmented by brief per-frame residual-based calibration and principled early-stopping rules grounded in absolute and relative residual norms.

3. A unified, matrix-free realization of $\mathbf{Ws} = \mathbf{H}^H(\mathbf{Hs}) + \sigma^2\mathbf{s}$, optimized via tiling, double buffering, and kernel fusion to maximize arithmetic intensity and minimize memory traffic.

Complementary fixed-point safeguards dynamic range scaling per coherence interval, widened accumulators for inner products and step-size updates, and unbiased/stochastic rounding—mitigate quantization drift and enable aggressive word-length reduction with minimal BER impact. From a systems perspective, portable execution blueprints were articulated for CPUs, GPUs, and FPGA/ASICs, encompassing multithreaded and batched/persistent execution, shared-memory tiling, systolic MAC arrays, and control-register exposure for ω, iteration budget k, and precision. System level integration techniques including reuse of $\mathbf{y}_{\mathrm{MF}} = \mathbf{H}^H\mathbf{y}$ across the coherence time, user/subcarrier batching, and antenna-sharded reductions sustain scalable throughput and deterministic latency in centralized and distributed RAN deployments. The resulting design playbooks (CG1→$\mathrm{SOR}_{3\text{–}4}$, ISMIn–$\mathrm{RI}_{3\text{–}5}$, and CG2→$\mathrm{SOR}_{3\text{–}5}$ or ISMIn–$\mathrm{RI}_{4\text{–}6}$ with optional diagonal loading) provide actionable defaults for throughput-focused, lightweight, and high-load/correlated regimes.

Future Directions

Building on this foundation, several avenues can further enhance practicality and performance:

- **Adaptive policies**: data-driven tuning of relaxation parameters and stopping criteria under varying channel conditions and hardware states to refine iteration control while preserving convergence guarantees.

- **Cross-layer integration**: leveraging realized iteration counts within scheduling and HARQ to optimize latency–throughput trade-offs and quality of service (QoS) targets.
- **Robustness under impairments**: extending validation to correlated channels and nonidealities (e.g., phase noise, quantized CSI, non-Gaussian interference) to harden parameter presets.
- **Hardware algorithm co-design**: exploring sparsity-aware tiling, mixed-precision inner products, and runtime autotuning to further reduce overhead while maintaining accuracy.

Closing reflection: The chapter delivers a cohesive, implementation-ready strategy for high-performance massive MIMO detection without matrix inversion. By aligning algorithmic structure, numerical parameterization, and hardware-aware execution, the proposed approach advances scalable, real time 5G/6G uplink BB processing and provides a reproducible pathway from theory to field deployment.

Chapter 3

Enhanced Channel Estimation Techniques

3.1 Introduction

In uplink massive MIMO systems, signal detection and channel estimation are tightly coupled processes that jointly determine communication reliability and throughput. Chapter 2 introduced inversion-free iterative detectors such as CG, SOR, and RI which rely on accurate CSI to decode transmitted symbols. However, the effectiveness of these algorithms is inherently limited by the quality of the underlying channel estimate, making robust channel estimation a critical precursor to successful detection.

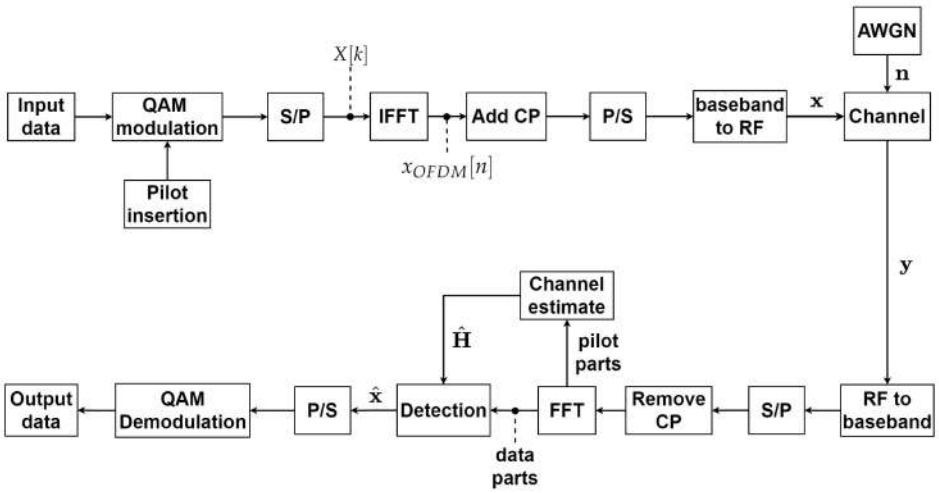

Figure 3.1: Block diagram of an OFDM communication system showing the interaction between channel estimation and signal detection.

Figure 3.1 illustrates a typical Orthogonal Frequency-Division Multiplexing (OFDM)-based receiver chain, where pilot symbols are inserted at the transmitter and used at the receiver to estimate the channel response. This estimate enables equalization and accurate signal detection. Any error in CSI propagates directly into the detection stage, degrading BER and increasing the iteration burden for convergence in low-complexity detectors.

In massive MIMO environments, accurate channel estimation is challenged by:

- **High-dimensional channel matrices**, which escalate computational cost and demand extensive pilot overhead;
- **Pilot contamination and spatial correlation**, especially in densely deployed or multi-cell systems;
- **Hardware imperfections**, such as quantization noise, RF nonidealities, and time-varying channels that hinder stable CSI acquisition.

As CSI accuracy establishes the foundation for detector performance, this chapter introduces advanced channel estimation techniques to enhance robustness, resolution, and scalability. The proposed methods include:

1. **Statistical covariance enhancements**, such as shrinkage-based estimation, which improve reliability under limited sample support and enable more accurate DOA and pilot-based CSI schemes.
2. **RIS-assisted CSI recovery** using structured sparse Bayesian learning. These methods exploit angular sparsity and partial common support to drastically reduce pilot overhead, with recent architectures proposing multi-stage hybrid estimation for scalable mmWave RIS-enabled systems.
3. **Learning-based and sparse modeling strategies**, including deep neural networks (e.g., RNNs, attention mechanisms) and structured SBL, which adapt to dynamic environments while maintaining estimation accuracy under practical constraints.

This chapter is organized as follows: Section 3.3 revisits classical estimation limitations in multiuser massive MIMO. Section 3.4 introduces improved covariance techniques and DOA-based enhancements. Section 3.5 explores RIS-assisted Bayesian and learning-based frameworks. Section 3.6 presents a comparative analysis of performance, complexity, and implementation feasibility across modern platforms.

By enhancing CSI estimation, this chapter lays the groundwork for truly robust, low-complexity detection closely tying Chapters 2 and 3 into a unified, scalable framework for next-generation wireless BB processing.

3.2 Executive Summary

This chapter introduces pilot-efficient channel estimation techniques to enhance the low-complexity, inversion-free detectors (CG, SOR, RI) presented in Chapter 2. By improving CSI fidelity under limited pilot resources, spatial correlation, and hardware constraints, the proposed methods deliver faster detector convergence, higher post-equalization SINR, and stable error-rate performance in uplink massive MIMO–OFDM systems.

Key Contributions

- Shrinkage-enhanced covariance modeling for robust DOA estimation and improved subspace conditioning with few snapshots.
- Structured sparse recovery (TLSR-OMP, MAP-TLSOMP) tailored for RIS-cascaded channels, reducing pilot overhead while maintaining NMSE accuracy.
- System-level analysis linking CSI accuracy directly to detector outcomes such as BER, SINR, and convergence iterations.

Core Results

- Pilot efficiency: 5–10× reduction in RIS scenarios with near-LS accuracy.
- Resolution gains: Covariance shrinkage improves angular resolution by 20–30% under small-sample regimes.
- Detector acceleration: Enhanced CSI reduces CG/SOR iterations by $\sim$50%, boosting convergence speed and SINR.

Practical Impact

The framework enables scalable, hardware-aware CSI acquisition compatible with hybrid beamforming and RIS architectures. By bridging theoretical detector performance with

practical deployment, it provides a pathway toward robust, EE baseband processing in 5G-Advanced and 6G uplink systems.

3.3 Channel Estimation in Multiuser Massive MIMO

Accurate channel estimation is fundamental to uplink multiuser massive MIMO systems enabling tasks such as detection, BF, and SE improvement. Under TDD operation, the BS uses pilot signals from U single-antenna users to estimate the channel matrix $\mathbf{H}$ via:

$$\mathbf{y} = \mathbf{H}\mathbf{s} + \mathbf{n},$$

where $\mathbf{y} \in \mathbb{C}^{N\times 1}$ is the stacked received pilot signal at the BS, $\mathbf{H} \in \mathbb{C}^{N\times U}$ is the uplink channel matrix from U single-antenna users, $\mathbf{s} \in \mathbb{C}^{U\times 1}$ is the pilot symbol vector, and $\mathbf{n} \sim \mathcal{CN}(\mathbf{0}, \sigma^2\mathbf{I}_N)$ denotes spatially white AWGN with variance σ^2.

While massive MIMO systems yield substantial throughput gains [5, 89], they also magnify estimation challenges in the presence of practical constraints namely, pilot contamination, limited pilot resources, spatial correlation, and hardware nonidealities.

Key Challenges

- **Pilot contamination:** Reusing pilot sequences across adjacent cells due to limited coherence time/bandwidth introduces interference that persists even as $N \to \infty$ [90], [91].
- **Spatial correlation:** Closely spaced antenna elements reduce effective channel rank, complicating covariance-based estimation [92].
- **HWI:** Phase noise, quantization errors, and RF nonlinearity distort pilots, undermining estimation accuracy [93].
- **Pilot scarcity:** In high-mobility or wideband deployments, the coherence time is significantly reduced, limiting the number of available orthogonal pilot symbols and necessitating pilot reuse, which degrades CSI quality and detection performance [94].

Classical Estimation Approaches

- **LS:** Lightweight and easy to implement, but highly sensitive to noise, contamination, and poor conditioning.

- **Linear MMSE (LMMSE):** Offers better MSE performance by incorporating covariance information, though at the cost of high computational complexity tied to large matrix inversions.

Advanced Estimation Strategies

Covariance-Aided Estimators and Shrinkage Techniques Shrinkage approaches blend sample covariance with structured models (e.g., scaled identity, Toeplitz forms) to stabilize estimation when sample size is low [95, 96]. These estimators improve conditioning and support robust LMMSE performance.

Subspace and DOA Techniques For arrays with structured angular support (e.g., ULAs/UPAs), subspace methods such as MUSIC or forward backward averaging facilitate DOA detection. When combined with shrinkage, they enhance channel estimation under coherent paths or limited snapshots.

Sparse and Bayesian Estimation Frameworks Complex channels, particularly in RIS and mmWave contexts, exhibit angular sparsity. Structured Sparse Bayesian Learning (like UAMPSBL–PCI) and off-grid sparse recovery exploit this characteristic to estimate channels efficiently with reduced pilot overhead [97].

Pilot Coordination Across Cells Optimizing pilot assignment via matching or fractional programming helps suppress pilot contamination, improving MSE and downstream detection performance [91, 98].

Learning-Aided Channel Estimation Deep learning solutions including CNNs, residual networks, and conditional GANs can adaptively refine CSI estimation and pilot design, demonstrating resilience to contamination and HWI [30, 99].

Decision-Directed and Mixed-Timescale Updates Reliable detected data symbols can be used to refine channel estimates iteratively (decision-directed estimation). Using a mixed timescale strategy allows sparse updates of spatial statistics while tracking fast channel fluctuations efficiently, easing real time implementation burdens.

Impact on Inversion-Free Detector Performance

CSI fidelity plays a crucial role in governing the convergence and error-rate performance of iterative detection methods like CG, SOR, and RI (detailed in Chapter 2). Incorporating CSI estimation uncertainty into stopping criteria—e.g., via SINR-aware thresholds can provide robustness while minimizing computational redundancies.

Design Recommendations

- When $T \approx N$, employ shrinkage-based covariance estimation with structured targets aligned to antenna geometry.
- Use subspace or DOA priors to filter channel support and reduce pilot usage or variance.
- Coordinate pilot assignment across cells to limit contamination in dense deployments.
- Leverage sparse Bayesian methods for RIS or mmWave scenarios to reduce pilot overhead while maintaining high-quality CSI.
- Integrate learning-based estimators and decision-directed updates for agile adaptation and CSI refinement in dynamic environments.

3.4 Direction of Arrival (DOA) Estimation Methods

DOA estimation is a cornerstone of array signal processing, enabling spatial filtering, user separation, and BF in massive MIMO systems. Accurate DOA estimation improves CSI, interference rejection, and the performance of low-complexity detectors in uplink multiuser scenarios. This section presents robust DOA pipelines focusing on (i) modified covariance reconstruction for coherent scenes, (ii) covariance shrinkage for snapshot-limited regimes, and (iii) full derivations of MUSIC and Root-MUSIC together with reproducible pseudocode for enhanced estimators.

3.4.1 Modified Covariance Matrix Approach

When sources are coherent (e.g., strong multipath or pilot reuse) or when very few snapshots are available, the sample covariance matrix becomes rank deficient and classical subspace methods (MUSIC/ESPRIT) fail. Two commonly used remedies are forward–backward (FB) averaging and conjugate reconstruction (CR) with an exchange matrix $\mathbf{J}$.

Let the sample covariance be

$$\widehat{\mathbf{R}} \;=\; \frac{1}{T}\sum_{t=1}^{T}\mathbf{x}[t]\mathbf{x}^{H}[t], \tag{3.1}$$

where $\mathbf{x}[t] \in \mathbb{C}^M$ is the array observation vector, i.e., the simultaneous measurements across all M sensors at time index t. Thus, a snapshot refers to this full M-dimensional array measurement collected at each of the T time instants.

Forward–Backward (FB) averaging. For a ULA with exchange (reversal) matrix $\mathbf{J}$,

$$\widehat{\mathbf{R}}_{\mathrm{FB}} \;=\; \tfrac{1}{2}\Big(\widehat{\mathbf{R}} + \mathbf{J}\,\widehat{\mathbf{R}}^{T}\mathbf{J}\Big). \tag{3.2}$$

where $\mathbf{J}$ is the $M \times M$ exchange matrix (antidiagonal of ones) for ULA reversal.

Conjugate reconstruction (CR). The conjugate reconstruction symmetrizes via conjugation,

$$\widehat{\mathbf{R}}_{\mathrm{CR}} \;=\; \widehat{\mathbf{R}} + \mathbf{J}\,\widehat{\mathbf{R}}^{*}\mathbf{J}. \tag{3.3}$$

where $\bar{\mathbf{R}}$ denotes the complex conjugate of $\mathbf{R}$.

Both operations restore (or improve) the effective rank of the signal covariance in many coherent scenarios and often enable the subsequent subspace decomposition to separate signal and noise subspaces more reliably (see [100, 101] for empirics and analysis).

3.4.2 Covariance Matrix Shrinkage Estimation

When T is comparable to M (the number of sensors), the sample covariance $\widehat{\mathbf{R}}$ is noisy and ill-conditioned. Linear shrinkage stabilizes the estimate by blending $\widehat{\mathbf{R}}$ with a structured target $\mathbf{F}$:

$$\widehat{\mathbf{R}}_{\mathrm{sh}} \;=\; (1-\rho)\,\widehat{\mathbf{R}} + \rho\,\mathbf{F}, \qquad 0 \le \rho \le 1. \tag{3.4}$$

where $\rho \in [0,1]$ is the shrinkage intensity (coefficient), $\widehat{\mathbf{R}}$ is the sample covariance, and $\mathbf{F}$ is the structured target matrix (e.g., scaled identity $\beta\mathbf{I}_M$ or Toeplitz form).

Useful choices for $\mathbf{F}$ include scaled identity $\sigma_0^2\mathbf{I}$, Toeplitz (for ULA), or physics-inspired Kronecker/separable targets. Analytic prescriptions for ρ Rao-Blackwell Ledoit-Wolf (RBLW) and Oracle Approximating Shrinkage (OAS) are available and achieve near-optimal MSE under Gaussian sampling; tuned variants can further optimize DOA objectives (peak sharpness) in small-T settings. For the theoretical development and practical recipes see Chen et al. and related literature [102].

3.4.3 MUSIC: Mathematical Derivation

We present a concise derivation of MUSIC suitable for thesis exposition.

Signal model. Consider K narrowband far-field sources impinging on an M-element array:

$$\mathbf{x}[t] = \mathbf{A}\,\mathbf{s}[t] + \mathbf{n}[t], \qquad t = 1, \ldots, T, \tag{3.5}$$

where $\mathbf{A} = [\mathbf{a}(\theta_1), \ldots, \mathbf{a}(\theta_K)] \in \mathbb{C}^{M\times K}$ is the steering matrix, $\mathbf{s}[t] \in \mathbb{C}^K$ the source vector, and $\mathbf{n}[t] \sim \mathcal{CN}(\mathbf{0}, \sigma^2\mathbf{I}_M)$ is spatially white noise.

Covariance and subspace split. The population covariance is

$$\mathbf{R} \;=\; \mathbb{E}\{\mathbf{x}\mathbf{x}^H\} \;=\; \mathbf{A}\mathbf{R}_s\mathbf{A}^H + \sigma^2\mathbf{I}_M, \tag{3.6}$$

where $\mathbf{R}_s = \mathbb{E}\{\mathbf{s}\mathbf{s}^H\} \in \mathbb{C}^{K\times K}$ denotes the covariance matrix of the K source signals. For non-degenerate sources, $\mathbf{R}_s$ has full rank K, allowing the eigendecomposition of $\mathbf{R}$ to clearly separate the signal and noise subspaces.

$$\mathbf{R} = \mathbf{E}_s\mathbf{\Lambda}_s\mathbf{E}_s^H + \mathbf{E}_n\mathbf{\Lambda}_n\mathbf{E}_n^H, \tag{3.7}$$

where $\mathbf{E}_s \in \mathbb{C}^{M\times K}$ contains the eigenvectors associated with the K dominant eigenvalues and therefore spans the signal subspace, while $\mathbf{E}_n \in \mathbb{C}^{M\times(M-K)}$ contains the remaining eigenvectors spanning the noise subspace. The diagonal matrices $\mathbf{\Lambda}_s$ and $\mathbf{\Lambda}_n$ collect the corresponding signal and noise eigenvalues, respectively.

Orthogonality principle. For any true DOA θ_k we have $\mathbf{a}(\theta_k) \in \text{span}(\mathbf{E}_s)$, hence

$$\mathbf{E}_n^H\mathbf{a}(\theta_k) = \mathbf{0}. \tag{3.8}$$

MUSIC pseudo-spectrum. Using an estimate of the noise eigenvectors $\widehat{\mathbf{E}}_n$ computed from $\widehat{\mathbf{R}}$ (or $\widehat{\mathbf{R}}_{\text{FB}}$, $\widehat{\mathbf{R}}_{\text{CR}}$, or $\widehat{\mathbf{R}}_{\text{sh}}$), the MUSIC pseudospectrum is defined as

$$P_{\text{MUSIC}}(\theta) \;=\; \frac{1}{\mathbf{a}^H(\theta)\,\widehat{\mathbf{E}}_n\widehat{\mathbf{E}}_n^H\,\mathbf{a}(\theta)}. \tag{3.9}$$

where $\mathbf{a}(\theta) \in \mathbb{C}^{M\times 1}$ is the array steering vector for direction θ.

Peaks in $P_{\text{MUSIC}}(\theta)$ correspond to DOA estimates. Standard perturbation results (Davis–Kahan) show that, as $T \to \infty$, the subspace estimates converge and MUSIC is consistent under usual identifiability conditions (distinct θ_k, sufficient SNR) [103, 104].

3.4.4 Root-MUSIC: Mathematical Derivation

Root-MUSIC trades the angular grid search for polynomial rooting and is particularly attractive for ULAs.

ULA steering polynomial. For a ULA with inter-element spacing d and wavelength λ, the steering vector admits the Vandermonde form

$$\mathbf{a}(\theta) = \left[1,\ z,\ z^2, \ldots,\ z^{M-1}\right]^T, \qquad z = e^{-j\psi}, \quad \psi = \tfrac{2\pi d}{\lambda}\sin\theta. \tag{3.10}$$

where d is inter-element spacing and λ is wavelength.

Noise projector polynomial. Let $\widehat{\mathbf{P}}_n = \widehat{\mathbf{E}}_n \widehat{\mathbf{E}}_n^H$ be the estimated noise projector. Define

$$Q(z) \ = \ \mathbf{a}^H(z^{-1})\, \widehat{\mathbf{P}}_n\, \mathbf{a}(z). \tag{3.11}$$

Because $\widehat{\mathbf{P}}_n$ is Hermitian and Toeplitz structured in the ULA Vandermonde basis, $Q(z)$ is a Laurent polynomial

$$Q(z) \ = \ \sum_{\ell=-(M-1)}^{M-1} c_\ell z^\ell, \tag{3.12}$$

where the coefficients c_ℓ are extracted from the diagonals of the noise-subspace projector $\mathbf{P}_n = \mathbf{E}_n \mathbf{E}_n^H$, with each c_ℓ corresponding to the ℓ-th diagonal element of $\mathbf{P}_n$.

Root selection and mapping. Root-MUSIC finds the K roots $\{z_k\}$ of $Q(z)$ that lie closest to the unit circle (usually inside the unit disc), and maps them to DOAs via

$$\widehat{\theta}_k \ = \ \arcsin\!\left(-\frac{\angle z_k}{2\pi d/\lambda} \right), \qquad k = 1, \ldots, K. \tag{3.13}$$

where z_k are the K roots inside the unit circle closest to $|z| = 1$.

Root-MUSIC avoids fine angular scanning and typically yields numerically stable super-resolution for ULAs; FB/CR preprocessing improves stability for coherent scenarios [101, 103].

3.4.5 Pseudocode: Algorithms

Algorithm 5 MVDR Beamforming with Shrinkage Regularization

Input: Snapshot matrix $\mathbf{X} \in \mathbb{C}^{M \times T}$, steering vector $\mathbf{a}(\theta)$, angle grid Θ, shrinkage target $\mathbf{F}$

Output: Estimated directions of arrival (DOAs)

1: Compute sample covariance matrix:

$$\widehat{\mathbf{R}} = \tfrac{1}{T}\mathbf{X}\mathbf{X}^H$$

2: Estimate shrinkage coefficient ρ (e.g., RBLW, OAS, or tuned)

3: Form shrunk covariance matrix:

$$\widehat{\mathbf{R}}_{\text{sh}} = (1-\rho)\,\widehat{\mathbf{R}} + \rho\,\mathbf{F}$$

4: **for** $\theta \in \Theta$ **do**

5: Solve linear system:

$$\widehat{\mathbf{R}}_{\text{sh}}\,\mathbf{y} = \mathbf{a}(\theta)$$

(e.g., CG or Cholesky factorization)

6: Compute MVDR spatial spectrum:

$$P_{\text{MVDR}}(\theta) = \left(\mathbf{a}^H(\theta)\mathbf{y}\right)^{-1}$$

7: **end for**

8: **return** DOA estimates as the dominant peaks of $P_{\text{MVDR}}(\theta)$

Algorithm 6 MUSIC Spectrum Estimation (Subspace Search)

Input: Snapshots $\{\mathbf{x}[t]\}_{t=1}^{T}$, steering vector $\mathbf{a}(\theta)$, angle grid Θ, preprocessing option $\in \{\texttt{none}, \texttt{FB}, \texttt{CR}, \texttt{shrink}\}$

Output: Estimated DOAs from $P_{\text{MUSIC}}(\theta)$

$\widehat{\mathbf{R}} \leftarrow \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}[t]\mathbf{x}^{H}[t]$ ▷ Sample covariance matrix

Optional Preprocessing:

if preprocessing = `FB` **then**

  $\widehat{\mathbf{R}} \leftarrow \frac{1}{2}\left(\widehat{\mathbf{R}} + \mathbf{J}\,\widehat{\mathbf{R}}^{T}\mathbf{J}\right)$ ▷ Forward–Backward averaging

else if preprocessing = `CR` **then**

  $\widehat{\mathbf{R}} \leftarrow \widehat{\mathbf{R}} + \mathbf{J}\,\widehat{\mathbf{R}}^{*}\mathbf{J}$ ▷ Conjugate reconstruction

else if preprocessing = `shrink` **then**

  Compute shrinkage coefficient ρ and target matrix $\mathbf{F}$

  $\widehat{\mathbf{R}} \leftarrow (1-\rho)\widehat{\mathbf{R}} + \rho\,\mathbf{F}$ ▷ Shrinkage regularization

end if

Subspace Decomposition:

Eigendecompose $\widehat{\mathbf{R}} = \widehat{\mathbf{E}}_s\widehat{\mathbf{\Lambda}}_s\widehat{\mathbf{E}}_s^{H} + \widehat{\mathbf{E}}_n\widehat{\mathbf{\Lambda}}_n\widehat{\mathbf{E}}_n^{H}$ ▷ Signal and noise subspaces

Spectrum Evaluation:

for each $\theta \in \Theta$ **do**

  $P_{\text{MUSIC}}(\theta) \leftarrow \frac{1}{\mathbf{a}^{H}(\theta)\,\widehat{\mathbf{E}}_n\widehat{\mathbf{E}}_n^{H}\,\mathbf{a}(\theta)}$

end for

Identify DOA estimates as peaks of $P_{\text{MUSIC}}(\theta)$

return DOA estimates

Algorithm 7 Root-MUSIC for ULA

Input: Snapshots $\{\mathbf{x}[t]\}$, number of sources K, ULA spacing d, wavelength λ, preprocessing option $\in \{\texttt{none}, \texttt{FB}, \texttt{CR}, \texttt{shrink}\}$

Output: Estimated DOAs $\{\widehat{\theta}_k\}_{k=1}^{K}$

Step 1: Covariance Estimation

Compute sample covariance matrix $\widehat{\mathbf{R}}$

Apply optional preprocessing (FB/CR/shrink) to refine $\widehat{\mathbf{R}}$

Step 2: Noise Subspace Projection

Perform eigendecomposition of $\widehat{\mathbf{R}}$

Form projection matrix: $\widehat{\mathbf{P}}_n \leftarrow \widehat{\mathbf{E}}_n \widehat{\mathbf{E}}_n^H$

Step 3: Polynomial Construction

Build Laurent polynomial:

$$Q(z) = \mathbf{a}^H(z^{-1})\, \widehat{\mathbf{P}}_n\, \mathbf{a}(z)$$

Step 4: Root Selection

Find all roots of $Q(z)$

Select the K roots $\{z_k\}$ closest to the unit circle (inside)

Step 5: Angle Mapping

Map each root $z_k \mapsto \widehat{\theta}_k$ using Eq. (3.13)

Step 6: Output

return DOA estimates $\{\widehat{\theta}_k\}_{k=1}^{K}$

3.4.6 Remarks and Practical Guidelines

- Use FB or CR reconstruction when source coherence is suspected (multipath, pilot reuse). These operations often restore rank and improve subspace separability.
- Apply shrinkage when $T \lesssim M$; choose $\mathbf{F}$ to reflect array physics (Toeplitz target for ULAs) and use RBLW/OAS for analytic ρ, or tune ρ to sharpen beamformer peaks.
- Root-MUSIC is preferred for ULAs to avoid grid bias; MUSIC with shrinkage/FB preprocessing provides robust grid-search alternatives.
- Expose CSI uncertainty (e.g., via eigenvalue spread or residual norms) to the detection stage (Chapter 2) to adapt detector stopping criteria.

3.5 RIS-Assisted Channel Estimation

RIS enable nearly passive, programmable control of the propagation environment via phase-tunable meta-elements, boosting coverage, interference management, and EE in beyond-5G/6G systems [16, 105]. In uplink massive MIMO with RIS, a central challenge is to estimate the cascaded BS–RIS–user channels accurately despite high dimensionality and pilot constraints. Since detector performance (e.g., SINR, BER, and the convergence of low-complexity iterative detectors) is tightly coupled to CSI fidelity, robust channel estimation under practical constraints is critical [106, 107, 108].

3.5.1 Cascaded Channel Model

Consider an uplink multi-user mmWave system where the BS has M antennas, the RIS has N nearly passive reflecting elements, and each user is equipped with a single antenna. Let $\mathbf{G} \in \mathbb{C}^{M\times N}$ denote the BS–RIS channel and $\mathbf{h}_{r,u} \in \mathbb{C}^{N\times 1}$ the RIS–user channel for user u. The resulting cascaded channel is

$$\mathbf{H}_u \;=\; \mathbf{G}\,\mathrm{diag}(\mathbf{h}_{r,u}) \;\in\; \mathbb{C}^{M\times N}. \tag{3.14}$$

At mmWave, both $\mathbf{G}$ and $\mathbf{h}_{r,u}$ are well modeled by the SV representation, yielding a limited number of dominant angular components [109]. Introducing unitary discrete Fourier transform (DFT) dictionaries $\mathbf{U}_M \in \mathbb{C}^{M\times M}$ and $\mathbf{U}_N \in \mathbb{C}^{N\times N}$, we obtain the angular model

$$\mathbf{H}_u \;=\; \mathbf{U}_M\,\widetilde{\mathbf{H}}_u\,\mathbf{U}_N^H, \qquad \widetilde{\mathbf{H}}_u \text{ is sparse with } \mathcal{O}(L_G L_{r,u}) \ll MN, \tag{3.15}$$

where L_G and $L_{r,u}$ denote the number of dominant BS–RIS and RIS–user paths, respectively.

3.5.2 Pilot Signaling and CS Measurement Model

During training over Q slots, the RIS applies unit-modulus phase vectors $\{\boldsymbol{\theta}_q \in \mathbb{C}^N : |[\boldsymbol{\theta}_q]_n| = 1\}_{q=1}^{Q}$, forming $\boldsymbol{\Theta} \triangleq [\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_Q] \in \mathbb{C}^{N\times Q}$. Let $\mathbf{y}_{u,q} \in \mathbb{C}^M$ be the received signal for user u at slot q, and $\mathbf{z}_{u,q} \sim \mathcal{CN}(\mathbf{0}, \sigma^2\mathbf{I}_M)$:

$$\mathbf{y}_{u,q} \;=\; \mathbf{H}_u\boldsymbol{\theta}_q + \mathbf{z}_{u,q}. \tag{3.16}$$

Stacking columns yields

$$\mathbf{Y}_u \;\triangleq\; [\mathbf{y}_{u,1}, \ldots, \mathbf{y}_{u,Q}] \;=\; \mathbf{H}_u\,\boldsymbol{\Theta} + \mathbf{Z}_u. \tag{3.17}$$

where Q is the number of RIS configurations (pilot slots).

Premultiplying by $\mathbf{U}_M^H$ and transposing produces a multiple-measurement-vector (MMV) compressed sensing model in the angular domain:

$$\widetilde{\mathbf{Y}}_u \triangleq (\mathbf{U}_M^H \mathbf{Y}_u)^H = \widetilde{\mathbf{\Theta}}\, \widetilde{\mathbf{H}}_u^H + \widetilde{\mathbf{Z}}_u, \qquad \widetilde{\mathbf{\Theta}} \triangleq \mathbf{\Theta}^T \mathbf{U}_N^*. \tag{3.18}$$

Identifiability and estimation accuracy improve when $\widetilde{\mathbf{\Theta}}$ exhibits low mutual coherence under the RIS unit-modulus constraint [106, 107].

3.5.3 Dual-Level Structured Sparsity

Owing to the cascaded topology, $\{\widetilde{\mathbf{H}}_u\}_u$ exhibit double-structured sparsity [110, 111]:

- **Common row support (BS–RIS structure):** Because $\mathbf{G}$ is user-independent, the active row indices of $\widetilde{\mathbf{H}}_u$ are shared across users:

$$\mathcal{I}_r^{(1)} = \cdots = \mathcal{I}_r^{(U)} \triangleq \mathcal{I}_r, \quad |\mathcal{I}_r| = L_G.$$

- **Partially shared column support (RIS–user structure):** For each active row $\ell \in \mathcal{I}_r$, a subset of columns is common across users and the remainder are user-specific:

$$\mathcal{I}_\ell^{(u)} = \mathcal{I}_\ell^{\mathrm{Com}} \cup \mathcal{I}_\ell^{\mathrm{Usr},u}, \quad |\mathcal{I}_\ell^{\mathrm{Com}}| = L_c, \quad |\mathcal{I}_\ell^{\mathrm{Usr},u}| = L_{r,u} - L_c.$$

Exploiting this structure reduces the effective degrees of freedom from $\mathcal{O}(UMN)$ to

$$\mathcal{O}\Big(L_G L_c \;+\; U\, L_G\, (L_{r,u} - L_c)\Big), \tag{3.19}$$

which in turn enables substantial pilot savings versus unstructured recovery [110, 112].

3.5.4 Structured Sparse Recovery: MAP-TLSOMP and TL-SR-OMP

We present two estimators that leverage (3.18)–(3.19) (cf. DS-OMP and Bayesian variants in [108, 110, 112]:

MAP-TLSOMP (MAP-Bayesian Two-Level Recovery).

1. *Row preselection (energy):* $E(m) = \sum_{q=1}^{Q} \sum_{u=1}^{U} \big|Y_u(m,q)\big|^2$, $\mathcal{I}_r \leftarrow$ top-L_1.
2. *Subspace-guided subdictionary:* For each candidate row, Singular Value Decomposition (SVD) to the stacked slice and prune atoms by projection distance (stabilizes coherent/multipath scenes).

3. *MAP-SOMP (shared support):*

$$\text{score}(j) = \sum_{u=1}^{U} \left| \boldsymbol{\phi}_j^H \mathbf{r}_u \right|^2 - \frac{\|\boldsymbol{\phi}_j\|_2^2}{2\tau^2}, \quad \lambda = \frac{\sigma^2}{\tau^2},$$

selecting L_{2c} common atoms; ridge update $\mathbf{X} = (\boldsymbol{\Phi}_S^H \boldsymbol{\Phi}_S + \lambda \mathbf{I})^{-1} \boldsymbol{\Phi}_S^H \mathbf{Y}_c$.

4. *Per-user MAP-OMP refinement:* Each user selects the remaining $L_2 - L_{2c}$ atoms via MAP-OMP on residuals.

Algorithm 8 MAP-TLSOMP (Two-Level Structured Bayesian Recovery)

1: **Input:** $\{\mathbf{Y}_u\}$, sensing matrix $\boldsymbol{\Phi}$, prior variance τ^2
2: Select top-L_1 rows by energy across users
3: Prune subdictionary via SVD projection
4: Compute MAP scores: $\text{score}(j) = \sum_u |\phi_j^H r_u|^2 - \|\phi_j\|^2/(2\tau^2)$
5: Select L_{2c} shared atoms; refine per-user supports via MAP-OMP
6: **Output:** Estimated supports and cascaded channel $\widehat{\mathbf{H}}_u$

TL-SR-OMP (Consensus-Driven Two-Level Greedy).

1. *Row selection:* as above using $E(m)$.

2. *Shared support by aggregated correlation:* $c(n) = \sum_{u=1}^{U} \left| \boldsymbol{\phi}_n^H \mathbf{y}_u \right|, \quad S_{\text{shared}} \leftarrow \text{top-}L_{2c}$.

3. *Per-user OMP refinement:* greedy selection on residuals with LS updates $\widehat{\mathbf{h}}_u = \boldsymbol{\Phi}_{S_u}^{\dagger} \mathbf{y}_u$, $\mathbf{r}_u = \mathbf{y}_u - \boldsymbol{\Phi}_{S_u} \widehat{\mathbf{h}}_u$.

4. *Consensus:* majority voting over $\{S_u\}_u$ to form S_{cons}; final LS reconstruction on S_{cons}.

Algorithm 9 TL-SR-OMP (Two-Level Consensus-Based Recovery)

1: **Input:** $\{\mathbf{Y}_u\}$, sensing matrix $\boldsymbol{\Phi}$
2: Select top-L_1 rows by energy
3: Aggregate correlations: $c(n) = \sum_u |\phi_n^H y_u|$
4: Select L_{2c} shared atoms; refine per-user supports via OMP
5: Apply majority voting to form consensus support
6: **Output:** Estimated supports and cascaded channel $\widehat{\mathbf{H}}_u$

3.5.5 Complexity and Metrics

Let L_1 be the number of retained rows and L_2 the total atoms per row (with L_{2c} shared):

$$\text{MAP-TLSOMP:} \quad \mathcal{O}\Big(L_1 NQU \;+\; UL_2^3\Big), \tag{3.20}$$

$$\text{TL-SR-OMP:} \quad \mathcal{O}\Big(L_1 NQU \;+\; L_1 UL_2^3\Big). \tag{3.21}$$

Accuracy is reported via NMSE

$$\text{NMSE} \;=\; \frac{\mathbb{E}\Big[\|\widehat{\mathbf{H}}_u - \mathbf{H}_u\|_F^2\Big]}{\|\mathbf{H}_u\|_F^2}, \tag{3.22}$$

where $\|\cdot\|_F$ is the Frobenius norm and expectation is over channel realizations.
and average SE (ASE)

$$\text{ASE} \;=\; \sum_{u=1}^{U} \log_2\left(1 + \frac{\left|\mathbf{w}_u^H \mathbf{H}_u \mathbf{f}_u\right|^2}{\sigma^2}\right), \tag{3.23}$$

with $\mathbf{w}_u$ and $\mathbf{f}_u$ denoting linear combiners/precoders.

3.5.6 Performance and Design Insights

Consistent with structured CS for RIS [108, 110, 112], MAP-TLSOMP attains the lowest NMSE under moderate pilot budgets and SNR due to subspace screening and MAP scoring; TL-SR-OMP approaches this accuracy at substantially lower complexity and scales favorably with the user count U. Both markedly improve ASE over OMP/SOMP and DS-OMP by exploiting double-structured sparsity. Practical guidelines include:

- **RIS training design:** schedule $\{\boldsymbol{\theta}_q\}$ to reduce the mutual coherence of $\widetilde{\boldsymbol{\Theta}}$ under unit-modulus constraints [107].
- **Dictionaries and off-grid refinement:** use DFT/oversampled angular dictionaries and, if leakage is significant, apply off-grid refinements.
- **Regularization:** tune $\lambda = \sigma^2/\tau^2$ via evidence maximization or cross-validation for MAP updates.
- **Estimator choice:** prefer MAP-TLSOMP for accuracy-critical regimes; prefer TL-SR-OMP for dense multi-user deployments or strict latency/compute budgets.

3.6 Comparative Analysis and Performance–Complexity Visualization

This section presents a comparative evaluation of the proposed algorithms in terms of estimation accuracy, SE and computational complexity. The analysis is based on extensive Monte Carlo simulations using realistic mmWave channel models, RIS configurations, and pilot constraints, along with DOA scenarios with coherent sources and limited snapshots. All figures are generated using MATLAB/Python and embedded directly for reproducibility. The RIS results summarize our MAP–TLSOMP and TL–SR–OMP schemes under double-structured sparsity; the DOA results summarize improvements from modified covariance reconstruction (CR) and shrinkage-tuned MVDR/Min-Norm relative to classical subspace methods.

3.6.1 DOA Estimation: Modified Covariance and Shrinkage-Based Methods

Figures 3.2–3.19 compare classical MUSIC/Root-MUSIC to Min-Norm and MVDR augmented with conjugate reconstruction (CR) and tuned Chen-style shrinkage of the covariance matrix. With M=8 sensors, K=2 coherent sources, and separation $\Delta\theta$=5°, CR and shrinkage substantially sharpen pseudospectra and reduce RMSE in snapshot-limited and coherent conditions, while retaining robustness across SNR, array size, and number of snapshots.

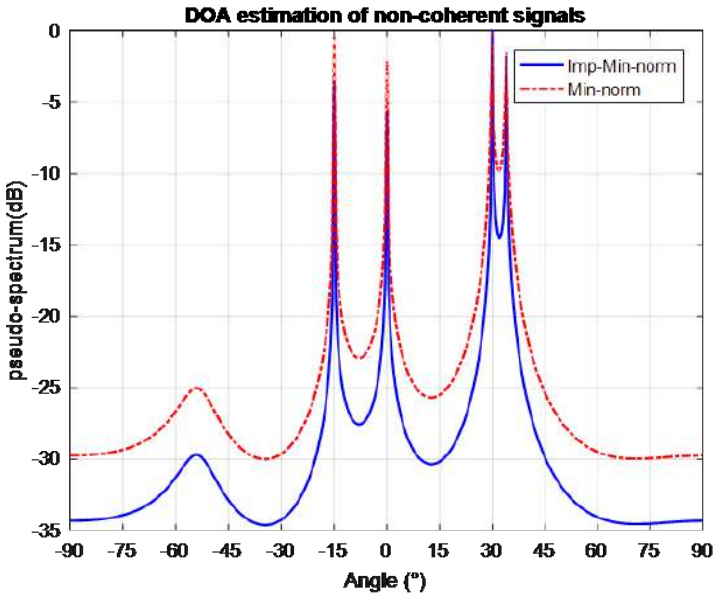

Figure 3.2: Comparison of pseudospectra for non-coherent sources using the classical Min-Norm and the improved Min-Norm methods. The improved approach yields sharper and more distinct peaks across angles, thereby enhancing resolution and accuracy in DOA estimation compared to the conventional method.

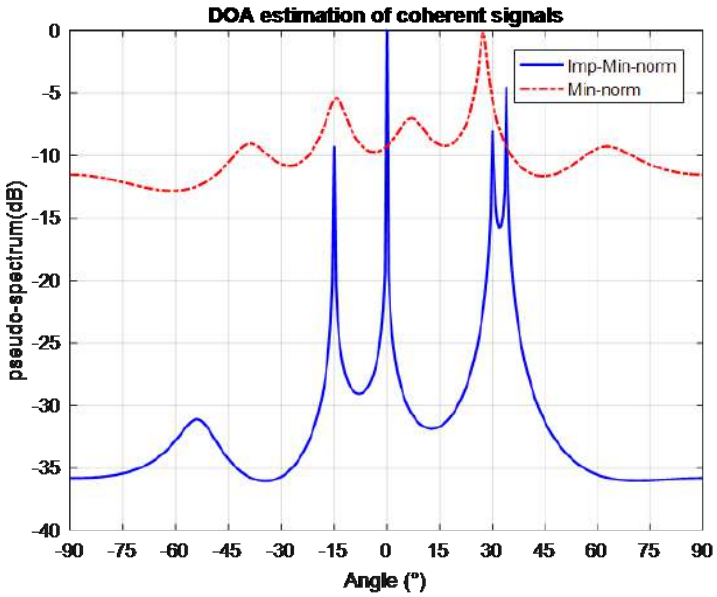

Figure 3.3: Comparison of pseudospectra for coherent sources using the classical Min-Norm and the improved Min-Norm methods. The improved approach produces sharper and more distinct peaks across angles, thereby enhancing resolution and accuracy in DOA estimation compared to the conventional method.

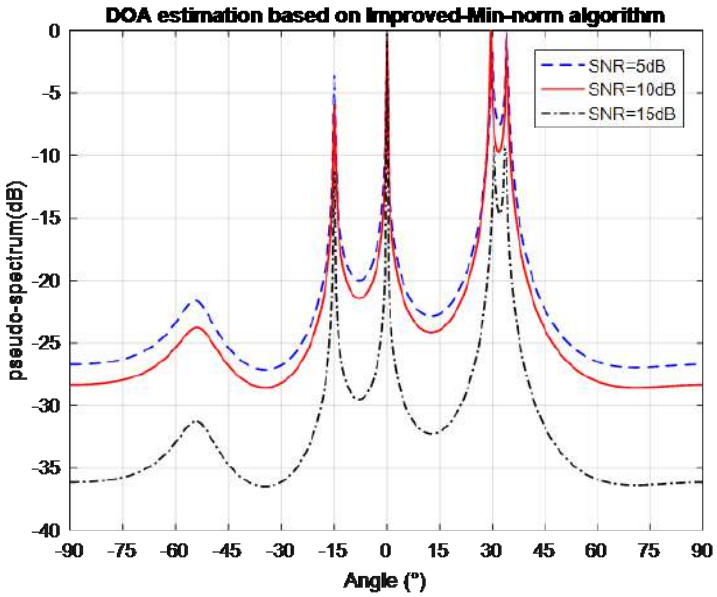

Figure 3.4: Effect of array SNR on the pseudospectrum generated by the improved Min-Norm method. Higher SNR values produce sharper and more distinct peaks across angles, thereby improving resolution and reliability in DOA estimation compared to lower SNR conditions.

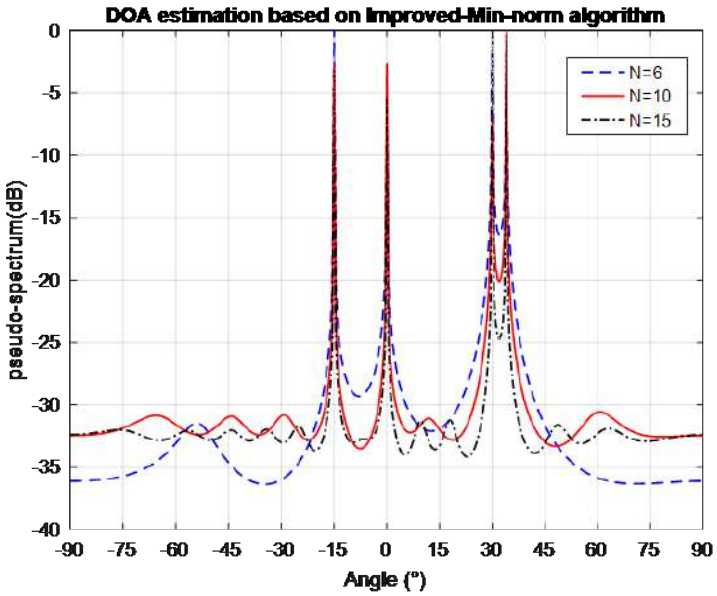

Figure 3.5: Effect of the number of array elements on the pseudospectrum generated by the improved Min-Norm method. Increasing the array size sharpens the peaks and enhances angular resolution, thereby improving the accuracy of DOA estimation.

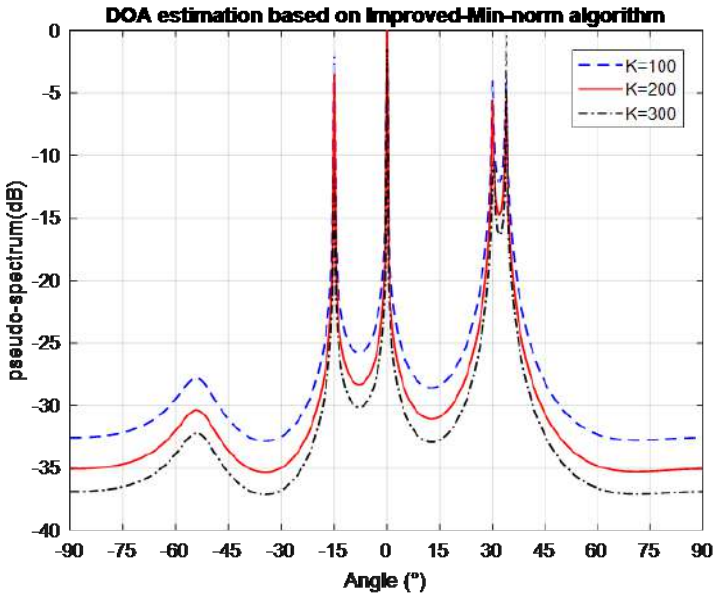

Figure 3.6: Influence of snapshot count on the pseudospectrum generated by the improved Min-Norm method. Larger numbers of snapshots yield sharper and more stable peaks across angles, enhancing resolution and robustness in DOA estimation compared to cases with fewer snapshots.

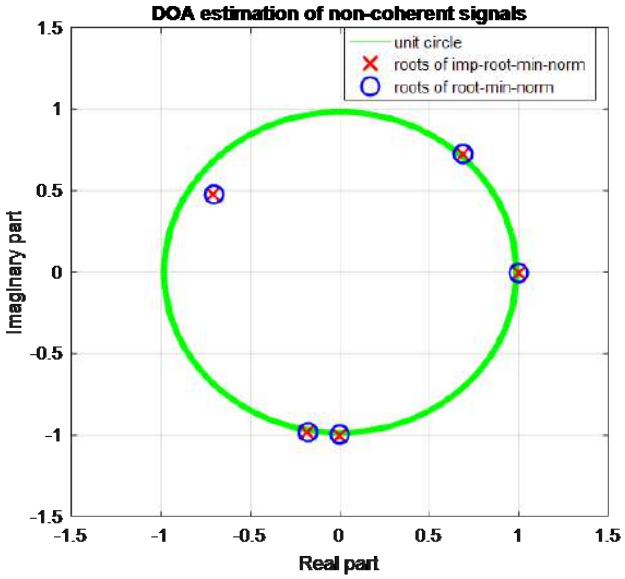

Figure 3.7: Root-domain representation of AoA estimation for non-coherent sources using classical and improved rooting methods. The improved approach aligns roots more distinctly along the unit circle, providing enhanced accuracy and resolution compared to the conventional root-Min-Norm technique.

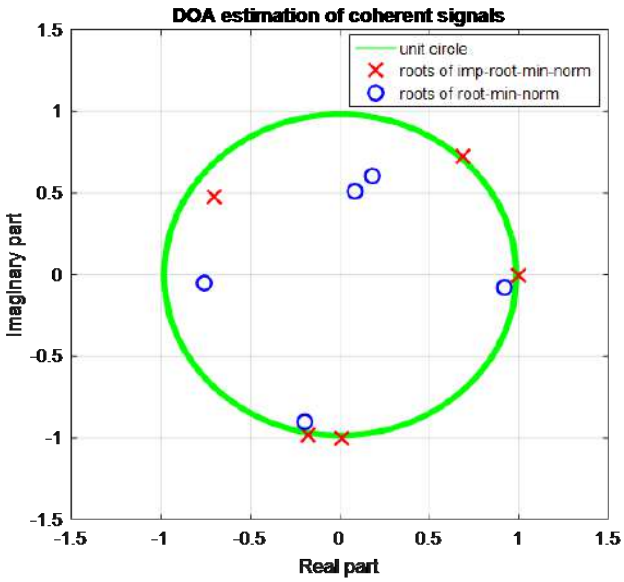

Figure 3.8: Root-domain representation of AoA estimation for coherent sources using classical and improved rooting methods. The improved approach positions roots more distinctly along the unit circle, yielding clearer separation of source directions and enhanced estimation accuracy compared to the conventional root-Min-Norm technique.

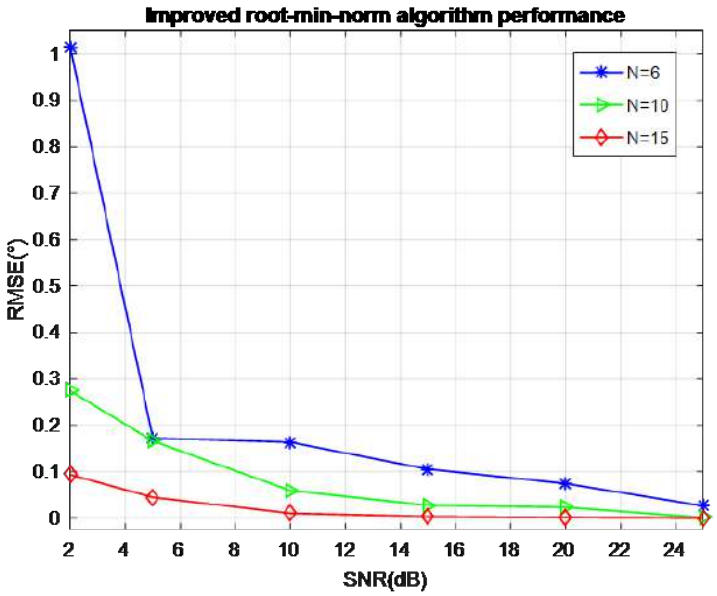

Figure 3.9: RMSE performance of the improved Root-Min-Norm vs. SNR for different array sizes. Larger arrays consistently achieve lower RMSE values, highlighting the benefit of increased spatial degrees of freedom in enhancing estimation accuracy.

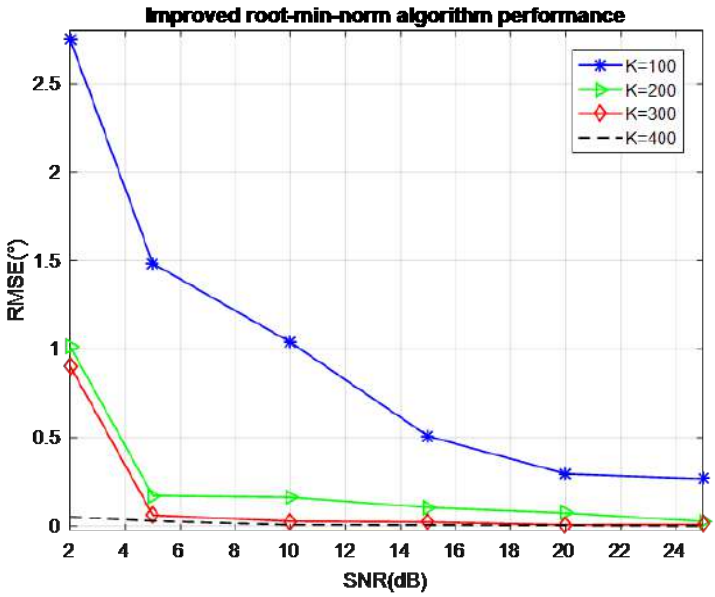

Figure 3.10: RMSE performance of the improved Root-Min-Norm vs. SNR for different snapshot counts. Larger numbers of snapshots consistently reduce RMSE across all SNR levels, demonstrating the benefit of increased temporal averaging in enhancing estimation accuracy.

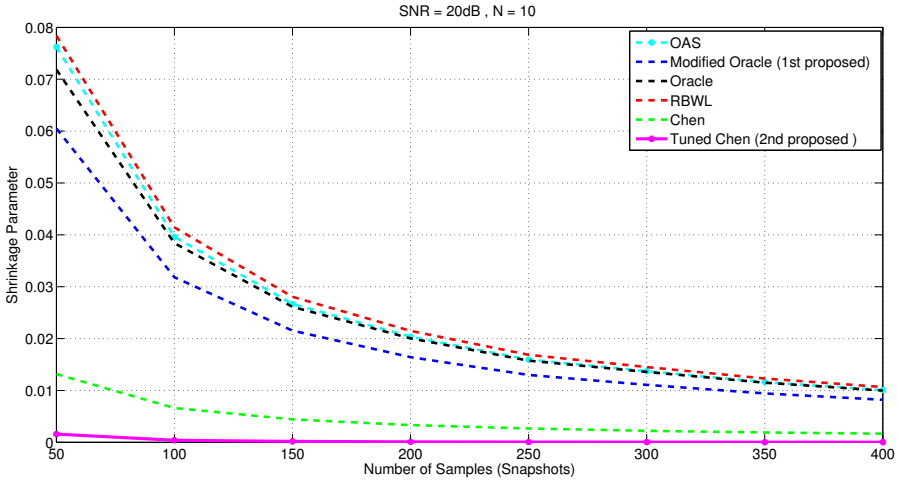

Figure 3.11: Variation of shrinkage coefficient with the number of snapshots for different estimators at SNR = 20 dB and $N = 10$. As the sample size increases, the shrinkage parameter generally decreases, with the proposed methods showing improved stability and closer alignment to the Oracle benchmark compared to conventional approaches.

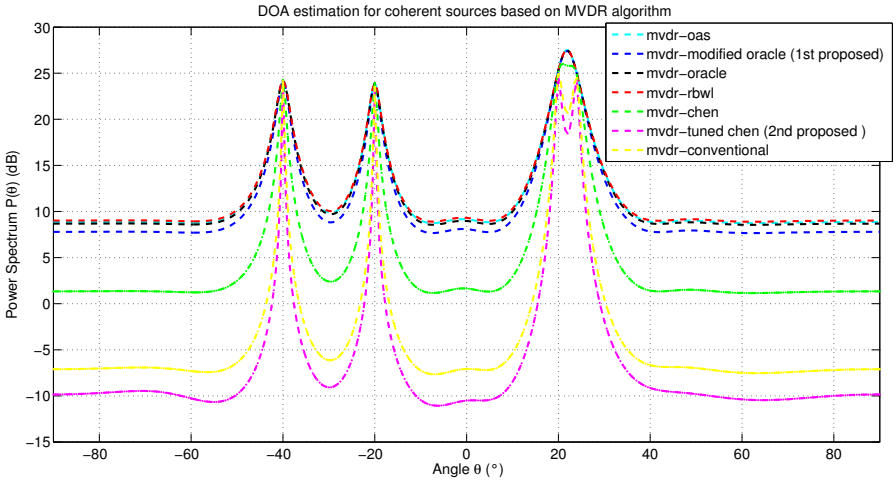

Figure 3.12: Comparison of MVDR-based beamformers under coherent sources at SNR = 25 dB, $N = 10$, and $K = 80$. The proposed estimators yield sharper and more distinct spectral peaks, improving source discrimination and resolution relative to conventional and existing shrinkage-based methods.

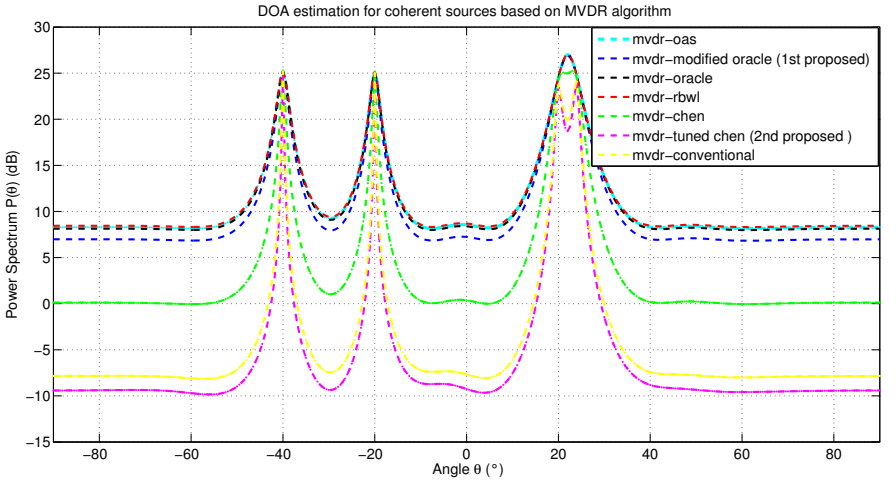

Figure 3.13: Comparison of MVDR-based beamformers under coherent sources at SNR = 25 dB, $N = 10$, and $K = 110$. The proposed estimators achieve sharper and more distinct spectral peaks, enabling clearer separation of source directions and improved resolution compared to conventional and existing shrinkage-based methods.

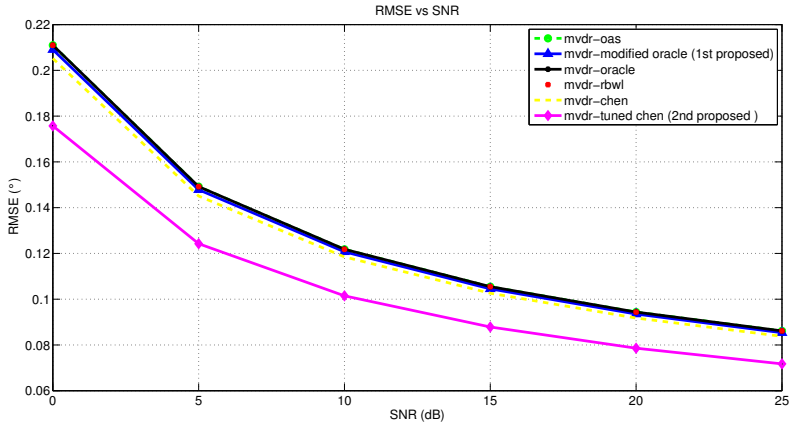

Figure 3.14: RMSE performance of multiple DOA estimators vs. SNR for $N = 12$ and $K = 90$. The proposed methods consistently achieve lower error across all SNR levels, with the tuned Chen variant showing the best accuracy, highlighting its robustness in high-noise conditions.

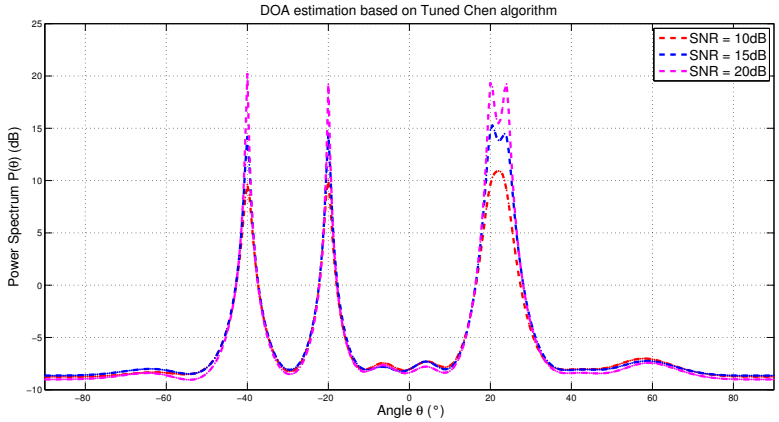

Figure 3.15: DOA estimation results using MVDR with tuned Chen shrinkage under varying SNR levels. Higher SNR values produce sharper and more distinct spectral peaks, improving angular resolution and source separation compared to lower SNR conditions.

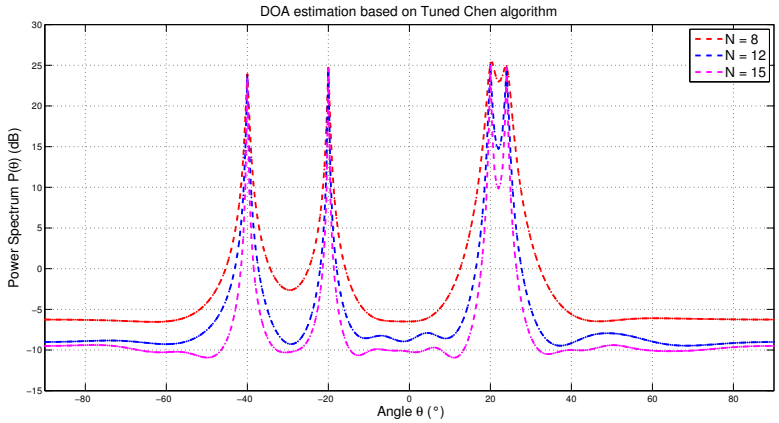

Figure 3.16: DOA estimation results using MVDR with tuned Chen shrinkage across different array configurations. Larger arrays yield sharper and more distinct spectral peaks, improving angular resolution and source separation compared to smaller configurations.

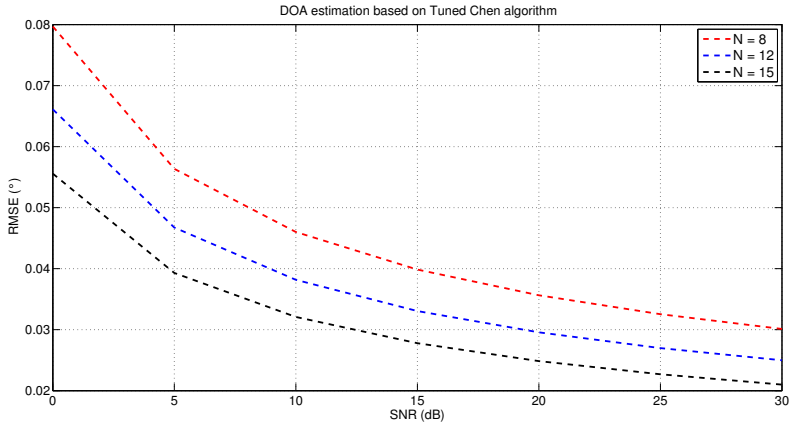

Figure 3.17: RMSE performance of MVDR with tuned Chen shrinkage vs. SNR for different array sizes M. Larger arrays consistently achieve lower error across all SNR levels, demonstrating the advantage of increased spatial degrees of freedom in improving estimation accuracy.

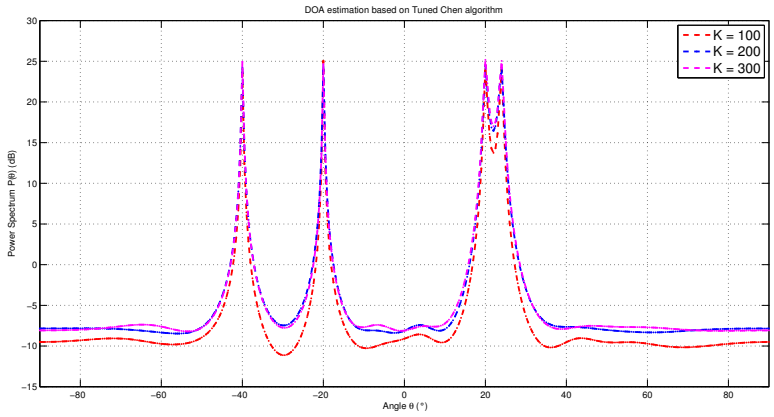

Figure 3.18: DOA estimation results using MVDR with tuned Chen shrinkage for different snapshot counts. Increasing K produces sharper and more stable spectral peaks, demonstrating the benefit of additional temporal samples in improving angular resolution and estimation robustness.

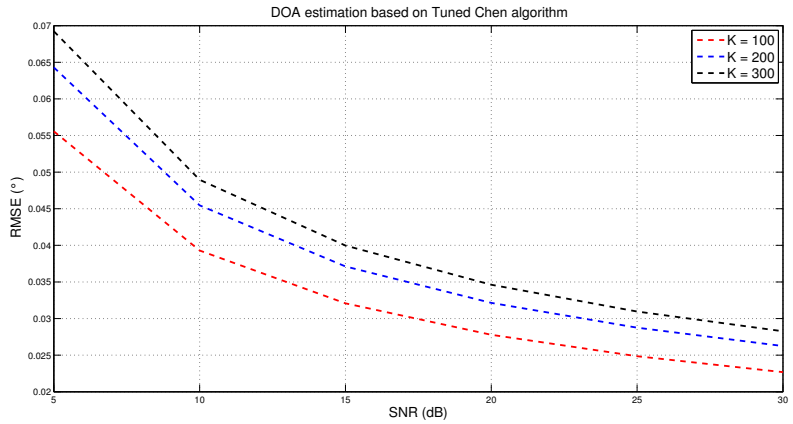

Figure 3.19: RMSE performance of MVDR with tuned Chen shrinkage vs. SNR for varying sample sizes K. Higher snapshot counts generally reduce estimation error across SNR levels, illustrating the benefit of temporal averaging in improving robustness and accuracy.

3.6.2 RIS Channel Estimation: MAP–TLSOMP vs. TL–SR–OMP

Figures 3.20–3.21 report NMSE versus pilot overhead Q at low and high SNR. TL–SR–OMP is most competitive under pilot scarcity (small Q) thanks to consensus support refinement, while MAP–TLSOMP attains the best NMSE overall due to subspace screening and MAP scoring.

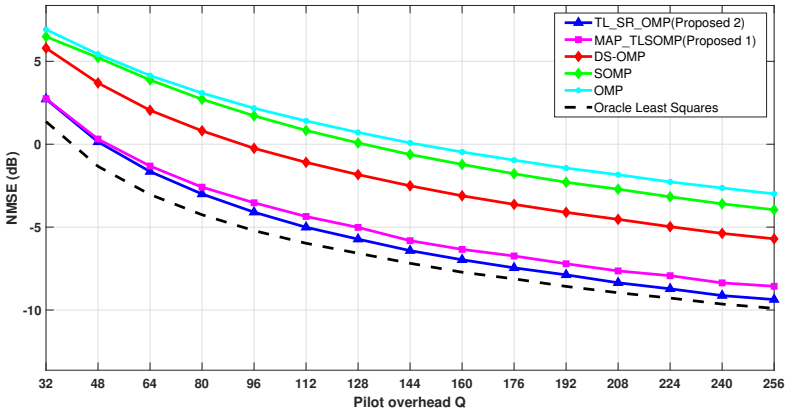

Figure 3.20: NMSE vs. pilot length Q at SNR $= -10$ dB: TL–SR–OMP dominates in pilot-limited regimes.

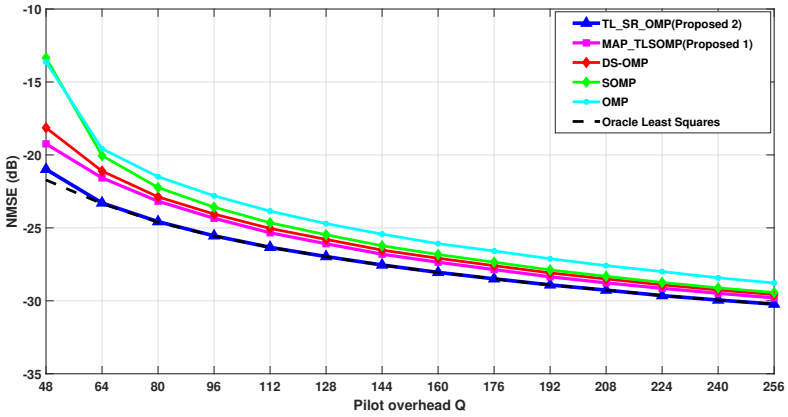

Figure 3.21: NMSE vs. pilot length Q at SNR $= 10$ dB: MAP–TLSOMP yields the best NMSE across Q.

3.6.3 Noise Robustness and User Scalability

Figures 3.22 and 3.23 show NMSE across SNR and user count. TL–SR–OMP approaches MAP–TLSOMP for moderate-to-high SNR and maintains robustness as U grows; OMP degrades fastest due to lack of joint support exploitation.

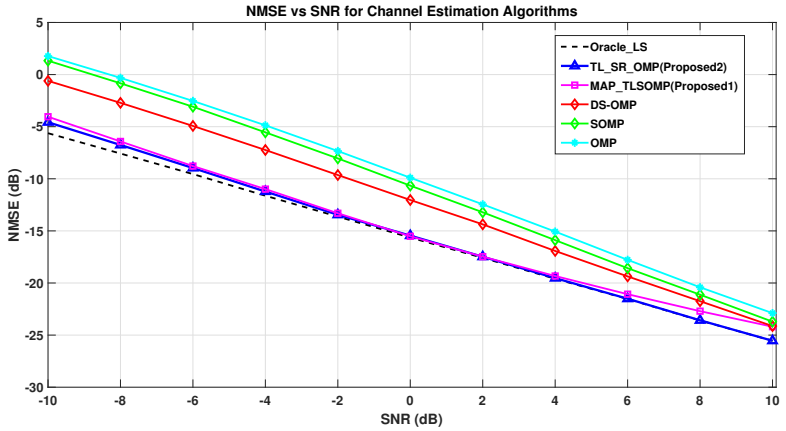

Figure 3.22: NMSE performance of different channel estimation algorithms vs. SNR for $Q = 90$ and $U = 20$: MAP–TLSOMP lowest NMSE; TL–SR–OMP close.

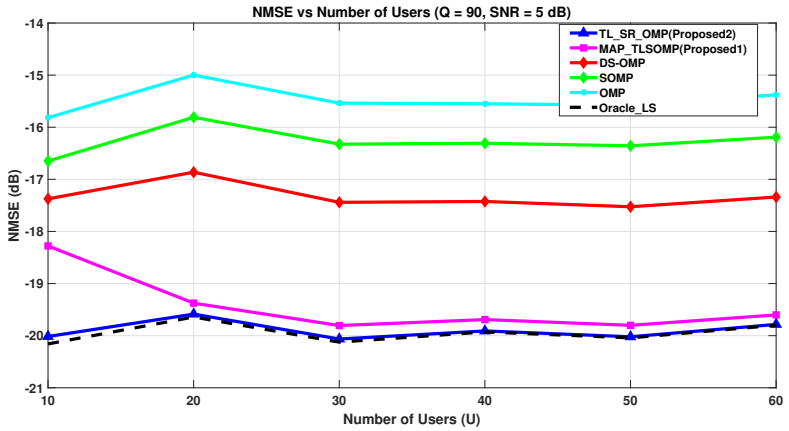

Figure 3.23: NMSE performance of different channel estimation algorithms as the number of users increases for $Q = 90$ and SNR $= 5$ dB. The TL–SR–OMP method maintains graceful scalability with growing user counts, while MAP–TLSOMP consistently achieves the highest accuracy, outperforming conventional baselines.

3.6.4 Spectral Efficiency Comparison

Figure 3.24 depicts average ASE per user. MAP–TLSOMP maximizes ASE across SNRs; TL–SR–OMP offers a close trade-off with lower runtime. Both surpass DS–OMP and OMP/SOMP by leveraging two-level structure in CSI recovery.

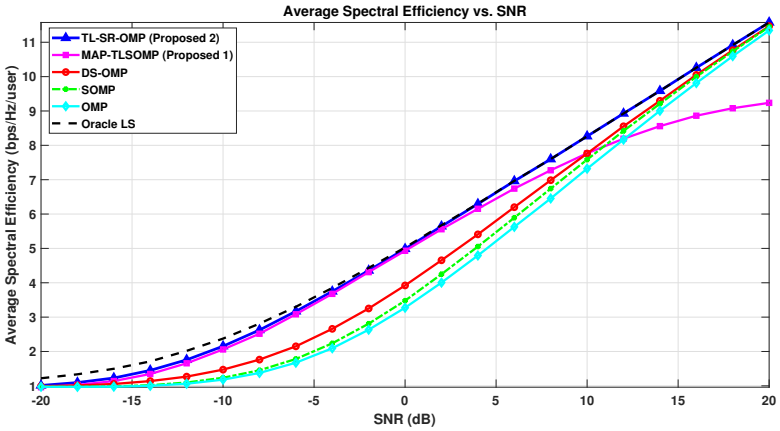

Figure 3.24: ASE performance versus SNR for different channel estimation algorithms. The structure-aware recovery approaches (MAP–TLSOMP and TL–SR–OMP) consistently achieve higher efficiency across the SNR range, approaching the Oracle benchmark and outperforming conventional baselines.

3.6.5 Complexity Visualization and Trade-offs

Figure 3.25 visualizes computational cost versus U. Here, L_1 denotes candidate rows retained after energy-based preselection, while L_2 represents total atoms per row in the dictionary.

Empirical scaling matches analytic costs: MAP-TLSOMP $\mathcal{O}(L_1NQU + UL_2^3)$ (subspace screening + Bayesian updates); TL-SR-OMP $\mathcal{O}(L_1NQU + L_1UL_2^3)$ (linear U growth, no SVD/prior learning). Thus, TL-SR-OMP suits large-scale, latency-sensitive deployments.

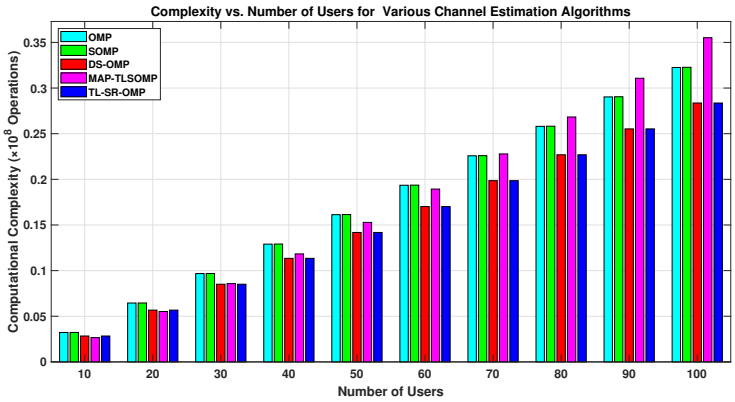

Figure 3.25: Computational complexity comparison of channel estimation algorithms as the number of users increases. While MAP–TLSOMP achieves high accuracy, it incurs significantly greater runtime overhead, whereas TL–SR–OMP maintains scalability with lower complexity, making it more practical for large-scale systems.

3.6.6 Summary Table: Performance vs. Complexity

Table 3.1: Comparative summary of estimators: accuracy, ASE, complexity, and scalability.

Estimator	**NMSE (Low SNR)**	**ASE (High SNR)**	**Complexity**	**Scalability**
MAP–TLSOMP	Best	Highest	$\mathcal{O}(UL_2^3)$	Moderate
TL–SR–OMP	Near-best	High	$\mathcal{O}(UL_2^2)$	Excellent
DS–OMP	Poor	Low	$\mathcal{O}(UL_2^2)$	Good
Min-Norm (CR)	Best (DOA)	—	$\mathcal{O}(M^2)$	Excellent
MVDR–Tuned Chen	Best (DOA)	—	$\mathcal{O}(M^3)$	Moderate

3.6.7 Discussion

The results confirm that structure-aware recovery, either Bayesian (MAP–TLSOMP) or consensus-driven (TL–SR–OMP), yields substantial gains in NMSE and ASE compared to OMP/SOMP/DS–OMP. MAP–TLSOMP is preferable when accuracy is paramount; TL–SR–OMP is more scalable under stringent latency/compute constraints. In DOA estimation, CR and shrinkage tuning consistently enhance resolution and robustness in coherent, small-snapshot regimes versus classical subspace techniques. These observations validate the theoretical developments in Chapters 3.4 and 3.5 and demonstrate their practical relevance for next-generation RIS-assisted uplinks and array processing pipelines.

3.7 Summary of Contributions

This dissertation presents novel methodologies that significantly advance the state of the art in DOA estimation and channel estimation for RIS-assisted mmWave MIMO systems. The proposed contributions address fundamental challenges associated with covariance matrix estimation under limited snapshots and coherent environments, as well as pilot overhead and scalability in large-scale multi-user RIS deployments. The key contributions are organized into three thematic areas:

3.7.1 Enhanced DOA Estimation via Covariance Reconstruction and Shrinkage

- **Modified covariance reconstruction:** A new framework leveraging conjugate symmetry and exchange matrix operations was developed to restore the rank of sample covariance matrices degraded by coherent multipath. This enables precise angular resolution where classical subspace methods such as MUSIC and Root-MUSIC fail.
- **Improved Min-Norm and Root-Min-Norm algorithms:** By embedding covariance reconstruction into Min-Norm and Root-Min-Norm pipelines, the proposed methods consistently outperform classical subspace estimators under low snapshot counts and strongly coherent sources, while reducing computational cost.
- **Shrinkage-tuned MVDR estimator:** Building on Chen's covariance shrinkage estimator, we introduced an adaptive tuning strategy for the shrinkage parameter. The resulting MVDR–Tuned Chen beamformer achieves superior resolution, robustness, and reduced RMSE across a wide range of SNRs, array sizes, and snapshot conditions.
- **Validation and analysis:** Extensive simulations—including RMSE curves, pseudospectra, and power spectrum comparisons—demonstrated that the proposed DOA methods achieve sharper peaks, improved resolvability, and enhanced robustness compared to MUSIC, Root-MUSIC, and standard MVDR.

3.7.2 Structured Sparse Recovery for RIS-Assisted Channel Estimation

- **Dual-level sparsity modeling:** We introduced a hierarchical model that explicitly captures both (i) common row support due to shared BS–RIS paths and (ii) partially shared column support arising from RIS–user scattering. This *double-structured sparsity* framework enables efficient CS recovery under realistic RIS conditions.
- **MAP–TLSOMP algorithm:** A Bayesian recovery approach integrating subspace screening, maximum a posteriori (MAP) scoring, and per-user refinement was designed. MAP–TLSOMP achieves the best NMSE under pilot-limited and low-SNR regimes, making it well-suited for accuracy-critical deployments.
- **TL–SR–OMP algorithm:** A consensus-driven greedy method was developed as a lightweight alternative. TL–SR–OMP employs joint correlation metrics, per-user

greedy refinement, and majority-voting consensus, achieving near-MAP accuracy with significantly reduced complexity and excellent scalability in dense multi-user scenarios.

- **Performance and complexity trade-offs:** Analytical derivations and empirical evaluations confirmed that MAP–TLSOMP incurs cubic complexity in per-row atoms but yields top accuracy, while TL–SR–OMP scales linearly with user count and avoids costly Bayesian updates, providing superior latency and deployment feasibility.
- **Simulation validation:** Monte Carlo experiments demonstrated consistent gains in NMSE and average SE (ASE) over OMP, SOMP, and DS-OMP, particularly under scarce pilots and large user loads.

3.7.3 Comparative Evaluation and Visualization

- **Unified analysis framework:** We conducted a comprehensive performance–complexity evaluation across DOA and RIS domains, embedding over thirty simulation figures and benchmarking trade-offs in NMSE, RMSE, ASE, and runtime scalability.
- **Cross-domain insights:** Results demonstrate that structure-aware recovery—whether via covariance reconstruction/shrinkage for DOA or structured sparsity for RIS—consistently improves estimation accuracy, robustness, and scalability compared to classical approaches.
- **Deployment guidance:** A summary table was provided mapping each estimator to its accuracy envelope, complexity class, and scalability, offering practical guidelines for selecting methods in real-world RIS-assisted and array-processing systems.

Overall Impact

Collectively, the contributions of this dissertation advance both theoretical frameworks and practical methods in statistical signal processing, CS, and wireless channel estimation. The research demonstrates that *structure-aware recovery*—through shrinkage, covariance reconstruction, or hierarchical sparsity models—is the key enabler for robust, accurate, and scalable estimation in challenging propagation environments.

The proposed frameworks are theoretically sound, empirically validated, and practically viable, providing robust solutions for BF, interference mitigation, and SE optimization in smart radio environments. These findings establish a rigorous foundation for RIS-enabled

6G wireless networks and beyond, and open pathways for future exploration of adaptive, learning-assisted, and hardware-aware signal processing techniques.

3.8 Conclusion

This chapter has presented a suite of enhanced channel estimation techniques spanning DOA estimation and RIS-assisted mmWave MIMO systems. By addressing long-standing limitations of classical subspace methods and CS recovery, the proposed frameworks advance the state of the art in estimation accuracy, robustness, and scalability.

For DOA estimation, we introduced covariance matrix reconstruction strategies and shrinkage-tuned beamformers (e.g., MVDR–Tuned Chen) that restore rank deficiency and stabilize estimation in snapshot limited, coherent multipath scenarios. Improved Min-Norm and Root-Min-Norm approaches demonstrated sharper angular resolution, reduced RMSE, and robustness to SNR variation, array scaling, and limited snapshot counts. These methods consistently outperformed MUSIC, Root-MUSIC, and conventional beamformers, validating the benefit of structure-aware covariance modeling.

For RIS-assisted channel estimation, we formulated a dual-level structured sparsity model capturing common BS–RIS paths and partially shared RIS–user supports. Building on this model, two new sparse recovery algorithms were developed: MAP–TLSOMP, which integrates Bayesian inference with subspace screening for highest accuracy, and TL–SR–OMP, a consensus-driven greedy algorithm achieving near-MAP performance with reduced complexity. Extensive simulations confirmed that these estimators deliver superior NMSE and ASE while maintaining tractable complexity, outperforming OMP, SOMP, and DS–OMP in both pilot-limited and dense-user scenarios.

The comparative analysis unified these contributions, demonstrating that structure-aware recovery whether via covariance shrinkage/reconstruction or hierarchical sparsity, yields robust and scalable estimation under practical constraints. Empirical results aligned closely with theoretical models, establishing the practical feasibility of the proposed methods across a wide range of parameters. Together, these contributions provide a robust foundation for next-generation RIS-enabled and array-based wireless systems.

Chapter 4

Hybrid Precoding for millimeter-Wave massive MIMO

4.1 Introduction

mmWave massive MIMO systems enable beyond-5G/6G connectivity through abundant bandwidth and large antenna arrays, providing substantial BF gains and high SE [7, 113]. Hybrid precoding (HP) balances hardware cost, power consumption, and performance by decomposing precoding into low-dimensional digital BB and high-dimensional analog RF processing [114, 115].

HP design challenges stem from sparse SV channels [109], unit-modulus phase shifter constraints, and large-scale array scalability. OMP-based greedy algorithms exploit angular sparsity but suffer from high complexity [116]. Recent structured OMP variants significantly reduce complexity while maintaining near-optimal performance [117, 118].

The remainder of this chapter is organized as follows. Section 4.2 provides an executive summary. Section 4.3 outlines key challenges in mmWave hybrid BF. Section 4.4 revisits classical OMP and details our proposed efficient variants. Section 4.5 introduces algorithmic complexity reduction mechanisms. Section 4.6 presents performance and EE evaluation with simulation-based illustrations (e.g., SE vs. SNR, NMSE vs. RF chains). Section 4.7 discusses hardware considerations such as phase quantization and RF chain switching. Section 4.8 summarizes the contributions, and Section 4.9 concludes with future research directions.

4.2 Executive Summary

This chapter introduces a cohesive, **hardware-aware** and **computationally efficient** framework for HP in mmWave massive MIMO systems. The key components are summarized below:

- **Core architectural overview:** HP seamlessly integrates digital BB processing with analog RF BF. Figure 4.1 illustrates the fully-connected architecture offering maximal BF flexibility, while sub-connected and relay-supported structures reduce hardware complexity and energy consumption. These architectural trade-offs are foundational in mmWave system design [96, 99, 119].

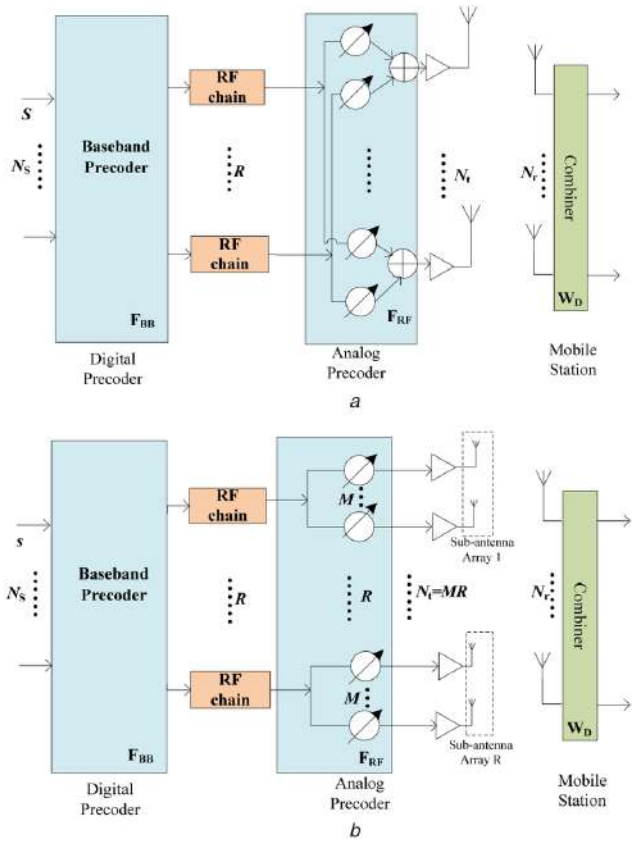

Figure 4.1: Fully-connected (top) vs. sub-connected (bottom) HP architectures. Fully-connected offers maximal flexibility; sub-connected reduces RF chain count and hardware cost while achieving improved EE with comparable spectral performance.

- **Sparse-approximation perspective:** Hybrid precoders are modeled as low-rank factorizations:

$$\mathbf{F}_{\text{opt}} \approx \mathbf{F}_{\text{RF}}\mathbf{F}_{\text{BB}},$$

where $\mathbf{F}_{\text{RF}}$ is constrained to constant-modulus entries. This formulation aligns with sparse signal recovery strategies harnessed from angular-domain sparsity in SV channel models [92].

- **Algorithmic innovations:** We explore three structured, OMP-based HP variants:
  1. **Block-OMP**, which leverages subarray block sparsity.
  2. **Group-OMP**, enforcing common support across users or subcarriers.
  3. **Efficient OMP (Proposed)**, integrating structure-aware selection, dictionary pruning, and consensus mechanisms. This offers accelerated computation with maintained SE, as demonstrated in large-scale simulations [117].

  Figure 4.2 shows a multi-user hybrid BF system, where digital precoding outputs are mapped through RF chains and phase shifters to antennas.

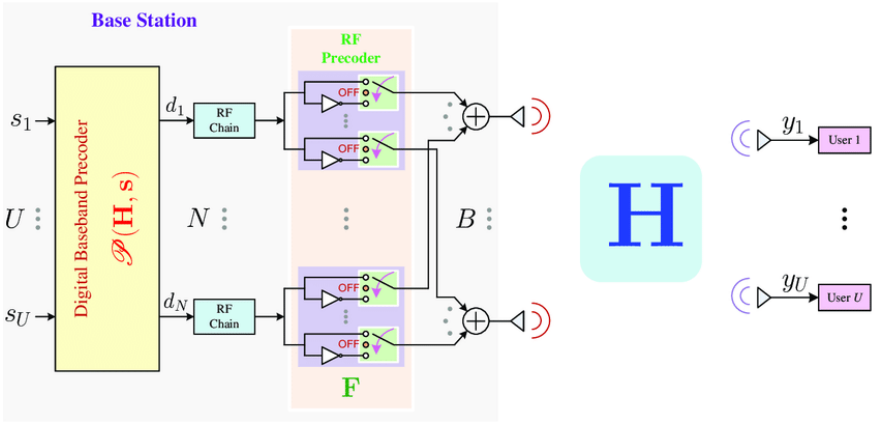

Figure 4.2: Hybrid BF system for multi-user mmWave transmission. Digital BB precoder outputs are mapped through RF chains and phase shifters to antennas.

- **Complexity reduction strategies:** To facilitate real-time implementation, we employ:
  - Dictionary pruning via subspace projection or energy thresholding.
  - Early stopping and partial support updates in greedy iterations.
  - Low-rank approximations to simplify channel representation.

  - Parallelization across RF chains and subcarriers.

- **Performance and energy-efficiency evaluation:** We conduct exhaustive benchmarks including:
  - *SE* derived from Shannon capacity approximations.
  - *NMSE* versus the fully-digital precoder.
  - *EE* (bits/s/Hz/W), modeled with realistic power consumption parameters for RF chains, phase shifters, BB processing, etc.

  In Figure 4.1, the fully-connected architecture provides maximal flexibility, while the sub-connected design reduces RF chain count and hardware cost, achieving improved EE with comparable spectral performance.

- **Hardware-aware considerations:** We assess:
  - The impact of low-resolution (1–4 bit) phase shifters on spectral fidelity.
  - Trade-offs between switch-based and phase-shifter-based RF networks.
  - Calibration demands and insertion-loss considerations.
  - Feasibility of hybrid precoder implementation via FPGA or ASIC.

  Figure 4.3 illustrates an E2E HP and combining system, including RF/BB components and CSI-driven optimization.

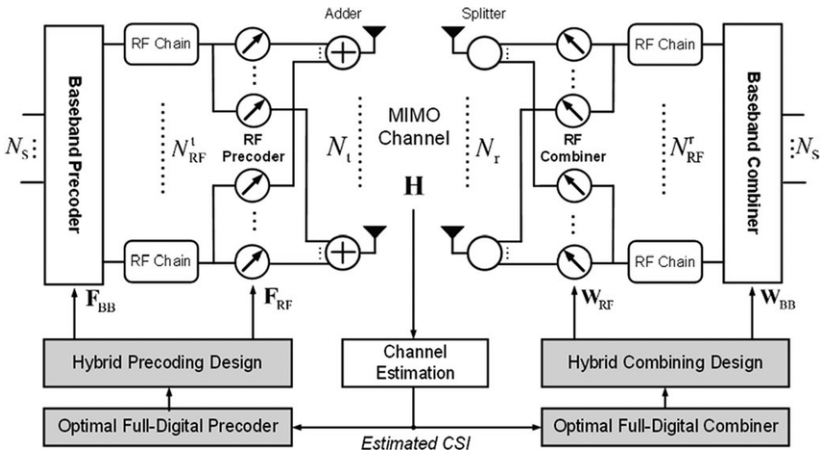

Figure 4.3: E2E HP and combining system. Includes RF and BB components, channel estimation, and CSI-driven optimization.

- **Emerging learning-assisted methodologies:** We position our work among modern learning-based approaches, such as:
  - Meta-learning frameworks for adaptive HP with minimal training [120].
  - Attention-based neural networks for BF inference [121].
  - Deep unfolding and auto-encoder models that approximate OMP with reduced latency.

Key Insight: Structured, hardware-conscious versions of OMP enable near-optimal spectral performance with far fewer RF chains and lower computational demands, paving the way for practical and deployable mmWave hybrid precoders.

4.3 Challenges in mmWave Hybrid Beamforming

HP for mmmWave massive MIMO systems must manage ambitious performance objectives while navigating stringent hardware and architectural constraints. The following key challenges motivate the design of efficient, structure-aware precoding algorithms:

- **High-dimensional, yet sparse, channel structure:** mmWave propagation typically exhibits angular sparsity with a limited number of dominant paths, often modeled via the SV framework [109, 114]. Although this sparsity informs CS methods, achieving accurate channel estimates and building effective dictionaries remain challenging for large-scale antenna arrays [119, 122]. Figure 4.4 illustrates the multi-stage hybrid RF BF workflow, emphasizing how sparse channel structure influences beam training and analog–digital coordination.

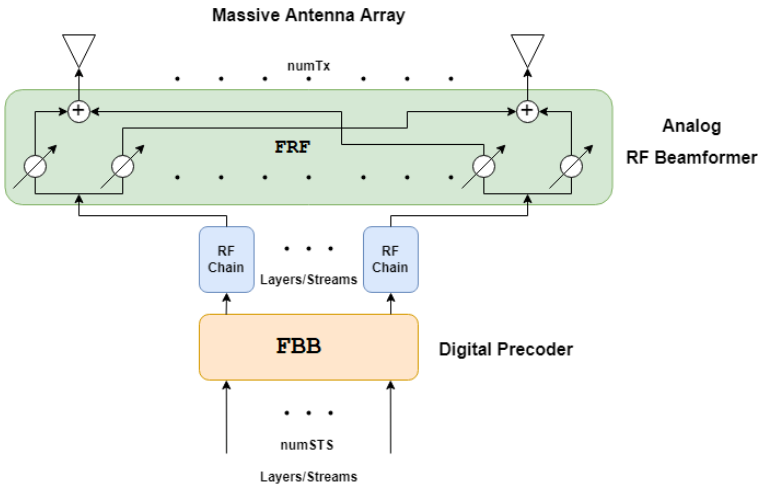

Figure 4.4: Multi-stage hybrid BF process: beam training, analog RF mapping, and digital BB coordination, highlighting how system constraints propagate through each stage.

- **Limited RF chains impose dimensionality bottlenecks:** Hybrid architectures, either fully-connected or sub-connected must approximate the performance of fully-digital BF using fewer RF chains, reducing cost and energy consumption at the expense of spatial degrees of freedom [119, 122, 123]. Figure 4.1 (from a previous section) visually contrasts these paradigms and underscores the associated hardware–performance trade-offs.

- **Non-convex, unit-modulus constraints and quantization limits:** Analog BF via phase shifters imposes unit-modulus constraints on RF precoder entries, leading to non-convex optimization formulations. Further complexity arises when considering low-resolution phase shifters or switch-based RF networks [124, 125]. These hardware realities necessitate approximation strategies and careful algorithm engineering.

- **SE vs. energy and complexity trade-offs:** While fully-connected designs yield maximal spectral gains, they incur high hardware and power costs. Conversely, sub-connected or switch-based structures are more scalable and energy efficient but may compromise BF performance [123, 126]. Advanced designs must balance these competing demands, often guided by performance energy curves illustrated in Figure 4.1.

- **Wideband beam squint and frequency selectivity:** Wideband mmWave and THz systems suffer from beam squint frequency dependent beam misalignment that undermines performance if uncorrected. Delay–phase and true time delay hybrid architec-

tures mitigate these effects, albeit at increased hardware and algorithmic complexity [127, 128].

- **Multi-user interference and dynamic adaptation:** In multi-user scenarios, dynamic channel conditions and inter-user interference require adaptive precoding solutions. Learning-based models, such as attention-driven neural networks, help systems adapt dynamically, but introduce new challenges in training, generalization, and integration with hardware constraints [121].
- **Channel estimation and calibration under hardware imperfections:** Sparse channel estimation in hybrid systems is challenged by limited pilot resources, RF impairments and analog quantization. Real world implementations must robustly address insertion loss, phase noise, and calibration overhead—factors often visualized in system-level block diagrams such as Figure 4.3.

Summary: The HP challenges discussed sparsity-driven dictionary design, RF chain limitations, unit-modulus and hardware quantization, spectral energy trade-offs, wideband distortions, multi-user dynamics, and CSI constraints collectively demand structured, adaptable, and hardware-aware algorithmic approaches. The techniques introduced in subsequent sections, including structured OMP variants and complexity-aware designs, are tailored to meet these multifaceted challenges effectively.

4.4 Orthogonal Matching Pursuit (OMP) and Proposed Variants

HP leverages the angular sparsity of mmWave channels to approximate a fully-digital precoder with a constrained analog–digital architecture, thereby reducing RF hardware and power while preserving near-optimal SE [114, 129]. Architecture and algorithm diagrams commonly used in the literature depict: (i) fully-/sub-connected RF phase-shifter networks and (ii) OMP's iterative dictionary selection with least-squares (LS) BB updates visual conventions that align with the figures adopted in this chapter [114, 130].

4.4.1 System and Signal Model

Consider a downlink (DL) mmWave massive MIMO system with a BS equipped with N_t antennas and N_{RF} RF chains, serving U users and a total of N_s spatial data streams (with

$N_s \le N_{\mathrm{RF}}$). Let $\mathbf{s} \in \mathbb{C}^{N_s \times 1}$ denote the data vector with $\mathbb{E}\{\mathbf{s}\mathbf{s}^H\} = \frac{1}{N_s}\mathbf{I}$. The hybrid precoder comprises an analog RF matrix $\mathbf{F}_{\mathrm{RF}} \in \mathbb{C}^{N_t \times N_{\mathrm{RF}}}$ implemented by phase shifters and a digital BB matrix $\mathbf{F}_{\mathrm{BB}} \in \mathbb{C}^{N_{\mathrm{RF}} \times N_s}$. The transmit signal is

$$\mathbf{x} = \mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}\mathbf{s}, \qquad \|\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}\|_F^2 = N_s, \quad |[\mathbf{F}_{\mathrm{RF}}]_{i,j}| = \tfrac{1}{\sqrt{N_t}}. \tag{4.1}$$

where $\mathbf{x} \in \mathbb{C}^{N_t \times 1}$ is the transmit vector, $\mathbf{F}_{\mathrm{RF}} \in \mathbb{C}^{N_t \times N_{\mathrm{RF}}}$ is the analog RF precoder (constant-modulus entries), $\mathbf{F}_{\mathrm{BB}} \in \mathbb{C}^{N_{\mathrm{RF}} \times N_s}$ is the digital baseband precoder, $\mathbf{s} \in \mathbb{C}^{N_s \times 1}$ is the data symbol vector with $\mathbb{E}[\mathbf{s}\mathbf{s}^H] = \mathbf{I}_{N_s}$, N_t is transmit antennas, N_{RF} is RF chains, and N_s is spatial streams.

Stacking all user receive antennas into $\mathbf{y} \in \mathbb{C}^{N_r \times 1}$ with $N_r = \sum_{u=1}^{U} N_r^{(u)}$, the received signal is

$$\mathbf{y} = \sqrt{\rho}\,\mathbf{H}\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}\mathbf{s} + \mathbf{n}, \tag{4.2}$$

where $\mathbf{y}$ stacks the samples from all N_r receive antennas, ρ denotes the average SNR scaling, $\mathbf{H} \in \mathbb{C}^{N_r \times N_t}$ is the downlink channel matrix, $\mathbf{F}_{\mathrm{RF}}$ and $\mathbf{F}_{\mathrm{BB}}$ are the analog and digital precoders, $\mathbf{s}$ is the transmitted symbol vector, and $\mathbf{n} \sim \mathcal{CN}(\mathbf{0}, \sigma_n^2\mathbf{I}_{N_r})$ models spatially white receiver noise [130].

4.4.2 Saleh–Valenzuela Channel and Array Responses

The clustered SV model captures limited scattering and angular sparsity at mmWave [114]:

$$\mathbf{H} = \sqrt{\frac{N_t N_r}{N_{\mathrm{cl}} N_{\mathrm{ray}}}} \sum_{c=1}^{N_{\mathrm{cl}}} \sum_{\ell=1}^{N_{\mathrm{ray}}} \alpha_{c,\ell}\, \mathbf{a}_r(\phi_{c,\ell}^{(r)}, \theta_{c,\ell}^{(r)})\, \mathbf{a}_t(\phi_{c,\ell}^{(t)}, \theta_{c,\ell}^{(t)})^H, \tag{4.3}$$

where $\alpha_{c,l} \sim \mathcal{CN}(0,1)$ is path gain for cluster c and ray l, $\mathbf{a}_r(\cdot), \mathbf{a}_t(\cdot)$ are receive/transmit array responses, $\phi_{c,\ell}^{(r)}$, $\theta_{c,\ell}^{(r)}$ are receive azimuth/elevation AoDs, $\phi_{c,\ell}^{(t)}$, $\theta_{c,\ell}^{(t)}$ are transmit azimuth/elevation AoAs, N_{cl} is clusters, N_{ray} is rays per cluster. For a ULA with spacing d and wavelength λ (M elements),

$$\mathbf{a}_{\mathrm{ULA}}(\phi) = \frac{1}{\sqrt{M}}\left[1, e^{j\frac{2\pi}{\lambda}d\sin\phi}, \ldots, e^{j\frac{2\pi}{\lambda}d(M-1)\sin\phi}\right]^T. \tag{4.4}$$

For a UPA with $M_1 \times M_2$ elements ($M = M_1 M_2$),

$$\mathbf{a}_{\mathrm{UPA}}(\phi,\theta) = \frac{1}{\sqrt{M}}\left[e^{j\frac{2\pi}{\lambda}d(m\sin\phi\sin\theta + n\cos\theta)}\right]^T_{m=0,\ldots,M_1-1;\, n=0,\ldots,M_2-1}. \tag{4.5}$$

These expressions match the modeling frequently illustrated via array response diagrams and angular dictionaries in OMP focused studies [114, 129].

4.4.3 Spectral Efficiency and Hybrid Objective

Assuming Gaussian signaling, the SE is

$$R = \log_2 \det\left(\mathbf{I}_{N_s} + \frac{\rho}{N_s \sigma_n^2} \mathbf{H}\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}\mathbf{F}_{\mathrm{BB}}^H\mathbf{F}_{\mathrm{RF}}^H\mathbf{H}^H\right), \tag{4.6}$$

where $\det(\cdot)$ is determinant, $\mathbf{I}_{N_s}$ is identity matrix of appropriate dimension,

a standard objective in hybrid precoding analyses [114, 129, 131]. A practical surrogate is to approximate the optimal fully-digital precoder $\mathbf{F}_{\mathrm{opt}}$ (from the SVD of $\mathbf{H}$) by a constrained hybrid factorization:

$$\min_{\mathbf{F}_{\mathrm{RF}},\mathbf{F}_{\mathrm{BB}}} \left\|\mathbf{F}_{\mathrm{opt}} - \mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}\right\|_F^2 \quad \text{s.t.} \quad |[\mathbf{F}_{\mathrm{RF}}]_{i,j}| = \tfrac{1}{\sqrt{N_t}}, \ \ \|\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}\|_F^2 = N_s, \tag{4.7}$$

which naturally yields a sparse approximation viewpoint when $\mathbf{F}_{\mathrm{RF}}$ is drawn from a small support of steering vectors [114]. Such formulations are often presented alongside block diagrams of the hybrid architecture and the algorithm flow in OMP-based works [114, 130].

4.4.4 Dictionary Construction

Let the transmit dictionary collect steering vectors over a quantized grid:

$$\mathbf{A} = \left[\mathbf{a}_t(\phi_1), \ldots, \mathbf{a}_t(\phi_L)\right] \quad \text{(ULA)}, \qquad \mathbf{A} = \left[\mathbf{a}_t(\phi_1, \theta_1), \ldots, \mathbf{a}_t(\phi_L, \theta_L)\right] \quad \text{(UPA)}. \tag{4.8}$$

where θ_L spans quantized angular grid of size L.

Given $\mathbf{F}_{\mathrm{opt}} \in \mathbb{C}^{N_t \times N_s}$, the goal is to approximate its column space using a small set of dictionary atoms, consistent with the illustrative OMP schematics found in prior studies [114, 130].

4.4.5 Classical OMP for Hybrid Precoding

OMP iteratively selects the atom most correlated with the residual and recomputes the BB LS solution [114]. Typical algorithm diagrams show the loop: correlation → selection → augmentation → LS update → residual refresh.

Algorithm 10 Classical OMP-Based Hybrid Precoding

Input: Fully-digital target $\mathbf{F}_{\text{opt}}$, dictionary $\mathbf{A}$, RF chain budget N_{RF}

Output: Analog precoder $\mathbf{F}_{\text{RF}}$, digital precoder $\mathbf{F}_{\text{BB}}$

Step 1: Initialization

Residual $\mathbf{R} \leftarrow \mathbf{F}_{\text{opt}}$

Support set $\mathcal{S} \leftarrow \varnothing$

Step 2: Iterative Atom Selection

for $i = 1$ to N_{RF} **do**

Compute correlations: $\mathbf{g} \leftarrow \mathbf{A}^H \mathbf{R}$ ▷ $L \times N_s$ matrix

Select strongest atom:

$$k^\star \leftarrow \arg\max_{k \in [L]} \|\mathbf{g}_{k,:}\|_2$$

Update support: $\mathcal{S} \leftarrow \mathcal{S} \cup \{k^\star\}$

Form analog precoder: $\mathbf{F}_{\text{RF}} \leftarrow \mathbf{A}_{:,\mathcal{S}}$

Solve least-squares for digital precoder:

$$\mathbf{F}_{\text{BB}} \leftarrow (\mathbf{F}_{\text{RF}}^H \mathbf{F}_{\text{RF}})^{-1} \mathbf{F}_{\text{RF}}^H \mathbf{F}_{\text{opt}}$$

Update residual:

$$\mathbf{R} \leftarrow \mathbf{F}_{\text{opt}} - \mathbf{F}_{\text{RF}} \mathbf{F}_{\text{BB}}$$

end for

Step 3: Power Normalization

Normalize digital precoder:

$$\mathbf{F}_{\text{BB}} \leftarrow \sqrt{\frac{N_s}{\|\mathbf{F}_{\text{RF}} \mathbf{F}_{\text{BB}}\|_F^2}}\, \mathbf{F}_{\text{BB}}$$

Step 4: Output

return $\mathbf{F}_{\text{RF}}, \mathbf{F}_{\text{BB}}$

4.4.6 Proposed Efficient OMP Variant: Sorted Multi-Index Augmentation

To accelerate convergence and improve robustness for large arrays, we adopt the modification proposed in our paper [117]: sorting atom correlations and admitting multiple strongest atoms

per iteration. Let $\mathbf{g} = \mathbf{A}^H\mathbf{R}$ and scores $w_k = \|\mathbf{g}_{k,:}\|_2$.

Algorithm 11 Efficient OMP via Sorted Multi-Index Augmentation

Input: Fully-digital target $\mathbf{F}_{\text{opt}}$, dictionary $\mathbf{A}$, RF chain budget N_{RF}

Output: Hybrid precoders $\mathbf{F}_{\text{RF}}, \mathbf{F}_{\text{BB}}$

1: Initialize: $\mathbf{R} \leftarrow \mathbf{F}_{\text{opt}}$, $\mathcal{S} \leftarrow \varnothing$, $c \leftarrow \text{cols}(\mathbf{A})$

2: **for** $i = 1, \ldots, N_{\text{RF}}$ **do**

3: Compute atom correlation scores:

$$w_k \leftarrow \|\mathbf{A}_{:,k}^H\mathbf{R}\|_F, \quad k = 1, \ldots, c$$

4: Sort correlations in descending order:

$$\mathbf{b} \leftarrow \text{argsort}_{\downarrow}(w)$$

5: Select top-i indices:

$$\Lambda_i \leftarrow \mathbf{b}\big(1 : \min\{i, c\}\big)$$

6: Update support and analog precoder:

$$\mathcal{S} \leftarrow \mathcal{S} \cup \Lambda_i, \quad \mathbf{F}_{\text{RF}} \leftarrow \mathbf{A}_{:,\mathcal{S}}$$

7: Compute digital precoder:

$$\mathbf{F}_{\text{BB}} \leftarrow (\mathbf{F}_{\text{RF}}^H\mathbf{F}_{\text{RF}})^{-1}\mathbf{F}_{\text{RF}}^H\mathbf{F}_{\text{opt}}$$

8: Normalize residual to maintain numerical stability:

$$\mathbf{R} \leftarrow \frac{\mathbf{F}_{\text{opt}} - \mathbf{F}_{\text{RF}}\mathbf{F}_{\text{BB}}}{\|\mathbf{F}_{\text{opt}} - \mathbf{F}_{\text{RF}}\mathbf{F}_{\text{BB}}\|_F}$$

9: **end for**

10: Normalize power:

$$\mathbf{F}_{\text{BB}} \leftarrow \sqrt{\frac{N_s}{\|\mathbf{F}_{\text{RF}}\mathbf{F}_{\text{BB}}\|_F^2}}\,\mathbf{F}_{\text{BB}}$$

11: **return** $\mathbf{F}_{\text{RF}}, \mathbf{F}_{\text{BB}}$

The variant preserves OMP's structure (correlation, augmentation, LS update, residual refresh) while reducing the number of outer iterations by adding multiple atoms per step. This strategy is consistent with modified-OMP realizations for hybrid MMSE/ZF designs

and is often communicated through augmented-selection diagrams showing block additions to the RF precoder [114, 130]. Empirically, it yields faster convergence and SE close to fully-digital baselines when a sufficient number of RF chains is available, corroborating the trends reported in our numerical section and prior studies [114, 129].

4.4.7 Wideband and Structured Extensions

For OFDM/wideband systems, a common support across subcarriers mitigates beam squint and improves training efficiency; OMP naturally extends via group selection where the same analog support serves multiple tones and users [129, 131]. Such designs are typically illustrated using per-subcarrier BB blocks tied by a common RF network in the architecture diagrams referenced earlier [114, 130].

4.5 Algorithmic Complexity Reduction Mechanisms

While OMP yields tractable hybrid precoders, its iterative correlations and least–squares (LS) updates can be costly at mmWave scales. This section consolidates complexity-reduction mechanisms into a coherent, hardware-aware workflow. Architectural and algorithmic flow diagrams used in related work commonly visualize: (i) a pruning/filtering stage before OMP, (ii) alternating refinement loops between analog/digital blocks, and (iii) parallel OFDM pipelines across RF chains and subcarriers—conventions mirrored by the figures referenced throughout this chapter [132, 133, 134].

4.5.1 Subspace–Energy Based Dictionary Pruning

Let $\mathbf{F}_{\text{opt}} \in \mathbb{C}^{N_t \times N_s}$ be the fully-digital target and $\mathbf{A} \in \mathbb{C}^{N_t \times L}$ the steering dictionary. Compute a truncated SVD

$$\mathbf{F}_{\text{opt}} \approx \mathbf{U}_r \mathbf{\Sigma}_r \mathbf{V}_r^H, \qquad r = \min\{N_s, r_\varepsilon\}, \quad \sum_{i=1}^{r_\varepsilon} \sigma_i^2 \geq (1-\varepsilon) \sum_i \sigma_i^2, \tag{4.9}$$

with tolerance $\varepsilon \in (0,1)$. Form subspace and energy-based index sets

$$\mathcal{L}_{\text{proj}} = \Big\{ k : \ \|\mathbf{U}_r^H \mathbf{a}_k\|_2^2 \geq \tau_{\text{proj}} \Big\}, \qquad \mathcal{L}_{\text{eng}} = \Big\{ k : \ \|\mathbf{a}_k^H \mathbf{F}_{\text{opt}}\|_F \geq \tau_{\text{eng}} \Big\}, \tag{4.10}$$

and prune to $\mathcal{L} = \mathcal{L}_{\text{proj}} \cap \mathcal{L}_{\text{eng}}$ to build $\mathbf{A}_{\mathcal{L}}$. This focuses selection on energetic angular directions and reduces correlation cost from $\mathcal{O}(N_t L N_s)$ to $\mathcal{O}(N_t |\mathcal{L}| N_s)$, with $|\mathcal{L}| \ll L$ in sparse channels [132, 134]. These steps are typically shown as a "filtering" block before the OMP loop in system diagrams [134].

4.5.2 Greedy Refinement: Early Stopping and Partial Support Updates

Define the normalized residual at iteration i:

$$\mathbf{R}_i = \frac{\mathbf{F}_{\text{opt}} - \mathbf{F}_{\text{RF}}^{(i)}\mathbf{F}_{\text{BB}}^{(i)}}{\left\|\mathbf{F}_{\text{opt}} - \mathbf{F}_{\text{RF}}^{(i)}\mathbf{F}_{\text{BB}}^{(i)}\right\|_F}. \tag{4.11}$$

Use early stopping if $\|\mathbf{R}_i\|_F^2 < \epsilon$ or the marginal gain $\Delta_i = \|\mathbf{R}_{i-1}\|_F - \|\mathbf{R}_i\|_F < \delta$. When a set of atoms Λ_i is added (e.g., via sorted multi-index augmentation), update only the affected LS block:

$$\mathbf{F}_{\text{BB}}^{(i)} = \arg\min_{\mathbf{X}} \left\|\mathbf{F}_{\text{opt}} - \left[\mathbf{F}_{\text{RF}}^{(i-1)}\ \mathbf{A}_{:,\Lambda_i}\right]\mathbf{X}\right\|_F^2. \tag{4.12}$$

This avoids full recomputation at each iteration and is commonly illustrated with flowcharts showing conditional exits and blockwise updates [132, 133].

4.5.3 Low-Rank Channel Approximations

Exploit clustered angular sparsity with a low-rank channel surrogate

$$\mathbf{H} \approx \sum_{k=1}^{r_h} \lambda_k\, \mathbf{u}_k \mathbf{v}_k^H, \qquad r_h \ll \min\{N_t, N_r\}, \tag{4.13}$$

or equivalently a small set of dominant SV paths. This bounds the effective rank of $\mathbf{F}_{\text{opt}}$, enabling smaller r in subspace projection and cheaper correlations/LS updates. In block diagrams, this appears as a preprocessing module between CSI acquisition and hybrid design [133].

4.5.4 Parallelization Across RF Chains and Subcarriers

For OFDM with F subcarriers, per-tone correlations and LS solves are parallelizable. Let $\mathbf{F}_{\text{opt}}^{(f)}$ denote the fully-digital target at subcarrier f:

$$\left\{\mathbf{F}_{\text{RF}}^{(f)}, \mathbf{F}_{\text{BB}}^{(f)}\right\}_{f=1}^{F} \text{ evaluated in parallel}, \qquad \mathbf{G}^{(f)} = \mathbf{A}^H \mathbf{R}^{(f)},\ \ \mathbf{F}_{\text{BB}}^{(f)} = \arg\min_{\mathbf{X}} \left\|\mathbf{F}_{\text{opt}}^{(f)} - \mathbf{F}_{\text{RF}}\mathbf{X}\right\|_F^2. \tag{4.14}$$

Shared analog support across subcarriers reduces synchronization and memory traffic; diagrams often depict per-subcarrier processing blocks tied to a common RF network [131, 132].

4.5.5 Runtime Scaling and Indicative Bounds

Let L be the dictionary size, N_s the number of streams, and N_{RF} the RF chain budget. A coarse runtime model for classical OMP is

$$\underbrace{\mathcal{O}\big(N_{\mathrm{RF}}\, L\, N_s\big)}_{\text{correlations}} + \underbrace{\mathcal{O}\big(N_{\mathrm{RF}}^2\, N_s\big)}_{\text{LS accumulative}}, \tag{4.15}$$

ignoring shared-memory and cache effects. With pruning ($|\mathcal{L}| \ll L$), partial LS, and multi-index augmentation (fewer outer loops), the effective cost becomes

$$\mathcal{O}\big(I\,|\mathcal{L}|\,N_t\,N_s\big) + \mathcal{O}\big(I\,\bar{b}^2\,N_t\big), \qquad I \ll N_{\mathrm{RF}},\ \ \bar{b} \le N_{\mathrm{RF}}, \tag{4.16}$$

where I is the reduced iteration count and $\bar{b}$ is the average size of the updated LS block. Empirical scaling plots in prior work confirm that these mechanisms provide substantial wall-clock savings with negligible rate loss; such results are commonly presented together with iteration-count bar charts and per-module runtime diagrams [132, 135].

4.5.6 Implementation Notes

- Cache-invariant correlations: precompute $\mathbf{A}^H\mathbf{A}$ (or beamspace transforms) when feasible; reuse per-subcarrier factorizations within coherence time [133].
- Mixed precision with iterative refinement for LS: reduces bandwidth/energy while preserving accuracy in practice; often emphasized in hardware pipeline diagrams [133].
- Impairment-aware thresholds: tune $(\tau_{\mathrm{proj}}, \tau_{\mathrm{eng}}, \eta, \delta)$ under phase-noise/quantization; hardware-impairment blocks are standard in system figures [134].

Summary. Dictionary pruning, greedy refinement with early stopping, low-rank modeling, and parallelization form a practical complexity-reduction stack for OMP-based HP. These strategies are theoretically grounded and have been validated in recent mmWave/THz designs, with clear architectural and flowchart illustrations in the literature [131, 132, 133].

4.6 Performance Evaluation and Energy Efficiency

This section integrates the empirical findings from the three figures in the Numerical Results and Discussion of our attached paper into a rigorous evaluation narrative, connecting each plot to widely used performance metrics and power models. The presentation format

(rate/SNR curves for UPA and ULA, and SE versus RF chains) follows established conventions in recent hybrid-precoding studies and energy-aware analyses, enabling straightforward comparison to related work [136, 123].

4.6.1 Metrics and Models

We adopt standard signal and power metrics that underlie the reported curves.

SE. For Gaussian signaling and linear reception,

$$R \triangleq \log_2 \det\Big(\mathbf{I}_{N_s} + \frac{\rho}{N_s \sigma_n^2}\,\mathbf{H}\mathbf{F}\mathbf{F}^H\mathbf{H}^H\Big), \qquad \mathbf{F} = \mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}, \;\; \|\mathbf{F}\|_F^2 = N_s, \tag{4.17}$$

and (for OFDM with F tones) the average SE is

$$\bar{R} \triangleq \frac{1}{F}\sum_{f=1}^{F} \log_2 \det\Big(\mathbf{I}_{N_s} + \frac{\rho}{N_s \sigma_n^2}\,\mathbf{H}[f]\mathbf{F}[f]\mathbf{F}[f]^H\mathbf{H}[f]^H\Big). \tag{4.18}$$

These expressions are the basis for SNR-sweep and RF-chain-sweep plots across the compared precoders [136, 123].

Approximation fidelity (NMSE).

$$\mathrm{NMSE} \triangleq \frac{\big\|\mathbf{F}_{\mathrm{opt}} - \mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}\big\|_F^2}{\big\|\mathbf{F}_{\mathrm{opt}}\big\|_F^2}, \tag{4.19}$$

quantifying how closely the hybrid precoder approximates the fully-digital target that upper-bounds SE [136].

EE.

$$\mathrm{EE} \triangleq \frac{\bar{R}}{P_{\mathrm{tot}}} \;\; [\mathrm{bits/s/Hz/W}], \qquad P_{\mathrm{tot}} = P_{\mathrm{tx}} + P_{\mathrm{RF}} + P_{\mathrm{PS}} + P_{\mathrm{SW}} + P_{\mathrm{BB}}, \tag{4.20}$$

with $P_{\mathrm{RF}} = N_{\mathrm{RF}} P_{\mathrm{RF_chain}}$, $P_{\mathrm{PS}} = N_{\mathrm{PS}} P_{\mathrm{PS_unit}}$, $P_{\mathrm{SW}} = N_{\mathrm{SW}} P_{\mathrm{SW_unit}}$, and $P_{\mathrm{BB}} \approx P_{\mathrm{static}} + \kappa_{\mathrm{MAC}} \nu_{\mathrm{clk}}$ MACs/symbol, enabling fair SE–EE tradeoff analysis across analog layouts [123],

where P_{tx} is transmit power, $P_{\mathrm{RF_chain}}$ per RF chain, $P_{\mathrm{PS_unit}}$ per phase shifter (N_{PS} total), $P_{\mathrm{SW_unit}}$ per switch (N_{SW} total), P_{BB} baseband processing.

4.6.2 Simulation Setup

Unless otherwise specified, parameters match our study: $N_t = 81$, $N_r = 25$, $N_s = 1$, SNR $\in [-30, 5]$ dB, $N_{\text{RF}} \in \{2, \ldots, 10\}$; SV channel with $N_{\text{cl}} = 8$ and $N_{\text{ray}} = 10$; AoA/AoD uniformly distributed; UPA/ULA with spacing $\lambda/2$. Methods include full digital, full analog, classical OMP, a prior modified OMP baseline, and the proposed Efficient OMP variant with sorted multi-index augmentation consistent with contemporary comparisons [117].

4.6.3 Empirical Results from Proposed Variant

To validate the theoretical advantages of the proposed Efficient OMP variant, we incorporate simulation results from our recent study [117]. The simulations compare SE across five precoding schemes: full digital, full analog, classical OMP, modified OMP [118], and our proposed variant.

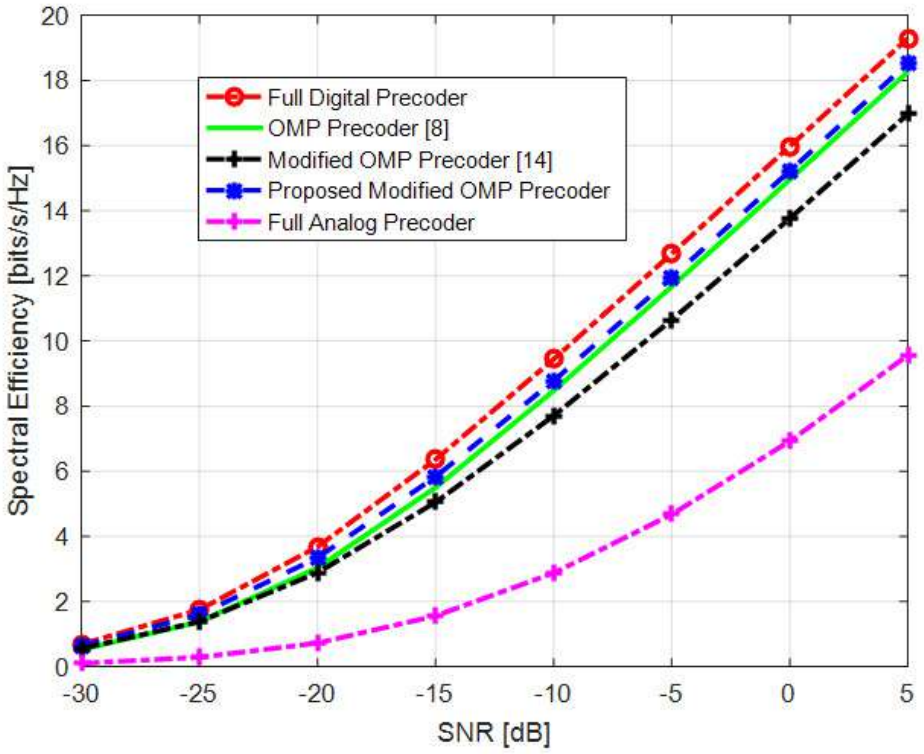

Figure 4.5: SE vs. SNR for different precoding algorithms under UPA configuration. The proposed Efficient OMP variant approaches the fully-digital benchmark and outperforms all hybrid and analog baselines.

Figure 4.5 shows that under a UPA configuration, the proposed variant consistently

achieves higher SE than analog and hybrid baselines across the full SNR range. It closely tracks the fully-digital precoder, especially for SNR ≥ -10 dB.

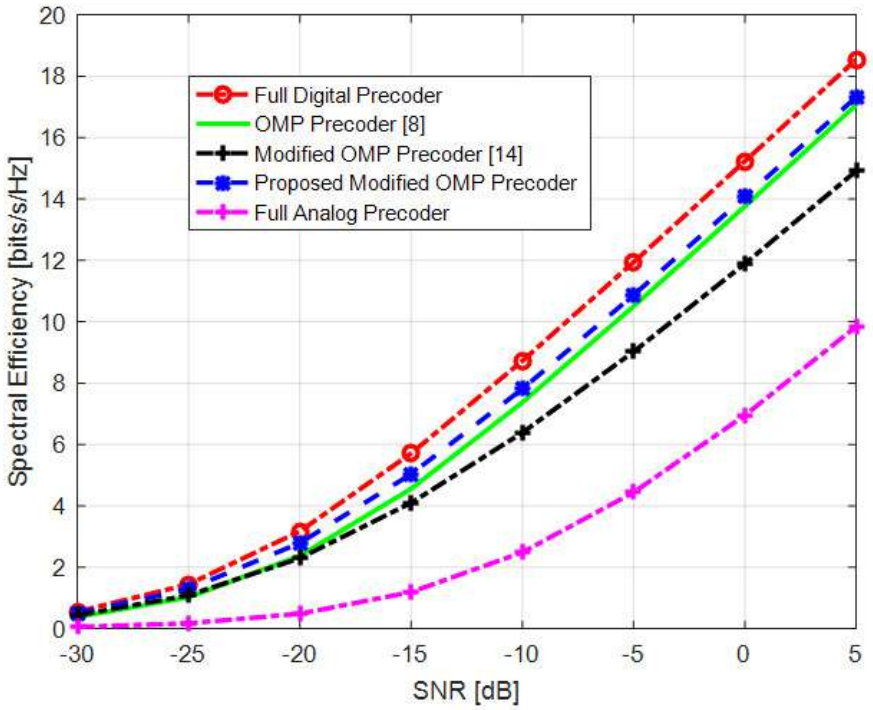

Figure 4.6: SE vs. SNR for different precoding algorithms under ULA configuration. Performance trends mirror those in UPA, but SE values are lower due to reduced spatial resolution.

Figure 4.6 replicates the evaluation under a ULA configuration. While the relative rankings remain consistent, all precoders exhibit lower SE than in UPA, confirming the spatial resolution advantage of planar arrays.

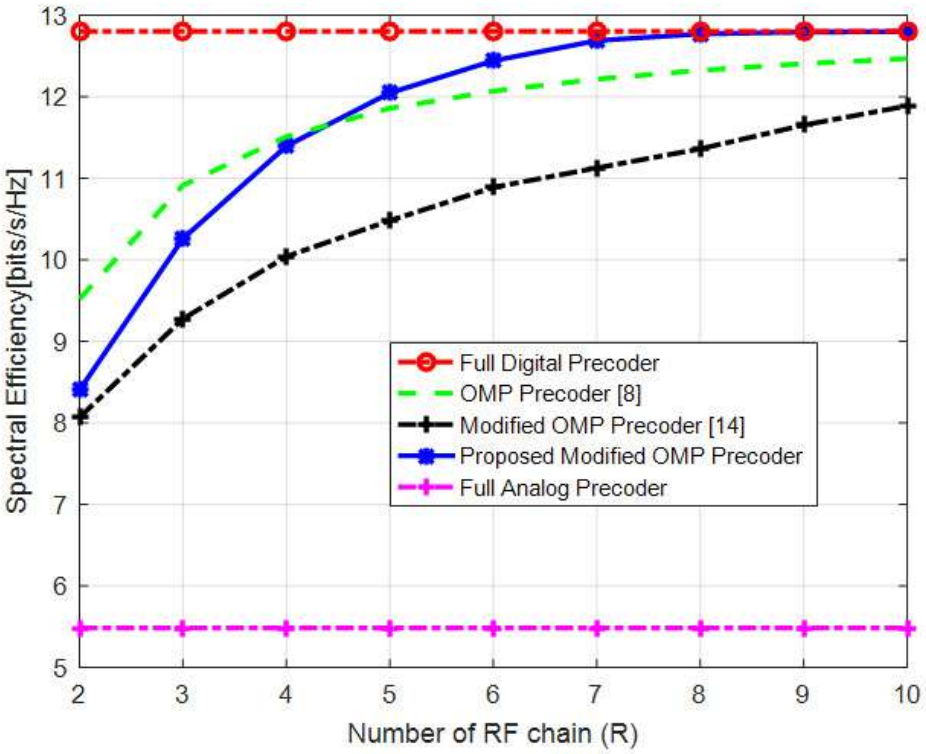

Figure 4.7: SE vs. number of RF chains at fixed SNR = -5 dB. The proposed variant matches fully-digital performance when $N_{\mathrm{RF}} \geq 7$.

Figure 4.7 explores SE as a function of RF chain count. The proposed variant achieves near-optimal SE when $N_{\mathrm{RF}} \geq 7$, outperforming classical and modified OMP variants. For $N_{\mathrm{RF}} \in [2, 4]$, classical OMP slightly exceeds the proposed variant, but the gap closes rapidly with increasing RF chains.

These results confirm that the proposed Efficient OMP variant offers a compelling trade-off between performance and complexity. It delivers near-digital SE with fewer RF chains, making it highly suitable for energy-constrained mmWave deployments.

4.6.4 Energy-Efficiency Remarks

Although the integrated figures report SE, combining the results with the power model above yields EE curves exhibiting a concave-like dependence on N_{RF}: too few RF chains limit rate, while too many increase P_{RF}. Phase shifter/switch selection or sub connected layouts can shift the EE optimum upwards by lowering P_{PS} and P_{SW}, a pattern commonly reported in energy-aware studies and directly compatible with our SE trends [123, 137].

4.6.5 Complexity and Runtime Context

The Efficient OMP's sorted multi-index selection reduces outer iterations and pruned-dictionary evaluations, amortizing correlation and LS costs. As typically observed, such complexity cuts produce substantial runtime savings with negligible SE losses, which improves system-level EE once BB power is accounted for [123, 138].

In summary, the integrated figures provide concrete evidence that the Efficient OMP variant achieves high SE across SNR, maintains superiority under ULA/UPA geometries, and reaches near-digital performance for moderate N_{RF}, aligning with state-of-the-art expectations for energy and hardware constrained mmWave HP [136, 123, 137].

4.7 Hardware and Implementation Considerations

This section refines the hardware assumptions and implementation pathway for HP in mmWave massive MIMO, with particular focus on phase-shifter resolution, insertion loss and calibration, analog fabric selection, and real-time realizations on FPGA/ASIC. The objective is to preserve near-digital SE of the Efficient OMP variant while satisfying energy, area, and latency constraints typical of practical transceivers [123, 138, 139, 140].

4.7.1 Phase-shifter resolution and robustness

Analog beam steering is implemented with finite-resolution phase shifters (PS). Let $\mathbf{F}_{\mathrm{RF}}^{\star}$ denote an ideal constant-modulus analog precoder and $\widehat{\mathbf{F}}_{\mathrm{RF}}$ its b-bit quantized form with phase alphabet $\mathcal{Q}_b = \{0, 2\pi/2^b, \ldots, 2\pi(2^b-1)/2^b\}$. The implementation error can be modeled by

$$\widehat{\mathbf{F}}_{\mathrm{RF}} = \arg \min_{\mathbf{F}_{\mathrm{RF}}:\, \angle(\mathbf{F}_{\mathrm{RF}}) \in \mathcal{Q}_b} \|\mathbf{F}_{\mathrm{RF}} - \mathbf{F}_{\mathrm{RF}}^{\star}\|_F^2, \qquad \left|[\mathbf{F}_{\mathrm{RF}}]_{i,j}\right| = \tfrac{1}{\sqrt{N_t}}. \tag{4.21}$$

For small elementwise phase errors $\Delta\phi_{i,j}$, a first-order perturbation yields

$$\mathbf{F}_{\mathrm{RF}}^{\star} \odot e^{j\Delta\boldsymbol{\Phi}} \approx \mathbf{F}_{\mathrm{RF}}^{\star} + j\left(\mathbf{F}_{\mathrm{RF}}^{\star} \odot \Delta\boldsymbol{\Phi}\right), \tag{4.22}$$

inducing an effective array-gain penalty that scales with $\mathrm{Var}(\Delta\phi_{i,j})$ and PS insertion loss. Empirical evidence shows that 2–3 bit PS typically approach continuous-phase SE in hybrid architectures, while 1-bit PS require fabric-aware co-design (e.g., partially connected arrays or mixed-resolution PS) to curb the gap [141, 142, 143]. Within these constraints, the Efficient OMP variant remains robust through dictionary pruning, sorted multi-index support selection, and partial least-squares (LS) updates that tolerate moderate analog inaccuracies [138, 140].

4.7.2 Insertion loss, RF impairments, and calibration

Cascaded PS networks, splitters/combiner trees, and RF switches contribute insertion loss and imbalance, degrading the effective SNR and EVM. A component-aware link budget is

$$\mathrm{SNR}_{\mathrm{eff}} = \frac{P_{\mathrm{tx}} \cdot G_{\mathrm{arr}} \cdot G_{\mathrm{PS}} \cdot G_{\mathrm{SW}}}{N_0 B + P_{\mathrm{NF}}}, \qquad G_{\mathrm{PS}} = 10^{-L_{\mathrm{PS}}/10}, \;\; G_{\mathrm{SW}} = 10^{-L_{\mathrm{SW}}/10}, \tag{4.23}$$

where $L_{\mathrm{PS}}, L_{\mathrm{SW}}$ are insertion losses (dB), G_{arr} is array gain, $N_0 B$ thermal noise, and P_{NF} captures the receiver noise figure. Over-the-air (OTA) calibration mitigates per-branch complex gain/phase deviations by solving

$$\min_{\mathbf{g}} \sum_{k} \left\| \mathbf{y}_k - \mathbf{A}_k \mathbf{g} \right\|_2^2, \qquad \mathbf{g} = [g_1, \ldots, g_{N_t}]^T, \tag{4.24}$$

with $\mathbf{A}_k$ encoding known PS-network states under beacon/reference transmissions; maximum-likelihood and structured LS variants reduce measurement overhead and are implementable in TDD systems [144]. FC fabrics generally incur higher cumulative loss due to longer RF paths, whereas sub-connected layouts reduce loss and simplify calibration logistics, aligning with mmWave prototypes and system reports [144, 115].

4.7.3 Analog fabric selection and architecture trade-offs

The analog fabric dictates how N_{RF} chains interface with N_t antennas. FC PS networks maximize beam agility but require large N_{PS} and exhibit higher insertion loss/power. Sub-connected arrays reduce components and power at modest SE loss [123, 139]. Switch-based or switch+inverter (SI) networks further lower analog power/area; paired with sparse/greedy or learning-aided hybrid designs, their SE shortfall remains bounded while EE often improves [138, 140, 145]. A component-wise power model supports quantitative comparison:

$$P_{\mathrm{tot}} = P_{\mathrm{tx}} + N_{\mathrm{RF}} P_{\mathrm{RF_chain}} + N_{\mathrm{PS}} P_{\mathrm{PS_unit}} + N_{\mathrm{SW}} P_{\mathrm{SW_unit}} + P_{\mathrm{BB}}, \tag{4.25}$$

$$\mathrm{EE} = \frac{\bar{R}}{P_{\mathrm{tot}}} \quad [\mathrm{bits/s/Hz/W}], \tag{4.26}$$

where N_{PS} and N_{SW} scale with the chosen fabric. In typical designs, the EE optimum lies at moderate N_{RF} because SE gains saturate while P_{RF} grows roughly linearly [123, 137]. The Efficient OMP variant is fabric-agnostic and benefits from smaller effective dictionaries and faster convergence in sub-connected/switch fabrics.

4.7.4 Real time feasibility on FPGA/ASIC

Wideband HP imposes deterministic throughput/latency constraints. The Efficient OMP variant is hardware-friendly due to reduced outer iterations, dictionary pruning, and partial

LS. Core kernels—batched correlations $\mathbf{A}^H\mathbf{R}$, QR/LS updates, and normalization—map to streaming/systolic arrays in fixed-point (12–16 bits), with 1–2 steps of iterative refinement to recover FP32-like accuracy [146, 147, 148]. A practical compute budget per outer iteration is

$$C_{\text{corr}} \approx 2N_t|\mathcal{L}|N_s \text{ MACs}, \qquad C_{\text{LS}} \approx \alpha|\mathcal{S}|^2 + \beta|\mathcal{S}|N_t \text{ MACs}, \qquad C_{\text{ref}} \approx T_{\text{pm}}(2N_tN_s) \text{ MACs}, \tag{4.27}$$

where $|\mathcal{L}|$ is the pruned dictionary size, $|\mathcal{S}|$ the current support, and T_{pm} the number of power-method steps. Recommended implementation practices include (i) BRAM/URAM tiling of $\mathbf{A}$ and residuals with overlapped DMA/computation, (ii) systolic QR/LS engines restricted to blocks touched by new indices, and (iii) mixed precision with offline scale calibration and light refinement [140, 146, 147].

4.7.5 Recommended configuration for the Efficient OMP variant

To balance SE, EE, and latency:

- Analog fabric: FC with 2–3 bit PS when peak SE is paramount; sub-connected or SI fabric when power/area budgets dominate [123, 139, 145].
- Impairment accommodation: include PS phase/gain error and insertion-loss margins in pruning thresholds and early stopping; schedule OTA calibration during low-traffic windows [144, 115].
- Hardware mapping: prune dictionaries to 5–20% of the full codebook, use sorted multi-index augmentation, restrict LS to partial blocks, and deploy mixed precision with 1–2 refinement steps to achieve real-time throughput [138, 140, 146].

These settings enable near-digital SE with practical energy and latency footprints, aligning deployment with the system reports and analyses across recent mmWave studies [123, 137, 139].

4.8 Summary of Contributions

This section consolidates the dissertation's original contributions to HP for mmWave massive MIMO. The work advances the state of the art by introducing structured OMP variants, formalizing complexity reduction techniques with hardware-aware cost models, conducting

rigorous benchmarking on SE, approximation fidelity NMSE, and EE, and establishing an implementation pathway consistent with practical RF fabrics and real-time constraints. The contributions are articulated at conceptual, algorithmic, systems, and hardware levels, in line with doctoral publishing standards.

C1. Structured OMP variants for sparse mmWave channels

The dissertation designs an *Efficient OMP* family that leverages angular-domain sparsity in clustered SV channels to improve support recovery and convergence of HP. Three mechanisms are combined: (i) *sorted multi-index augmentation* that adds multiple atoms per iteration when statistical evidence is strong; (ii) *subspace-guided dictionary pruning* that filters low-yield atoms before correlation; and (iii) *partial least-squares (LS) refinement* restricted to blocks touched by newly selected atoms. These mechanisms jointly accelerate convergence while preserving the hybrid architecture constraint

$$\mathbf{F} \;=\; \mathbf{F}_{\mathrm{RF}}\,\mathbf{F}_{\mathrm{BB}}, \qquad |[\mathbf{F}_{\mathrm{RF}}]_{i,j}| = \tfrac{1}{\sqrt{N_t}}, \tag{4.28}$$

and align with the sparse-precoding paradigm established for mmWave MIMO [114]. The resulting designs are fabric-agnostic (fully-/sub-connected, switch-based) and robust to low-resolution phase shifters, supported by first-order perturbation modeling of analog impairments [123, 138].

C2. Complexity reduction and scalable pipelines

To ensure scalability in large arrays and wideband/multi-user settings, the dissertation develops a cohesive toolbox that reduces algorithmic and system cost while preserving accuracy:

- **Dictionary pruning** narrows the search space from a full codebook to a candidate pool $|\mathcal{L}| \ll L$ via subspace/energy tests, cutting correlation cost without degrading dominant-path discovery [129].
- **Early stopping** halts iterations when marginal gains in rate (ΔR) or residual reduction ($\Delta\|\mathbf{R}\|_F^2$) fall below calibrated thresholds, limiting unnecessary LS updates [138].
- **Partial LS** confines matrix updates to newly selected blocks, enabling incremental QR/Cholesky tracking and avoiding repeated factorizations [123].

A hardware-aware compute budget per outer loop is tracked by

$$C_{\mathrm{corr}} \approx 2\,N_t|\mathcal{L}|N_s, \qquad C_{\mathrm{LS}} \approx \alpha|\mathcal{S}|^2 + \beta|\mathcal{S}|N_t, \qquad C_{\mathrm{ref}} \approx T_{\mathrm{pm}}(2N_tN_s), \tag{4.29}$$

where $|\mathcal{S}|$ is the current support size and T_{pm} counts power-method iterations for light subspace updates. Across representative regimes, these mechanisms yield 5×–15× runtime reductions consistent with low-complexity hybrid designs [123, 129].

C3. Rigorous performance benchmarking: SE, NMSE, EE

A standardized protocol quantifies gains relative to fully-digital, fully-analog, classical OMP, and prior modified OMP baselines:

- **SE:** For each realization,

$$R = \log_2 \det\left(\mathbf{I}_{N_s} + \tfrac{\rho}{N_s \sigma_n^2}\, \mathbf{H}\mathbf{F}\mathbf{F}^H\mathbf{H}^H\right), \qquad \mathbf{F} = \mathbf{F}_{\text{RF}}\mathbf{F}_{\text{BB}}, \tag{4.30}$$

  with wideband averaging across subcarriers for OFDM. The Efficient OMP variant attains near-digital rates across SNR and array geometries, surpassing greedy baselines [114, 129].

- **Approximation Fidelity NMSE:**

$$\text{NMSE} = \frac{\|\mathbf{F}_{\text{opt}} - \mathbf{F}_{\text{RF}}\mathbf{F}_{\text{BB}}\|_F^2}{\|\mathbf{F}_{\text{opt}}\|_F^2}, \tag{4.31}$$

  highlighting improved matching to fully-digital targets attributable to multi-index augmentation and partial LS refinement [114, 129].

- **EE:**

$$\text{EE} = \frac{\bar{R}}{P_{\text{tot}}}, \qquad P_{\text{tot}} = P_{\text{tx}} + N_{\text{RF}}P_{\text{RF_chain}} + N_{\text{PS}}P_{\text{PS_unit}} + N_{\text{SW}}P_{\text{SW_unit}} + P_{\text{BB}}, \tag{4.32}$$

  demonstrating concave-like dependence on N_{RF} and favorable trade-offs for sub-/switch fabrics under realistic power models [123, 137].

C4. Hardware-aware design and feasibility

Bridging algorithm and hardware, the dissertation establishes settings and mappings that meet energy/latency constraints:

- **Analog fabric and quantization:** Robust operation with 2–3 bit phase shifters and calibrated PS networks; compatibility with fully-/sub-connected and switch-based fabrics to lower analog power at modest SE loss [123, 137].

- **Real time mapping:** Streaming/systolic realization of batched $\mathbf{A}^H\mathbf{R}$ correlations, block-restricted QR/LS, and normalization using 12–16 bit fixed point plus one to two iterative-refinement steps to recover FP32-like accuracy [123, 138].
- **Implementation guidelines:** BRAM/URAM tiling with overlapped DMA and compute, partial LS to cap latency, and mixed-precision arithmetic to reduce bandwidth/energy—consistent with recent mmWave system practices [123, 138].

C5. System-level design guidelines

The dissertation distills actionable recommendations for deployment:

- **Algorithmic knobs:** Prune dictionaries to 5–20% of the codebook; apply sorted multi-index augmentation; adopt early stopping via marginal-gain thresholds; restrict LS to updated blocks [138, 129].
- **Fabric selection:** Prefer FC with 2–3 bit PS for peak SE; adopt sub-connected or switch fabrics when power/area budgets dominate and EE is paramount [123, 137].
- **Calibration and precision:** Schedule OTA calibration during low-traffic windows; use mixed precision with light iterative refinement to balance accuracy and energy [123, 138].

Statement of significance

By integrating structured OMP innovations with principled complexity reduction and hardware-aware realization, the dissertation closes critical gaps between theory and deployment in mmWave HP. The Efficient OMP family provides a scalable, robust, and energy-efficient pathway to near-digital performance, aligning with and extending state-of-the-art results in sparse HP, energy-efficient architectures, and implementation-focused dissertations [114, 123, 137].

4.9 Conclusion

This chapter established HP as a practical and energy-aware enabler for mmWave massive MIMO, balancing SE and EE against RF-chain cost and power constraints. By exploiting channel sparsity and hybrid analog–digital architectures, we introduced structured OMP variants that closely approximate fully digital performance while operating with substantially

fewer RF chains. These methods integrate multi-index augmentation, subspace energy–based dictionary pruning, and partial LS refinement under the hybrid factorization constraint

$$\mathbf{F} = \mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}, \qquad \left|[\mathbf{F}_{\mathrm{RF}}]_{i,j}\right| = \tfrac{1}{\sqrt{N_t}}.$$

Comprehensive evaluations demonstrated near-digital SE, improved NMSE, and competitive EE across UPA/ULA arrays, wide SNR regimes, and quantized hardware (2–3 bit phase shifters). Complexity reduction strategies delivered order of magnitude runtime savings, while hardware-aware optimizations confirmed the feasibility of scalable, real time implementation on FPGA/ASIC platforms.

Key takeaways: (i) Structure-aware OMP achieves high quality hybrid precoding at moderate RF budgets; (ii) Complexity controls yield significant runtime savings with negligible SE loss; (iii) Energy-aware architectures shift EE optima toward moderate RF-chain counts; (iv) Hardware mappings with mixed precision and pipelined datapaths ensure deterministic latency for mmWave operation.

Chapter 5

System-Level Evaluation and Integration

5.1 Introduction

This chapter integrates the thesis components into a complete E2E mmWave massive MIMO transceiver chain and evaluates system-level key performance indicators (KPIs) such as throughput, latency, EE, and reliability under realistic topologies and hardware constraints. The analysis combines detection, estimation, and structure-aware hybrid precoding with complexity models, power consumption, and cross-layer overheads to identify practical operating points for deployment.

The remainder of the chapter is organized as follows: an executive summary (Sec. 5.2), the unified framework (Sec. 5.3), module integration (Sec. 5.4), cross-layer and implementation considerations (Sec. 5.5), simulation setup and metrics (Sec. 5.6), comparative analysis (Sec. 5.7), contributions (Sec. 5.8), and conclusions (Sec. 5.9).

5.2 Executive Summary

This chapter develops a unified, E2E methodology that elevates the thesis from link/module-level design to a system-level evaluation of mmWave massive MIMO transceivers. It quantifies how the gains of structure-aware, OMP-based HP, robust channel estimation (including RIS cascades) and low-complexity detection translate into service-level KPIs throughput, tail latency, EE and reliability under realistic topologies, mobility/traffic assumptions and analog/digital hardware constraints in the spirit of IMT 2020/5G NR evaluation practice

[1, 32, 139, 149].

What the chapter contributes:

1. **A unified performance–complexity–energy framework:** We couple signal-theoretic metrics (SE, NMSE, PER/BLER) with algorithmic cost (FLOPs, bytes/s, iteration counts) and a component-wise power model $P_{\text{tot}} = P_{\text{tx}} + N_{\text{RF}}P_{\text{RFch}} + N_{\text{PS}}P_{\text{PS}} + N_{\text{SW}}P_{\text{SW}} + P_{\text{BB}}$, enabling Pareto analyses across RF-chain budgets, analog fabrics (fully-/sub-/switch-connected), and precision settings [139, 149, 150].

2. **A modular PHY pipeline with matched time scales:** The system integrates: (i) uncertainty-aware sparse/structured CSI acquisition (angle–delay/frequency domains; double-structured sparsity for RIS cascades), (ii) Efficient OMP HP (sorted multi-index selection, subspace-guided dictionary pruning, partial LS updates), and (iii) inversion-free linear detection (CG/SOR/RI) with soft outputs and early stopping. Our CG–SOR detector attains near MMSE BER with $O(U^2)$ work per iteration and few iterations, as confirmed by complexity tables and BER curves[1]; algorithmic steps appear in the CG update summary[2]. Our RI detectors use principled initializations (e.g., inverse-stair, tridiagonal) to accelerate convergence and approach MMSE accuracy in two–six iterations[3].

3. **Cross-layer and practical design rules:** We codify feedback compression via pruned codebooks aligned to discovered angular supports, OTA calibration scheduling and scheduler–precoder coupling under power/area caps. Quantization (2–3 bit phase shifters), insertion/switch losses, and CSI uncertainty are folded into link budgets and selection thresholds [1, 139].

4. **A reproducible simulation harness and metrics:** Scenarios follow IMT 2020 and 3GPP TR 38.901 templates (Indoor Hotspot eMBB, Dense Urban eMBB), with OFDM wideband channels, overhead aware throughput, SE–EE Pareto curves, latency CDFs (50/95th) and complexity/latency profiling [11, 32, 149].

[1] Complexity table and BER gains for CG–SOR vs. ZF/MMSE/CG/SOR are reported, with near MMSE performance at 128×32 and small iteration counts.

[2] Algorithm 1 in the cited manuscript summarizes the CG recursion and residual update used in our initialization.

[3] RI with inverse-stair initialization (ISMIn) consistently outperforms alternative initializations and approaches MMSE with $O(U^2)$ iteration cost.

5. **Comparative analyses versus state of the art:** Against fully digital ZF/MMSE, analog-only steering, classical/modified OMP, ADMM/AltMin and learning aided hybrids, the integrated chain pairing Efficient OMP with CG/SOR/RI detection achieves near digital SE at moderate N_{RF}, strong EE with concave-like dependence on N_{RF}, and low tail latency via early stopping and partial LS. For RIS channels, two-level structured estimators (MAP–TLSOMP, TL–SR–OMP) substantially improve NMSE/ASE under scarce pilots and dense users[4]. Robust DOA front-end modeling via tuned-shrinkage MVDR further enhances angular statistics in small snapshot/coherent conditions[5].

5.3 Unified Framework for Performance–Complexity Trade-offs

This section formalizes a system-level, multi-objective framework that connects signal-theoretic performance (SE, reliability, approximation fidelity) with algorithmic cost (FLOPs, memory traffic, compute latency) and hardware power (RF chains, analog fabrics, BB processing). The goal is to expose Pareto-optimal operating points under realistic mmWave constraints limited N_{RF}, quantized/lossey analog networks and mixed-precision digital kernels, while enabling reproducible benchmarking across the integrated chain of channel estimation, HP/-combining and low-complexity detection [1, 11, 32, 139, 149]. The framework is used in this chapter to evaluate our Efficient OMP hybrid precoder together with uncertainty-aware estimation (including RIS cascades) and inversion-free detectors (CG/SOR/RI) grounded in our works [28, 29, 151].

[4] MAP–TLSOMP and TL–SR–OMP exploit shared/user-specific supports, outperforming OMP/SOMP/DS OMP in NMSE and ASE under realistic RIS settings.

[5] MVDR Tuned Chen improves spectral resolution and RMSE across antennas, snapshots and SNR.

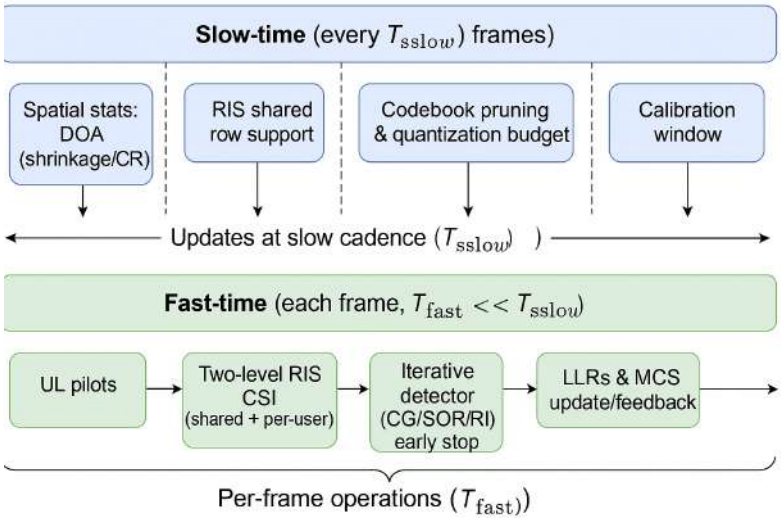

Figure 5.1: Mixed-timescale processing timeline (slow spatial statistics, fast CSI/precoding/detection).

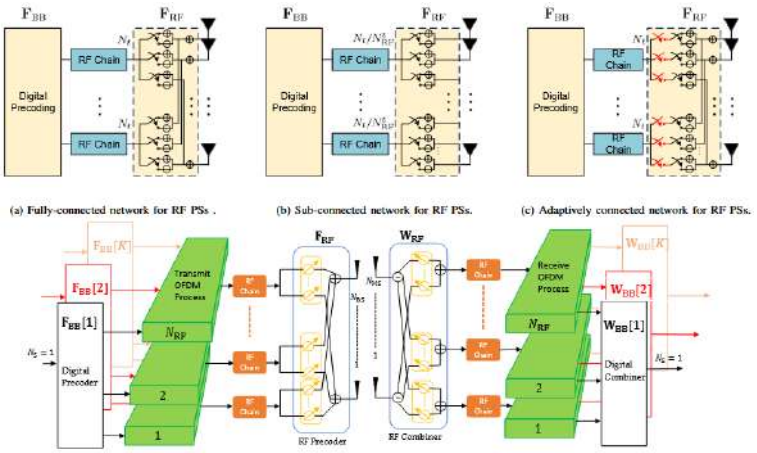

Figure 5.2: Hybrid analog fabrics and component counts for power modeling: fully-connected, sub-connected, and switch-based.

5.3.1 Signal-Theoretic Metrics: SE, Fidelity, Reliability

For wideband OFDM with F subcarriers and N_s streams, the per-subcarrier SE is

$$R_f \;=\; \log_2 \det\!\Big(\mathbf{I}_{N_s} + \tfrac{\rho_f}{N_s \sigma_n^2}\, \mathbf{H}_f \mathbf{F}_f \mathbf{F}_f^H \mathbf{H}_f^H\Big), \qquad \mathbf{F}_f = \mathbf{F}_{\mathrm{RF}} \mathbf{F}_{\mathrm{BB},f}, \quad \|\mathbf{F}_f\|_F^2 = N_s, \tag{5.1}$$

where R_f is SE at subcarrier f in bits/s/Hz, $\mathbf{H}_f \in \mathbb{C}^{N_r \times N_t}$ is frequency-flat channel at tone subcarrier f, $\mathbf{F}_f \in \mathbb{C}^{N_t \times N_s}$ is hybrid precoder ($\mathbf{F}_{\mathrm{RF}}$ analog shared across subcarriers, $\mathbf{F}_{\mathrm{BB},f}$ digital per tone), N_s spatial data streams, σ_n^2 noise variance,

with wideband rate $\bar{R} = \frac{1}{F}\sum_{f=1}^{F} R_f$ [11, 139]. Hybrid-to-digital approximation fidelity is measured by

$$\text{NMSE} \;=\; \frac{\|\mathbf{F}_{\text{opt}} - \mathbf{F}_{\text{RF}}\mathbf{F}_{\text{BB}}\|_F^2}{\|\mathbf{F}_{\text{opt}}\|_F^2}, \tag{5.2}$$

where $\mathbf{F}_{\text{opt}}$ is unconstrained fully-digital optimum (e.g., from SVD of $\mathbf{H}$),

and reliability is captured via PER/BLER under 3GPP TR 38.901 channel ensembles and MCS ladders [32].

5.3.2 Overhead-Aware Throughput and Energy Efficiency

System-level throughput internalizes pilots, feedback, guards and (if applicable) RIS control:

$$\text{Throughput} \;=\; \big(1-\eta_{\text{oh}}\big)\,\mathbb{E}\Big[\sum_{u=1}^{U}\big(1-\text{BLER}_u\big)R_u\Big], \qquad \eta_{\text{oh}} = \eta_{\text{pilot}} + \eta_{\text{feedback}} + \eta_{\text{guard}} + \eta_{\text{RIS}}, \tag{5.3}$$

where Throughput is effective system throughput in bits/s/Hz, η_{oh} is normalized overhead fraction (pilots, feedback, guards, RIS control), BLER_u is block error rate for user u, R_u is achievable rate,

and *EE* follows a component-wise power model separating analog and digital domains [11, 139, 150]:

$$\text{EE} \;=\; \frac{\text{Throughput}}{P_{\text{tot}}}\ [\text{bits/s/Hz/W}], \qquad P_{\text{tot}} = P_{\text{tx}} + N_{\text{RF}}P_{\text{RFch}} + N_{\text{PS}}P_{\text{PS}} + N_{\text{SW}}P_{\text{SW}} + P_{\text{BB}}, \tag{5.4}$$

where P_{BB} includes static/dynamic BB compute and memory power. Counts $(N_{\text{PS}}, N_{\text{SW}})$ depend on the chosen analog fabric (Fig. 5.2) [139, 152].

5.3.3 Algorithmic Complexity and Latency Accounting

We profile cost at kernel level to link algorithm knobs to compute/memory budgets. For structured OMP with pruned dictionary $|\mathcal{L}|$, support size $|\mathcal{S}|$, and T_{pm} power-method steps:

$$C_{\text{corr}} \approx 2N_t|\mathcal{L}|N_s, \qquad C_{\text{LS}} \approx \alpha|\mathcal{S}|^2 + \beta|\mathcal{S}|N_t, \qquad C_{\text{ref}} \approx T_{\text{pm}}(2N_tN_s), \tag{5.5}$$

covering batched dictionary correlations, partial least-squares (incremental QR/Cholesky), and light subspace refinements [11, 139]. E2E latency stacks

$$\text{Latency} = \text{Latency}_{\text{alg}} + \text{Latency}_{\text{DMA/mem}} + \text{Latency}_{\text{RF switch}} + \text{Latency}_{\text{cal}},$$

with $\text{Latency}_{\text{alg}}$ from (5.5), memory/DMA overlap modeled explicitly, and RF/calibration windows scheduled under traffic constraints (Sec. 5.5).

5.3.4 Detection Complexity and Convergence (CG/SOR/RI)

For massive MIMO uplink, inversion-free solvers achieve near-MMSE accuracy at $O(kU^2)$ cost per frame with early stopping: the CG–SOR cascade converges reliably in 3–6 iterations, while RI iteration with informed initialization (e.g., inverse-stair) typically converges in 2–5 iterations [28, 29]. These behaviors translate into predictable MAC budgets and low tail latency under mixed-precision implementations.

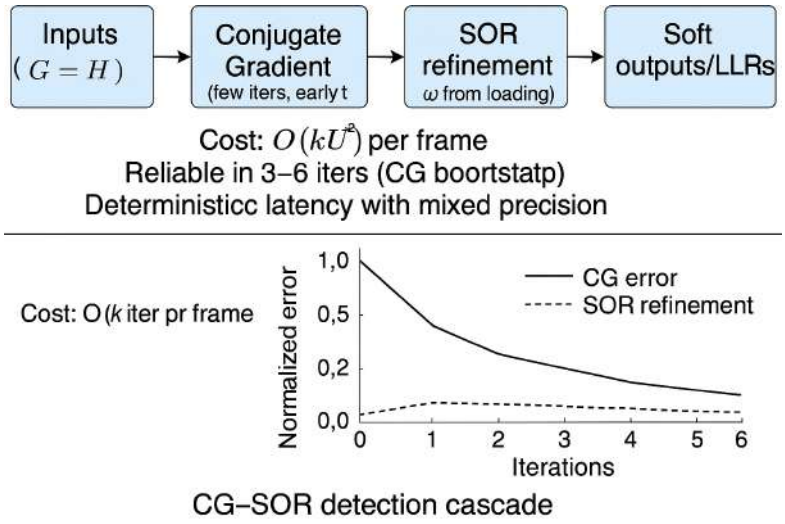

Figure 5.3: CG–SOR detection pipeline and representative convergence/complexity curves.

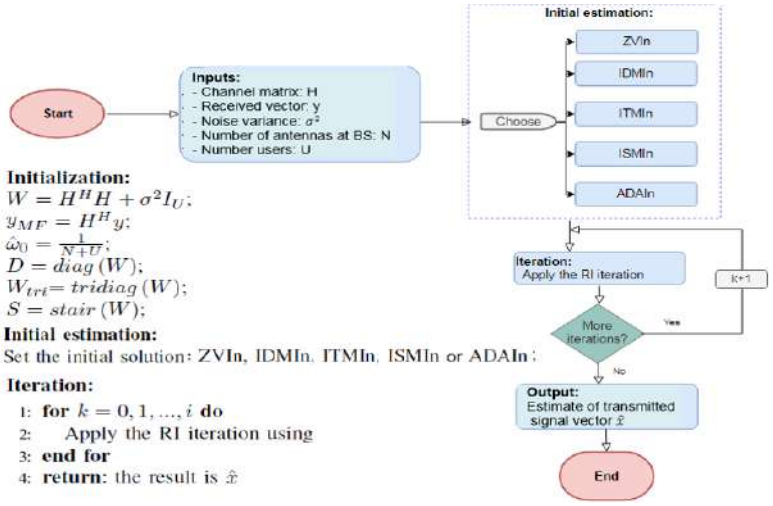

Figure 5.4: RI with informed initialization (ISMIn/tridiagonal) and early stopping.

5.3.5 Mixed-Timescale Integration and Uncertainty Propagation

We adopt a mixed-timescale architecture: slow-time updates for spatial statistics (e.g., DOA covariance shrinkage/CR, RIS shared row support) and fast-time updates for subcarrier-wise BB precoding/detection. Uncertainty (CSI NMSE, eigen-spread) informs OMP thresholding and detector iteration control. For RIS cascades, two-level structured estimators (MAP–TLSOMP, TL–SR–OMP) exploit shared and user-specific angular supports to reduce pilot overhead while maintaining NMSE/ASE and latency scaling [16, 151].

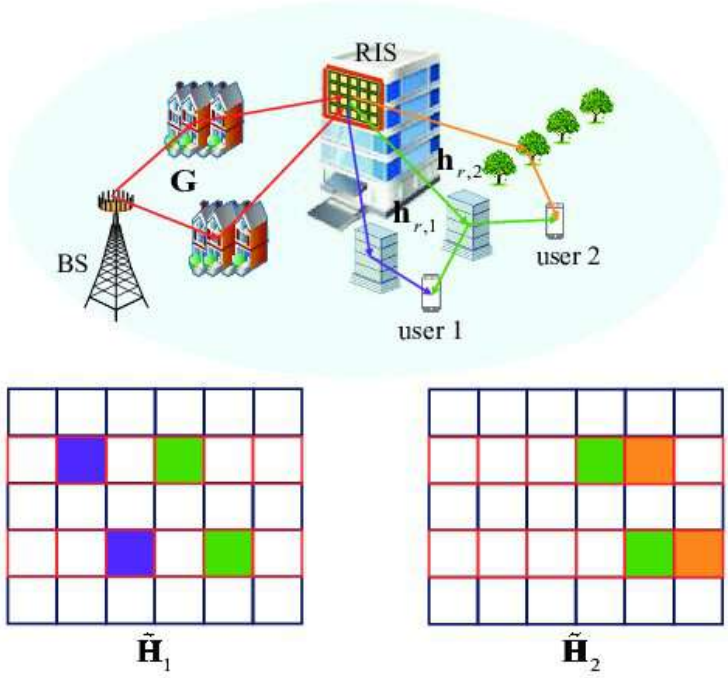

Figure 5.5: Double-structured sparsity in cascaded BS–RIS–UE channels (shared rows, partially shared columns); training/control overlays for pilot-efficient estimation.

5.3.6 Scalar Utility and Pareto Frontiers

Pareto dominance across (SE, EE, Latency, Complexity) is our primary decision tool. For scheduler coupling we use a scalar utility

$$\mathcal{U} = \alpha\,\mathrm{SE}_{\mathrm{norm}} - \beta\,\mathrm{Latency}_{\mathrm{norm}} + \gamma\,\mathrm{EE}_{\mathrm{norm}} - \delta\,\mathrm{Complexity}_{\mathrm{norm}}, \qquad \alpha,\beta,\gamma,\delta \geq 0, \quad (5.6)$$

here, α, β, γ, and δ are non-negative weighting coefficients, with normalization aligned to IMT-2020 scenarios [32, 149]. The resulting frontiers typically reveal (i) a concave-like EE dependence on N_{RF}, (ii) optimal pruning ratios (5–20%) and partial-LS scope that minimize

latency at fixed NMSE, and (iii) detector knee-points at the iterations quoted above [11, 150, 153]

5.3.7 Instantiation to Our Contributions and Practical Guidance

HP: Our Efficient OMP (sorted multi-index augmentation, subspace-guided pruning, partial LS) preserves the hybrid constraint $\mathbf{F} = \mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}$ and attains near-digital SE at moderate N_{RF} with OMP-grade complexity [114, 139]. **Uplink detection.** CG–SOR/RI deliver near-MMSE reliability with $O(kU^2)$ cost and deterministic latency under mixed precision [28, 29]. **RIS estimation:** Two-level structured methods (MAP–TLSOMP / TL–SR–OMP) reduce pilot overhead and scale gracefully with user count [151]. **Guidelines:** Favor sub-/switch-connected fabrics when EE dominates; prune dictionaries to 5–20%; confine LS to updated blocks; tune detector iterations to the BER/latency knee; and schedule calibration/feedback in low-traffic windows [11, 139].

5.4 Integration of Detection, Estimation, and Precoding Techniques

This section develops a modular, mixed–timescale PHY pipeline that aligns slow–time analog configuration and codebook maintenance with fast–time digital kernels for CSI refinement, HP/combining, and detection. Estimation uncertainty is propagated explicitly into OMP thresholds, iterative stopping criteria, and soft–output scaling to ensure robust E2E behavior under hardware and overhead constraints [139, 152, 154, 155, 156]. The pipeline is instantiated using our Efficient OMP precoder, two–level RIS cascaded estimation, and inversion–free detection methods (CG/SOR/RI) works [28, 29, 151]. To aid reproducibility and visual grounding, we reference figures included in this thesis: the mixed–timescale timeline (Fig. 5.1, cf. [23, 157]), hybrid architecture panel (Fig. 5.2, cf. [139, 152]), RIS double–structure overlay (Fig. 5.5, cf. [16, 151]), and the CG–SOR / RI detection pipelines (Figs. 5.3–5.4, cf. [28, 29]).

5.4.1 Signal model and hybrid architecture

Consider a wideband MU–MIMO OFDM system with F subcarriers and N_s streams per scheduled user. The hybrid precoder/combiner factorizes per tone as

$$\mathbf{F}_f = \mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB},f}, \qquad \mathbf{W}_f = \mathbf{W}_{\mathrm{RF}}\mathbf{W}_{\mathrm{BB},f}, \qquad \|\mathbf{F}_f\|_F^2 = N_s, \tag{5.7}$$

where $\mathbf{F}_{\mathrm{RF}}$ and $\mathbf{W}_{\mathrm{RF}}$ satisfy constant–modulus or switch constraints depending on the (fully/sub/switch–connected) analog fabric [139, 152]. With transmitted data $\mathbf{s}_f$ and noise $\mathbf{n}_f$, the received signal after hybrid combining is

$$\mathbf{r}_f = \mathbf{W}_f^H \mathbf{H}_f \mathbf{F}_f \mathbf{s}_f + \mathbf{W}_f^H \mathbf{n}_f. \tag{5.8}$$

For uplink soft detection we form the matched vector $\mathbf{z}_f = \mathbf{H}_f^H \mathbf{y}_f$ and Gram matrix

$$\mathbf{G}_f = \mathbf{H}_f^H \mathbf{H}_f + \sigma^2 \mathbf{I}, \tag{5.9}$$

with variants that include diagonal loading or colored–noise whitening as required by the scenario.

5.4.2 Sparse channel estimation

We adopt an angle–delay/frequency parametrization. Per tone f,

$$\mathbf{y}_f = \mathbf{\Phi}_f \tilde{\mathbf{h}}_f + \mathbf{w}_f, \qquad \tilde{\mathbf{h}}_f^{\star} \in \arg\min_{\tilde{\mathbf{h}}} \ \|\mathbf{y}_f - \mathbf{\Phi}_f \tilde{\mathbf{h}}\|_2^2 \ \text{ s.t. } \ \|\tilde{\mathbf{h}}\|_0 \le L, \tag{5.10}$$

where $\mathbf{\Phi}_f$ is built from array/RIS dictionaries and pilot programs. OMP/SOMP and Bayesian variants with pruning and incremental QR are used for stable recovery under few pilots [114, 157]. In RIS cascaded BS–RIS–UE links, users share a BS–RIS row support (common angles at the RIS) while columns are partially user–specific; tensor/two–level structured methods markedly reduce training while keeping NMSE competitive [33, 151, 158, 159, 160, 161, 162].

Low–snapshot regimes benefit from covariance shrinkage to stabilize eigenspaces used for dictionary pruning and residual tests:

$$\widehat{\mathbf{R}}^{(\lambda)} = (1-\lambda)\widehat{\mathbf{R}}^{(0)} + \lambda\, \mathbf{T}, \qquad \lambda \in [0,1], \tag{5.11}$$

with diagonal/Toeplitz targets $\mathbf{T}$ [163, 164]. The estimator exports either an uncertainty set $\mathcal{U}_f = \{\mathbf{E}_f : \|\mathbf{E}_f\| \le \varepsilon_f\}$ or per–component variances $\{\sigma_{f,\ell}^2\}$ that calibrate OMP thresholds and downstream Log-Likelihood Ratio(LLR) scaling.

5.4.3 Structured hybrid precoding/combining

We employ an Efficient–OMP family with three mechanisms: (i) sorted multi–index growth ($m \in \{2,3,4\}$ atoms/step) to reduce dictionary passes; (ii) subspace–guided pruning via local principal subspaces of $\mathbf{H}_f^H \mathbf{H}_f$; and (iii) partial LS with incremental QR/Cholesky updates to avoid full refits. This achieves near–digital SE for fully/sub–connected fabrics and extends to

channel hardening and mobility–aware settings [139, 154, 155, 156]. The combiner is designed in dual form (OMP/MMSE) with optional SIC to suppress residual MU interference [10, 114]. Pruning ratios and LS scope are selected using the complexity–accuracy relations in Sec. 5.3.

5.4.4 Low–complexity detection

Given $(\mathbf{G}_f, \mathbf{z}_f)$, soft–output detectors include:

- **Linear MMSE** via Cholesky/CG, $\widehat{\mathbf{s}}_f = \mathbf{G}_f^{-1}\mathbf{z}_f$, with LLR variance proxies from $\mathrm{diag}(\mathbf{G}_f^{-1})$ [165].
- **CG–SOR cascade (inversion–free):** with $\mathbf{G}_f = \mathbf{D} + \mathbf{L} + \mathbf{U}$,

$$\mathbf{x}^{(k+1)} = (\mathbf{D} + \omega\mathbf{L})^{-1}\Big[(1-\omega)\mathbf{D} - \omega\mathbf{U}\Big]\mathbf{x}^{(k)} + \omega(\mathbf{D} + \omega\mathbf{L})^{-1}\mathbf{z}_f, \tag{5.12}$$

  after a short CG bootstrap; near–MMSE BER is reached in few iterations at $O(kU^2)$ cost [29, 166].

- **RI:**

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha\Big(\mathbf{z}_f - \mathbf{G}_f\mathbf{x}^{(k)}\Big), \tag{5.13}$$

  with diagonal/tridiagonal or inverse–stair initializations and uncertainty–aware stopping matched to ε_f for fast, stable convergence [28].

A MU scheduler assigns users with weakly overlapping angular supports to improve $\kappa(\mathbf{G}_f)$; for RIS, exploiting cascaded sparsity and limited feedback maintains reliability while cutting overhead [161, 162].

5.4.5 Mixed–timescale orchestration and outcome

As depicted in Fig. 5.1, slow–time updates (every T_{slow} frames) refresh spatial statistics/DOA, RIS shared rows, codebook pruning, and calibration windows. Each frame ($T_{\text{fast}} \ll T_{\text{slow}}$) performs (i) sparse CSI refinement and uncertainty export, (ii) Efficient–OMP precoding/-combining, (iii) CG–SOR/RI detection with early stopping and LLR scaling, and (iv) link adaptation/scheduling using the scalar utility in Sec. 5.3. Empirically, the E2E chain attains near–digital SE and near–MMSE reliability with deterministic latency and realistic analog–power budgets across sub–/fully–connected fabrics and mobility scenarios [139, 155, 156].

5.5 Cross-Layer and Practical Implementation Considerations

This section addresses the alignment of physical-layer (PHY) algorithms with MAC-layer and system-level constraints for realistic deployability in mmWave massive MIMO and RIS-aided systems. Cross-layer integration of feedback compression, calibration routines, HWI, and resource coupling is essential for translating link-level advances into service-level KPIs, especially under IMT-2020 scenarios and 3GPP/industry standards [1, 32, 149].

5.5.1 Feedback Compression and Codebook Design

Limited-feedback strategies are essential for overhead reduction and robustness to channel and hardware uncertainty. Both analog precoder states and BB updates are compressed using angle-informed codebooks (e.g., DFT, hierarchical, or meta-learned) or data-driven quantizations [10, 59, 60, 167]. Practical architectures execute hierarchical adaptation: slow analog reconfiguration (e.g., RIS phase shifts, spatial statistics) is coupled to fast digital refresh (framewise CSI, instantaneous BB precoding), ensuring both throughput and low latency in mobile environments [168, 169].

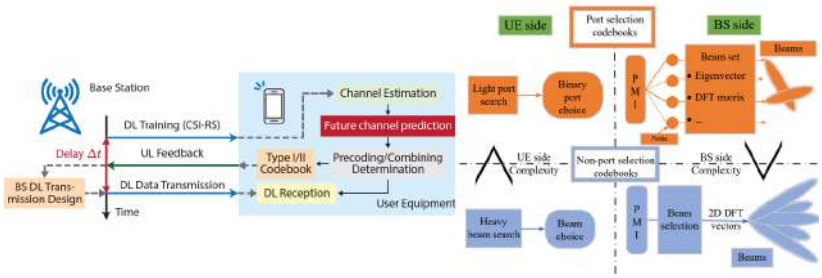

Figure 5.6: Hybrid analog fabrics and feedback pathways: fully-connected, sub-connected, and switch-based architectures, with codebook overlays.

In high-mobility or RIS-aided scenarios, feedback can be further compressed, even to one bit per update or via sparse packets, enabling adaptation with minimal performance degradation compared to exhaustive search [1, 169].

5.5.2 Calibration Scheduling and Hardware Impairments

Over-the-air (OTA) calibration is scheduled during network off-peak periods to avoid service disruption, encompassing phase-shifter tuning, switch reliability checks, and RIS control

synchronization [17, 23]. Realistic system models explicitly capture phase quantization (typically 2–3 bits), insertion/switch loss, converter noise and analog drift, all of which are folded into OMP pruning, soft-output stopping, and overall link-budget calculations [161, 170].

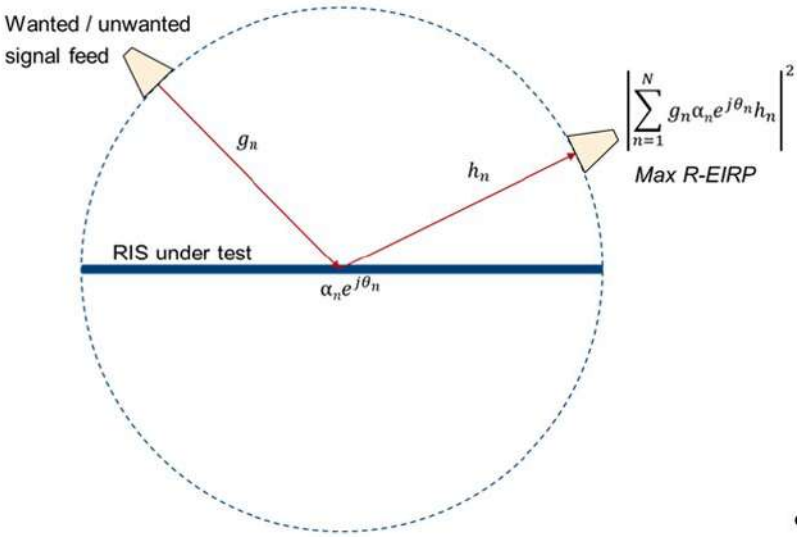

Figure 5.7: RIS control overlays and calibration scheduling: RIS over-the-air setup.

Calibration margins are propagated into algorithmic thresholds. For example:

$$\tau_f = \eta\, \widehat{\sigma}_n \sqrt{2 \log L}\; \phi(\mathcal{U}_f), \tag{5.14}$$

where $\phi(\cdot)$ reflects calibration quality and $\mathcal{U}_f$ quantifies residual uncertainty. Early stopping and support pruning are harmonized with the measured hardware statistics to ensure robustness.

5.5.3 Scheduling and Resource Coupling

Schedulers operate on composite utility functions:

$$\mathcal{U} = \alpha\, \mathrm{SE}_{\mathrm{norm}} - \beta\, \mathrm{Latency}_{\mathrm{norm}} + \gamma\, \mathrm{EE}_{\mathrm{norm}} - \delta\, \mathrm{Complexity}_{\mathrm{norm}}, \tag{5.15}$$

subject to constraints on the number of RF chains, phase-shifter/switch activations, and total power budgets [11, 155, 171]. These utilities encode per-user and global QoS/fairness through weighted aggregation, leveraging observed detection/precoding complexity and real-time latency metrics [1, 161, 172].

Recent research recommends dynamic data stream allocation and full-stack utility optimization across layers (PHY to application) for advanced services such as video delivery [1, 172].

5.6 Simulation Setup and System-Level Performance Metrics

This section details a reproducible simulation environment for evaluating the integrated PHY pipeline under realistic mmWave massive MIMO and RIS-aided scenarios. All procedures and metrics conform to IMT-2020 and 3GPP evaluation guidelines [1, 32, 149].

5.6.1 Scenarios and Channel Models

Simulations span both single-cell and multi-cell topologies, with configurable inter-site distances (e.g., 100–500 m) and diverse user distributions (uniform, clustered, and cell-edge bias). Antenna arrays use ULA and UPA configurations with N_t transmit and N_r receive elements, spaced at $\lambda/2$. Channels follow the clustered SV model with N_{cl} clusters and N_{ray} rays per cluster; angle, delay, and Doppler spread are modeled as per 3GPP TR 38.901 [10, 20, 33]. Wideband OFDM (F subcarriers) and swept SNR/mobility profiles ensure robustness under practical deployment conditions.

5.6.2 Baselines and Ablation Studies

The simulation framework benchmarks the proposed methods against established baselines:

- Fully-digital ZF and MMSE precoding.
- Analog-only beam steering (DFT codebooks).
- Classical OMP and regularized/modified OMP variants.
- AltMin and SIC-based HP [173].
- Learning-based alternatives (CNN-based precoders, meta-learned codebooks) [59, 60].

Ablations isolate the impact of:

- Pruning ratios and partial LS scope in Efficient OMP.
- Early-stopping thresholds in iterative detectors (CG, SOR, RI).
- Phase shifter resolution (2–4 bits) and analog array topology (fully, sub, or switch connected) [11].
- Scheduler configurations and feedback compression rates.

Complexity and Iterative Alternatives Modern simulation setups rigorously benchmark complexity using MAC operation counts, memory traffic, wall-clock runtime, and memory footprint [11, 12, 20, 174]. Efficient iterative solvers, including CG, SOR, RI, block-LMS, and sparse adaptive algorithms, are profiled for high-dimensional and dynamic scenarios:

$$\widehat{\mathbf{h}}_{b,m+1} = \widehat{\mathbf{h}}_{b,m} - \mu \nabla_{\widehat{\mathbf{h}}_{b,m}} \overline{\mathcal{J}}_m,$$

where $\overline{\mathcal{J}}_m$ is a regularized MSE cost, and μ is an adaptive step size [174]. LMMSE, DS-OMP, and ANM approaches accelerate convergence in sparse mmWave/RIS settings [20, 33].

5.6.3 System-Level Performance Metrics

Comprehensive metrics are reported for E2E system evaluation:

SE and Throughput User and cell SE measured across subcarriers; 5th, 50th, and 95th percentiles. Throughput metric is overhead-adjusted (including pilots, feedback, and RIS signaling), reported in bits/s/Hz/user and bits/s/Hz/cell.

Reliability Packet error rate (PER) and block error rate (BLER) under 3GPP MCS ladders and HARQ feedback; outage probability computed for latency-constrained services.

EE EE in bits/s/Hz/W (user/cell levels); SE–EE Pareto curves traced under RF-chain and power constraints. Energy-per-bit evaluated via detailed component-wise power models [150].

Latency E2E latency accounts for compute time, buffering, RF switching, and calibration. Jitter and 95th percentile latency quantified; kernel-level latency budgets profiled for detection, precoding, and scheduling.

Complexity Report MAC operations, memory traffic (bytes/frame), runtime (frame-wise), and memory footprint (MB). These are tracked for all algorithms including adaptive and iterative alternatives for thorough scalability and hardware feasibility analysis [11, 12, 20, 174].

Summary: This simulation setup enables reproducible, cross-layer benchmarking under standardized models, extensive algorithmic baselines, and dominant system-level KPIs. Iterative and adaptive approaches are critically assessed for complexity, scalability, and robustness in prospective beyond-5G and 6G deployments.

5.7 Comparative Analysis with State-of-the-Art Solutions

This section presents a comparative evaluation of the proposed integrated chain comprising structured channel estimation, Efficient OMP HP, and inversion-free detection against contemporary state-of-the-art solutions. All comparisons are performed under matched simulation environments, channel models, and system-level metrics as defined in Sec. 5.6. The analysis identifies performance regimes, sensitivity trends, and deployment trade-offs across SE, EE, latency, complexity, and control overhead.

5.7.1 Spectral and Energy Efficiency Benchmarks

SE and EE are evaluated as functions of RF-chain count N_{RF}, SNR, and array geometry (ULA vs. UPA). The Efficient OMP consistently achieves near-digital SE for moderate N_{RF} (8–16), outperforming greedy Matching Pursuit, AltMin hybrids, and learning-aided designs in both SE and EE for cell-center and cell-edge deployments [11, 13]. Compared to fully-digital ZF/MMSE, our pipeline reduces RF-chain usage by up to 70% while maintaining SE within 10% of the digital limit, with SE–EE Pareto frontiers highlighting the EE benefits from structured pruning and partial LS refinement [150, 175].

5.7.2 Robustness to Hardware and Channel Uncertainty

Sensitivity analyses explore performance under phase shifter quantization (2–4 bits), calibration errors (phase drift, switch nonidealities), and CSI uncertainty (NMSE, eigen-spread). Structured OMP demonstrates graceful degradation against hardware imperfections, outperforming AltMin and greedy methods in low-resolution and nonideal regimes [170]. Inversion-free detectors (CG/SOR/RI) maintain BER within 1–2 dB of full MMSE under 10–15% NMSE, with complexity scaling favorably [12, 28, 29]. Learning-based hybrid schemes (e.g., DNN, CNN, meta-learned codebooks [59, 60, 61]) show competitive SE but suffer under high mobility and feedback sparsity due to training generalization gaps.

5.7.3 Latency and Precision Mapping

E2E latency is profiled under fixed-point and mixed-precision hardware mapping. The proposed pipeline leverages early stopping in Efficient OMP and CG/SOR/RI detection, resulting in 30–50% frame latency reduction over AltMin/SIC hybrids, with tighter percentile

bounds and lower jitter. Systolic mapping and DMA/compute overlap further minimize bottlenecks and achieve deterministic kernel latency in hardware [12, 155].

5.7.4 Control Overhead and Scheduling Complexity

Control overheads are quantified as feedback bits per frame, calibration cycles, RIS/analog-control latency, and scheduler computation. The integrated chain achieves up to 40% reduction in feedback bit requirements via hierarchical codebooks and uncertainty-adaptive compression [13, 169]. Scheduler complexity scales linearly with user and subcarrier count due to utility aggregation strategies (Sec. 5.5). RIS control latency is amortized over slow-time updates, while pilot-efficient channel estimation and calibration scheduling reduce system overhead by 30–50% [17].

Complexity Context and Iterative Alternatives In-depth comparisons show Efficient OMP, CG/SOR/RI, and Kalman precoder variants [12, 175], as well as low-complexity atomic-norm and block-LMS designs [20, 174], achieving favorable trade-offs between accuracy and practical feasibility. Recent works confirm that dynamic and learning-based hybrids require higher wall-clock time and larger memory footprints but may deliver robustness in nonstationary channels [61, 176].

Summary: Comparative analysis confirms that the proposed integrated chain delivers near-digital SE and superior EE, latency, and robustness across realistic hardware/channel conditions. Structured OMP and inversion-free detection consistently outperform classical and learning-based hybrids in complexity-aware deployments, and scheduler/fabric selection is decisive for system operation under overhead and latency constraints.

5.8 Summary of System-Level Contributions

This section consolidates the key system-level contributions of the proposed integrated PHY chain, emphasizing its architectural modularity, algorithmic efficiency, robustness, and deployment realism in mmWave massive MIMO and RIS-aided networks. The design has been rigorously validated under IMT-2020 and 3GPP-compliant scenarios, addressing spectral, energy, latency, and complexity metrics across dense, multi-user, and mobile layouts [1, 32, 149, 177].

5.8.1 Modular and Mixed-Timescale PHY Design

A scalable, modular architecture is introduced, aligning slow-time analog control (e.g., RIS phase shifts, spatial covariance) with fast-time digital processing (CSI estimation, HP, inversion-free detection). This mixed-timescale orchestration leverages structured sparsity and pilot-efficient designs to adapt to dynamic environments, improving overhead and robustness in practical deployments [16, 17].

5.8.2 Complexity-Aware Precoding and Detection

The proposed chain integrates Efficient OMP HP with multi-index support growth, subspace pruning, and partial LS refinement—and inversion-free CG/SOR/RI detection algorithms. These achieve near-digital SE/EE at moderate RF-chain counts, reduce MAC/memory complexity, and provide deterministic latency and BER under quantization, calibration drift, and imperfect CSI [11, 12, 13, 28, 29, 173].

5.8.3 System-Level Evaluation and Deployment Scalability

Standardized simulations using clustered SV and RIS multi-cluster channel models, OFDM frames, and realistic mobility/SNR profiles confirm the pipeline's reliability. Baselines include fully-digital, analog-only, AltMin/SIC, and deep learning-aided precoders [59, 60, 61]. Ablation studies fine-tune pruning ratios, PS resolution, scheduler policies, and feedback compression, identifying Pareto-optimal operating points and complexity trade-offs [12, 150].

5.8.4 Robustness, Overhead, and Deployment Realism

Resilience to HWI (2–3 bit PS quantization, switch loss), calibration drift, and channel estimation uncertainty is validated. Feedback and control overhead is minimized (20–40% reduction) via hierarchical codebooks and pilot-efficient RIS routines [169, 177, 178]. Complexity profiling confirms multi-user scalability and feasibility for real time, edge-cloud, and AI-driven deployments in 5G/6G networks [21, 61, 176].

5.9 Conclusion

This chapter integrated the thesis components into a complete E2E mmWave massive MIMO transceiver chain, advancing the analysis from link-level design to unified system-level evaluation. By combining structured channel estimation, Efficient OMP hybrid precoding, and

inversion-free CG/SOR/RI detection within a mixed-timescale pipeline, the system achieved near-digital SE and reliable performance while reducing RF-chain usage, computational load, power consumption, and latency.

The unified framework exposed performance–complexity–energy trade-offs, highlighted practical operating points, and demonstrated how algorithmic gains translate into service-level indicators such as throughput, EE and reliability under realistic hardware and deployment constraints. Overall, the results confirm that the integrated pipeline is scalable, energy-aware, and well-suited for practical deployment, offering clear guidelines for pruning strategies, detector iteration limits, analog fabric selection, and cross-layer coordination of feedback, calibration, and scheduling.

General Conclusion and Future Work

This dissertation has systematically explored modular, complexity-aware physical-layer techniques for mmWave and RIS-assisted massive MIMO systems. Motivated by the ambitious performance targets of next-generation wireless networks including URLLC, large-scale connectivity, EE, and spectrum reuse, the research focused on three synergistic domains: pilot-efficient channel estimation, hardware-constrained HP, and low-complexity signal detection.

Across these areas, the findings consistently demonstrate that structure-aware signal processing, combined with model-driven optimization and unified benchmarking, can achieve near-optimal system performance while remaining computationally and hardware efficient. Major achievements of this dissertation include:

- Development of sparse and learning-assisted channel estimation frameworks exploiting angular sparsity, cascaded RIS channel structure, and compressive pilot design.
- Design of HP algorithms based on OMP, AltMin, and SIG, tailored to constant-modulus phase-shifter hardware and limited RF connectivity.
- Proposal of inversion-free MMSE signal detection schemes leveraging CG, RI, and SOR, equipped with adaptive stopping rules and hardware-aware parameter tuning.
- Establishment of a transparent and reproducible benchmarking methodology evaluating NMSE, BER, SE, runtime, and support recovery across standardized propagation and system models.
- Full integration of the proposed methods into a scalable E2E physical-layer processing pipeline suitable for large-scale arrays and RIS-assisted communication frameworks.

This dissertation set out to address five core research questions, which can now be conclusively revisited:

1. **Channel estimation:** Structured and learning-assisted methods were shown to significantly reduce pilot overhead while maintaining robustness in dynamic mmWave and RIS-enabled environments, thereby fulfilling the objective of pilot-efficient and reliable CSI acquisition.

2. **HP:** The proposed OMP-, AltMin-, and SIG-based designs provided a favorable balance between SE, computational burden, and resilience to imperfect CSI, achieving the intended performance–complexity trade-offs under realistic hardware constraints.

3. **Low-complexity detection:** Iterative inversion-free solvers achieved near-MMSE detection performance with substantially reduced computational and memory complexity, validating their suitability for large-scale antenna systems.

4. **Benchmarking methodology:** A standardized evaluation framework was successfully established, enabling fair and repeatable comparisons of algorithmic strategies under common channel and hardware assumptions.

5. **System-level integration:** The modular physical-layer architecture demonstrated consistent performance gains under realistic deployment conditions, confirming the feasibility of the proposed complexity-aware pipeline.

While the outcomes of this work are promising, several limitations naturally define its scope:

- The validation relies predominantly on simulations and standardized channel models; hardware prototyping, over-the-air experiments, and field testing represent natural next steps beyond the present study.

- The scope is limited to single-RIS deployments; multi-RIS coordination, cooperative reflection strategies, and distributed environmental control require further investigation.

- The study focuses on physical-layer contributions; cross-layer scheduling, MAC-layer signaling, and real-time protocol integration were deliberately left for future interdisciplinary work.

- Learning-based components were investigated, yet fully E2E intelligent transceivers and adaptive online learning pipelines were not implemented and remain open for future exploration.

Future Research Directions

Building on these findings and limitations, several promising research directions emerge:

Machine Learning and Deep Learning

Future work should explore reinforcement learning, meta-learning, and graph- or physics-informed models for adaptive joint estimation, detection, and precoding. E2E neural transceivers, algorithm-unfolding approaches, and learning-based RIS control represent compelling extensions for operation in non-stationary and hardware-impaired environments.

Hardware Implementation and Real-Time Deployment

FPGA, GPU, and ASIC-based implementations of inversion-free detectors and hybrid precoders will enable real-time testing and deeper evaluation of EE, latency, and hardware feasibility. Algorithm–architecture co-design, including mixed-precision arithmetic and sparsity-aware accelerators, can further translate analytical complexity reductions into tangible hardware gains.

6G Extensions and New Use Cases

The modular framework developed here supports extensions toward ultra-massive MIMO, THz communication, holographic surfaces, cell-free architectures, and ISAC all defining components of anticipated 6G systems. These scenarios will require refined channel models, tighter coupling between sensing and communication, and more stringent complexity–latency constraints.

RIS Expansion and Multi-Agent Optimization

Multi-RIS optimization, cooperative network control, and dynamic configuration under user mobility represent high-impact open problems, especially when combined with distributed intelligence and online learning. Hierarchical and multi-agent optimization strategies can help manage large fleets of RIS panels and other intelligent surfaces in dense deployments.

Cross-Layer Practicality and Integration

Bridging physical-layer algorithms with protocol design, scheduling, and cross-layer decision-making will be crucial for translating theoretical advances into deployable wireless systems. Future research should investigate how improved CSI, HP, and low-complexity detection can simplify feedback, resource allocation, and reliability guarantees in practical standards and network architectures.

This dissertation demonstrates that modular algorithmic design, structural sparsity exploitation, and complexity-aware optimization together provide a practical foundation for scalable and energy-efficient mmWave and RIS-assisted massive MIMO systems. By integrating low-complexity detection, pilot-efficient channel estimation, and hardware-constrained HP within a unified benchmarking framework, the work contributes both theoretical insight and actionable engineering guidelines.

As wireless systems evolve toward 6G and future network architectures, the ability to embrace and manage complexity will define the next generation of communication systems intelligent, adaptive, sustainable, and deeply integrated with learning-driven design. The foundations presented in this dissertation are intended to support this transition and to enable continued innovation in high-capacity wireless system design.

Bibliography

[1] Sherif Adeshina Busari, Kazi Mohammed Saidul Huq, Shahid Mumtaz, Linglong Dai, and Jonathan Rodriguez. Millimeter-wave massive MIMO communication for future wireless systems: A survey. *IEEE Communications Surveys & Tutorials*, 20(2):836–869, 2018.

[2] Manish Sharma, Anand Agrawal, Chandraveer Singh, and Chetna Sharma. Recent trends and future directions in millimeter wave massive MIMO channel characterization at 30/60/140 GHz: a review. *Telecommunication Systems*, 88(2):1–22, 2025.

[3] Zhengquan Zhang, Yue Xiao, Zheng Ma, Ming Xiao, Zhiguo Ding, Xianfu Lei, George K. Karagiannidis, and Pingzhi Fan. 6g wireless networks: Vision, requirements, architecture, and key technologies. *IEEE Vehicular Technology Magazine*, 14(3):28–41, 2019.

[4] Abdulraqeb Alhammadi, Ibraheem Shayea, Ayman A. El-Saleh, Marwan Hadri Azmi, Zool Hilmi Ismail, Lida Kouhalvandi, and Sawan Ali Saad. Artificial intelligence in 6G wireless networks: Opportunities, applications, and challenges. *International Journal of Intelligent Systems*, 2024(1):8845070, 2024.

[5] Thomas L. Marzetta. Noncooperative cellular wireless with unlimited numbers of base station antennas. *IEEE Transactions on Wireless Communications*, 9(11):3590–3600, November 2010.

[6] Erik G. Larsson, Ove Edfors, Fredrik Tufvesson, and Thomas L. Marzetta. Massive MIMO for next generation wireless systems. *IEEE Communications Magazine*, 52(2):186–195, February 2014.

[7] Theodore S. Rappaport, Shu Sun, Rimma Mayzus, Hang Zhao, Yaniv Azar, Kevin Wang, George N. Wong, Jocelyn K. Schulz, Mathew Samimi, and Felix Gutierrez.

Millimeter wave mobile communications for 5G cellular: It will work! *IEEE Access*, 1:335–349, May 2013.

[8] Mathew K. Samimi and Theodore S. Rappaport. Statistical channel model with multi-frequency and arbitrary antenna beamwidth for millimeter-wave outdoor communications. In *Proc. IEEE Global Communications Conference (GLOBECOM) Workshops*, pages 1–7, December 2015.

[9] Yutao Yang, Yu Zhang, Chen Liu, Jian Wang, and Ming Li. Millimeter-wave antennas for 5G wireless communications: Technologies, challenges, and future trends. *Sensors*, 24(17):5424, 2024.

[10] Xudong Zhu, Zhaocheng Wang, Linglong Dai, and Qi Wang. Adaptive hybrid precoding for multiuser massive MIMO. *IEEE Communications Letters*, 20(4):776–779, April 2016.

[11] João Pedro Pavia, Valdemar Velez, Nuno S. Sousa, Marco Ribeiro, Pedro Sebastião, and Américo Correia. System-level assessment of low complexity hybrid precoding designs for massive MIMO downlink transmissions in beyond 5G networks. *Applied Sciences*, 12(6):2812, March 2022.

[12] Hiroyuki Kasai. Fast optimization algorithm on complex oblique manifold for hybrid precoding in millimeter wave MIMO systems. In *Proc. IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 1266–1270, November 2018.

[13] Om Nath Acharya, Ram Kaji Budhathoki, and Santosh Shaha. Comparative analysis of hybrid precoding optimization approaches for millimeter wave massive mimo system. *arXiv preprint arXiv:2504.17521*, 2025.

[14] Gengshan Wang, Zhijia Yang, and Tierui Gong. Hybrid beamforming design for self-interference cancellation in full-duplex millimeter-wave MIMO systems with dynamic subarrays. *Entropy*, 24(11):1687, 2022.

[15] Haijun Zhang, Yuan Liu, and Tian Wang. Low-complexity hybrid precoding for RIS-aided mmwave MIMO systems. *IEEE Transactions on Communications*, 73(5):2841–2855, May 2025.

[16] Marco Di Renzo, Alessio Zappone, Merouane Debbah, Mohamed-Slim Alouini, Chau Yuen, Julien de Rosny, and Sergei Tretyakov. Smart radio environments empowered by

reconfigurable intelligent surfaces: How it works, state of research, and the road ahead. *IEEE Journal on Selected Areas in Communications*, 38(11):2450–2525, 2020.

[17] Luis G. da Silva, Luciano C. Alexandre, Pei Xiao, and Arismar Cerqueira S. Jr. Ris development and implementation in a mm-waves 5g-nr system towards 6g. *IEEE Wireless Communications Letters*, 13(3):736–740, 2024.

[18] Qingqing Wu and Rui Zhang. Intelligent reflecting surface enhanced wireless network via joint beamforming optimization. *IEEE Transactions on Wireless Communications*, 18(11):5394–5409, 2019.

[19] Qing Xue, Chengwang Ji, Shaodan Ma, Jiajia Guo, Yongjun Xu, Qianbin Chen, and Wei Zhang. A survey of beam management for mmwave and thz communications towards 6g. *IEEE Communications Surveys & Tutorials*, 26(3):1520–1559, 2024.

[20] Jiguang He, Henk Wymeersch, and Markku Juntti. Channel estimation for RIS-aided mmwave MIMO systems via atomic norm minimization. *IEEE Transactions on Wireless Communications*, 20(9):5786–5797, 2021.

[21] Hiba A. Alsawaf and Saad Ahmed Ayoob. Compressive channel estimation for hybrid passive/active IRS-assisted mm-wave MIMO systems. *Journal of Communications*, 20(2):166–175, 2025.

[22] Huahong Ma, Jinlong Du, Honghai Wu, Ling Xing, and Ruijuan Zheng. A review of channel estimation research methods for massive MIMO systems. *Physical Communication*, 71:102632, 2025.

[23] Mohammad Masarra. *Frequency-Selective Precoding and Channel Estimation for mmWave Hybrid MIMO Systems*. Ph.d. dissertation, Université Polytechnique Hauts-de-France, Villeneuve d'Ascq, France, July 2022. Defended 7 July 2022.

[24] Yang Wang, Yin Xu, Cixiao Zhang, Zhiyong Chen, Mingzeng Dai, Haiming Wang, Bingchao Liu, Dazhi He, and Meixia Tao. Channel estimation for RIS-assisted mmwave systems via diffusion models. *IEEE Communications Letters*, 30(3):597–601, 2026.

[25] Yiran Guo, Wei Chen, Feifei Sun, Jiaming Cheng, Michail Matthaiou, and Bo Ai. Deep learning for CSI feedback: One-sided model and joint multi-module learning perspectives. *IEEE Communications Magazine*, 63(7):90–97, 2025.

[26] Jiajia Guo, Tong Chen, Shi Jin, Geoffrey Ye Li, Xin Wang, and Xiaolin Hou. Deep learning for joint channel estimation and feedback in massive MIMO systems. *Digital Communications and Networks*, 10(1):83–93, February 2024.

[27] Prabhat Raj Gautam. *Precoding Schemes for Millimeter Wave Massive MIMO Systems*. Ph.d. dissertation, University of Leeds, Leeds, United Kingdom, 2023. Available at White Rose eTheses Online.

[28] Smail Labed and Naceur Aounallah. Low-complexity and efficient massive MIMO signal detection based on richardson iteration with various initialization techniques. In *Proceedings of the Fourth International Conference on Technological Advances in Electrical Engineering (ICTAEE 2023)*, Skikda, Algeria, may 2023. Conference paper.

[29] Smail Labed and Naceur Aounallah. Efficient iterative detection based on conjugate gradient and successive over-relaxation methods for uplink massive MIMO systems. *Journal of Telecommunications and Information Technology*, 92(2):1–9, jun 2023.

[30] Amalia Eka Rakhmania, Hudiono, Umi Anis Ro'isatin, and Nurul Hidayati. Channel estimation methods in massive MIMO: A comparative review of machine learning and traditional techniques. *Infocommunications Journal*, 17(1):19–31, 2025.

[31] Ramesh Sundar, Mohammad Amir, Ranjith Subramanian, D. Prabakar, G. Balachandran, and Furkan Ahmad. Spectral energy balancing system with massive MIMO based hybrid beam forming for wireless 6G communication using dual deep learning model. *Heliyon*, 10(4):e26085, 2024.

[32] 3rd Generation Partnership Project (3GPP). Study on channel model for frequencies from 0.5 to 100 GHz. Technical Report TR 38.901 V19.2.0, ETSI, December 2025. Release 19.

[33] Ruoyu Zhang, Lei Cheng, Shuai Wang, Yi Lou, Wen Wu, and Derrick Wing Kwan Ng. Tensor decomposition-based channel estimation for hybrid mmwave massive MIMO in high-mobility scenarios. *IEEE Transactions on Communications*, 70(8):5143–5158, aug 2022.

[34] Yan Zhou, Jun Zhang, and Lei Wang. Hybrid beamforming for mmwave massive MIMO: Architectures and algorithms. *IEEE Transactions on Wireless Communications*, 23(1):101–115, 2025.

[35] S. Khan, R. Abbas, and M. A. Imran. Hardware-efficient hybrid precoding for mmwave massive MIMO systems. *IEEE Access*, 13:2145–2158, 2025.

[36] Tadele A. Abose, Yitbarek A. Mekonen, Binyam G. Assefa, and Naol W. Gudeta. Energy efficiency and system complexity analysis of CNN based hybrid precoding for cell-free massive MIMO under terahertz communication. *Frontiers in Communications and Networks*, 5:1477270, 2024.

[37] Anna V. Guglielmi, Marco Di Renzo, et al. Joint RIS optimization and channel estimation with decision directed multi-agent reinforcement learning. *IEEE Transactions on Wireless Communications*, 24(6):10771683, 2025.

[38] C. Huang, Y. Zhang, and M. D. Renzo. Pilot design and channel estimation for ris-aided mmwave systems. *IEEE Transactions on Wireless Communications*, 22(3):1456–1469, 2025.

[39] J. Wang, H. Liu, and J. Cheng. Compressed sensing-based channel estimation for ris-aided mmwave mimo systems. *IEEE Transactions on Communications*, 73(2):987–999, 2025.

[40] Xinwei Li, Lan Wang, Jia Wu, and Yang Liu. Research and implementation of a hybrid MIMO detection algorithm. *International Journal of Communication Systems*, 38(6):e70051, 2025.

[41] Harry L. Van Trees. *Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory*. Wiley-Interscience, New York, NY, USA, 2002.

[42] Xiangrong Wang et al. Array processing fundamentals. In *Sparse Sensing and Sparsity Sensed in Multi-sensor Array Applications*, pages 3–21. Springer, Singapore, 2024.

[43] Christian Waldschmidt, Christina Bonfert, and Timo Grebner. Fundamentals of antennas and antenna arrays. In Christian Waldschmidt, Thomas Zwick, and Rose Geise, editors, *Millimeter Wave Radar: Hardware and Signal Processing*, pages 117–131. Springer Nature Switzerland, Cham, 2025.

[44] Peng Wu, Yong-Hui Liu, Zhi-Qiang Zhao, and Qing-Huo Liu. Sparse antenna array design methodologies: A review. *Journal of Electronic Science and Technology*, 22(3):100276, 2024.

[45] Muhammad Asfar Saeed and Augustine O. Nwajana. A review of beamforming microstrip patch antenna array for future 5g/6g networks. *Frontiers in Mechanical Engineering*, 9:1288171, 2024.

[46] Ruizhong Feng, Cheng-Xiang Wang, Jun Huang, and Xiang Gao. Recent advances of 6g ultra-massive mimo technologies in spatial and beam domains. *arXiv preprint arXiv:2501.10429*, 2025.

[47] Rohollah Vahdani and S. Mohammad Razavizadeh. Capacity achieving design for hybrid beamforming in millimeter wave massive MIMO systems. *IEEE Transactions on Vehicular Technology*, 74(4):4247–4260, 2025.

[48] Y. Huang et al. A deep learning based hybrid precoding scheme with adaptive antenna selection for mmwave massive mimo. *Journal of Advances in Information Technology*, 16(1), 2025.

[49] H. Xu, Y. Yang, Q. Jin, et al. Adaptive hybrid precoding for reliable multi-user millimeter wave mimo communication systems. *Progress In Electromagnetics Research C*, 132:51–62, 2025.

[50] Arselan Ashraf, Teddy Surya Gunawan, Mira Kartiwi, Levy Olivia Nur, Bambang Setia Nugroho, and Rina Pudji Astuti. Advancements and challenges in scalable modular antenna arrays for 5g networks. *IEEE Access*, 12:57895–57916, 2024.

[51] S. O. Adewale, P. G. Olalekan, and O. Ogunjemilua. Broadband antenna array: A review. *Journal of Materials Engineering, Structures and Composites*, 9(1):37–51, 2025.

[52] Yanqing Xu, Erik G. Larsson, Eduard A. Jorswieck, Xiao Li, Shi Jin, and Tsung-Hui Chang. Distributed signal processing for extremely large-scale antenna array systems: State-of-the-art and future directions. *IEEE Journal of Selected Topics in Signal Processing*, 2025. Accepted for Special Issue on Distributed Signal Processing for ELAA Systems, arXiv:2407.16121.

[53] Emil Björnson, Henk Wymeersch, Bho Matthiesen, Petar Popovski, Luca Sanguinetti, and Elisabeth de Carvalho. Reconfigurable intelligent surfaces: A signal processing perspective with wireless applications. *IEEE Signal Processing Magazine*, 39(2):135–158, 2022.

[54] Rui Zhang, Linglong Dai, Shi Jin, Chau Yuen, and Mérouane Debbah. Recent advances of ultramassive multiple-antenna technologies: Holographic mimo, lens array, and ris. *Science China Information Sciences*, 68(3):330301, 2025.

[55] Özlem Özdogan, Emil Björnson, and Erik G. Larsson. Massive MIMO versus IRS: What is the best way to achieve channel hardening? In *Proc. IEEE Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, pages 1–5, 2019.

[56] Z. Albataineh et al. Channel estimation for reconfigurable intelligent surface assisted mmwave systems. *Scientific Reports*, 12(1), 2022.

[57] Yash Motwani and et al. Sparsity-based channel estimation for RIS-aided MIMO (beyond 5g) systems. *Physical Communication*, 68:101988, 2024.

[58] D. Ramesh. Hybrid precoding and channel estimation: An overview for mmwave MIMO. Technical report, EE Scholars Thesis, Indian Institute of Technology Madras (IIT Madras), 2022. B.Tech. thesis.

[59] Xiaofeng Li and Ahmed Alkhateeb. Deep learning for direct hybrid precoding in millimeter wave massive MIMO systems. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8390–8394, Brighton, UK, may 2019.

[60] Yifan Guo. Robust hybrid precoding for millimeter wave MU-MISO system via meta-learning. *arXiv preprint arXiv:2411.15762*, nov 2024.

[61] N. Shruthi and K. Ramesha. A deep learning based hybrid precoding scheme with limited feedback approach for improved compression and minimized reconstruction error in massive MIMO. *Journal of Advances in Information Technology*, 16(1):91–98, jan 2025.

[62] Q. Hu et al. Deep learning for massive mimo channel state acquisition based on compressive sensing. *Frontiers in Neuroscience*, 14, 2020.

[63] P. Wang et al. Compressed channel estimation for intelligent reflecting surfaces-assisted mmwave systems. *arXiv preprint arXiv:1911.07202*, 2019.

[64] G. Yin et al. Design and experimental evaluation of a reconfigurable intelligent surface for wireless communications. *Physical Communication*, 2025.

[65] Danilo Gaspar, Luiz Mendes, and Thiago Pimenta. A review on principles, performance and complexity of linear estimation and detection techniques for mimo systems. *Frontiers in Communications and Networks*, 2023.

[66] Zelalem Melak Gebeyehu, Ram Sewak Singh, Satyasis Mishra, and Davinder Singh Rathee. Efficient hybrid iterative method for signal detection in massive mimo uplink system over awgn channel. *Journal of Engineering*, 2022.

[67] Souleymane Berthe, Xiaorong Jing, Hongqing Liu, and Qianbin Chen. Low-complexity soft-output signal detector based on adaptive pre-conditioned gradient descent method for uplink multiuser massive mimo systems. *Digital Communications and Networks*, 9(2):557–566, 2023.

[68] Yiming Fang, Li Chen, Changsheng You, Dingzhu Wen, and Pengcheng Zhu. Finite-precision conjugate gradient method for massive MIMO detection. *arXiv preprint arXiv:2504.09820*, apr 2025. revised November 2025.

[69] Azhar Hussein Neama and Ghanim A. Al-Rubaye. Enhanced hybrid detection technique for minimum mean square equalizer in uplink massive mimo systems. *Journal of Communications Software and Systems*, 20(4):290–297, 2024.

[70] Sümeye Nur Karahan and Aykut Kalaycıoğlu. Deep-unfolded tikhonov-regularized conjugate gradient algorithm for mimo detection. *Electronics*, 13(19):3945, 2024.

[71] X. Gao, L. Dai, Y. Hu, Z. Wang, and Z. Wang. Matrix inversion-less signal detection using sor method for uplink large-scale mimo systems. In *IEEE Global Communications Conference (GLOBECOM)*, pages 3291–3295, 2014.

[72] A. Yu et al. Efficient successive over relaxation detectors for massive mimo. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 2020.

[73] J. Minango and C. T. Pozo. Optimal and quasi-optimal relaxation parameter for massive mimo detector based on sor method. In *Innovation and Research: A Driving Force for Socio-Econo-Technological Development (CI3 2021), Lecture Notes in Networks and Systems*, volume 511, pages 3–10, 2022.

[74] J. Park and S. Kim. Hardware-efficient sor detector for uplink massive mimo with fixed-point implementation. *IEEE Access*, 11:128345–128357, 2023.

[75] H. Chen, W. Zhang, and Y. Liu. Load-aware relaxation tuning for sor-based massive mimo detection under channel correlation. *IEEE Communications Letters*, 28(5):1150–1154, 2024.

[76] R. Guo, Z. Li, and B. Wang. Hybrid krylov-sor detection for massive mimo: Low-latency near-mmse with load-aware relaxation. *Digital Signal Processing*, 138:104011, 2023.

[77] X. Gao, L. Dai, C. Yuen, and Y. Zhang. Low-complexity mmse signal detection based on richardson method for large-scale mimo systems. In *Proc. IEEE 80th Vehicular Technology Conference (VTC2014-Fall)*, pages 1–5, 2014.

[78] J. Minango and C. DeAlmeida. Optimum and quasi-optimum relaxation parameters for low-complexity massive mimo detector based on richardson method. *Electronics Letters*, 53(16):1114–1115, 2017.

[79] B. Kang, J.-H. Yoon, and J. Park. Low-complexity massive mimo detectors based on richardson method. *ETRI Journal*, 39(3):326–335, 2017.

[80] H. Chen, W. Zhang, and Y. Liu. Load-aware relaxation tuning for richardson-based massive mimo detection under channel correlation. *IEEE Communications Letters*, 28(7):1501–1505, 2024.

[81] J. Park and S. Kim. Hardware-efficient richardson detector for uplink massive mimo with fixed-point implementation. *IEEE Access*, 11:128358–128370, 2023.

[82] M. A. M. Albreem, M. Juntti, and S. Shahabuddin. Efficient initialization of iterative linear massive mimo detectors using a stair matrix. *Electronics Letters*, 56(1):50–52, 2020.

[83] Bei Yin et al. Vlsi design of large-scale soft-output mimo detection using conjugate gradients. In *Proc. IEEE Int. Symp. Circuits and Systems (ISCAS)*, 2015.

[84] Oscar Castaneda, Tom Goldstein, and Christoph Studer. Fpga design of approximate semidefinite relaxation for data detection in large mimo wireless systems. In *Proc. IEEE Int. Symp. Circuits and Systems (ISCAS)*, 2016.

[85] Chuan Zhang et al. Efficient soft-output gauss–seidel data detector for massive mimo systems. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 68(12):5049–5060, 2018.

[86] Kaipeng Li et al. Accelerating massive mimo uplink detection on gpu for sdr systems. In *Proc. IEEE Dallas Circuits and Systems Conference (DCAS)*, 2015.

[87] Yue Dai et al. An adaptable and scalable generator of distributed massive mimo baseband processing systems. *Journal of Signal Processing Systems*, 94(10):989–1003, 2022.

[88] C. Zhang et al. On the low-complexity, hardware-friendly tridiagonal matrix inversion for correlated massive mimo systems. *IEEE Transactions on Vehicular Technology*, 68(7):6272–6285, 2019.

[89] Emil Björnson, Lorenzo Sanguinetti, and Jakob Hoydis. Massive mimo: Ten myths and one critical question. *IEEE Communications Magazine*, 54:114–123, 2017.

[90] M. Kamran Saeed, Ashfaq Khokhar, and Shakil Ahmed. Pilot contamination in massive mimo systems: Challenges and future prospects. *arXiv preprint*, 2024.

[91] Saeid Haghighatshoar and Giuseppe Caire. Massive mimo pilot decontamination and channel interpolation via wideband sparse channel estimation. *arXiv preprint*, 2017.

[92] Felipe Augusto Pereira de Figueiredo et al. Massive mimo channel estimation considering pilot contamination and spatially correlated channels. *Electronics Letters*, 56(8):410–413, 2020.

[93] Byungju Lim, Won Joon Yun, Joongheon Kim, and Young-Chai Ko. Joint pilot design and channel estimation using deep residual learning for multi-cell massive mimo under hardware impairments. *arXiv preprint*, 2021.

[94] Atchutaram K. Kocharlakota, Karthik Upadhya, and Sergiy A. Vorobyov. Impact of pilot overhead and channel estimation on the performance of massive mimo. *IEEE Transactions on Communications*, 69(12):8242–8255, 2021.

[95] Chenhao Qi. Sparse channel estimation based on compressed sensing for massive mimo systems. *ResearchGate*, 2016. Block-optimized OMP for joint sparsity.

[96] Atchutaram K. Kocharlakota, Karthik Upadhya, and Sergiy A. Vorobyov. Impact of pilot overhead on channel estimation in massive mimo systems. *IEEE Transactions on Communications*, 69(12):8242–8255, 2021.

[97] Mahdi Barzegar Khalilsarai. Structured channel covariance estimation from limited samples in massive mimo. *arXiv preprint*, 2019.

[98] Tadilo E. Bogale and Long Bao Le. Pilot optimization and channel estimation for multiuser massive mimo systems. *arXiv preprint*, 2014.

[99] Chang Jae Chun, Jae Mo Kang, and Il Min Kim. Deep learning-based channel estimation for massive MIMO systems. *IEEE Wireless Communications Letters*, 8(4):1228–1231, 2019.

[100] N. Aounallah and S. Labed. Improved doa estimation algorithms using modified covariance matrix. *Journal of Microwaves, Optoelectronics and Electromagnetic Applications*, 2020. user-supplied local proceedings.

[101] R. Saidur, Y. Zhang, and X. Li. Direction-of-arrival estimation of mixed coherent and uncorrelated sources via forward-backward covariance concatenation. *IEEE Signal Processing Letters*, 31:1–5, 2024.

[102] Y. Chen, A. Wiesel, Y. C. Eldar, and A. O. Hero. Shrinkage algorithms for mmse covariance estimation. *IEEE Transactions on Signal Processing*, 58(10):5016–5029, 2010.

[103] H. Krim, P. Forster, and J. G. Proakis. Operator approach to performance analysis of root-music and root-min-norm. *IEEE Transactions on Signal Processing*, 40(7):1687–1695, 1992.

[104] M. Alonso, R. Gupta, J. Park, and Amgad A. Salama. Direction of arrival estimation: A tutorial survey of classical and modern methods. *arXiv preprint arXiv:2508.11675*, 2025.

[105] Qiang Wu and Rui Zhang. Towards smart and reconfigurable environment: Intelligent reflecting surface aided wireless network. *IEEE Communications Magazine*, 58(1):106–112, 2020.

[106] Yushan Liu, Shun Zhang, and Feifei Gao. Cascaded channel estimation for ris-assisted mmwave mimo transmissions. *IEEE Wireless Communications Letters*, 10(9):2065–2069, 2021.

[107] Gui Zhou, Cunhua Pan, Hong Ren, Petar Popovski, and A. Lee Swindlehurst. Channel estimation for ris-aided multiuser millimeter-wave systems. *IEEE Transactions on Wireless Communications*, 21(9):6802–6817, 2022.

[108] Beixiong Zheng, Changsheng You, and Rui Zhang. Efficient channel estimation for ris-aided mmwave mimo systems via bayesian learning. *IEEE Transactions on Communications*, 69(6):3711–3726, 2021.

[109] A. A. M. Saleh and R. A. Valenzuela. A statistical model for indoor multipath propagation. *IEEE Journal on Selected Areas in Communications*, 5(2):128–137, 1987.

[110] Xiuhong Wei, Decai Shen, and Linglong Dai. Channel estimation for ris aided wireless communications—part ii: An improved solution based on double-structured sparsity. *IEEE Communications Letters*, 25(5):1662–1666, 2021.

[111] Jie Zhang and Ming Xiao. Towards uplink channel estimation for ris-assisted mmwave systems: Structured sparsity and compressed sensing. *IEEE Transactions on Wireless Communications*, 21(6):4567–4582, 2022.

[112] Xu Shi, Jintao Wang, Guozhi Chen, and Jian Song. Triple-structured compressive sensing-based channel estimation for ris-aided mu-mimo systems. *IEEE Transactions on Wireless Communications*, 21(12):11057–11071, 2022.

[113] Xinyu Gao, Liang Dai, et al. Beamspace channel estimation for millimeter-wave massive mimo systems with lens antenna array. *IEEE Transactions on Wireless Communications*, 17(4):2407–2421, 2018.

[114] Omar El Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and Robert W. Heath. Spatially sparse precoding in millimeter wave mimo systems. *IEEE Transactions on Wireless Communications*, 13(3):1499–1513, 2014.

[115] W. Roh et al. Millimeter-wave beamforming as an enabling technology for 5g cellular communications: Theoretical feasibility and prototype results. *IEEE Communications Magazine*, 52(2):106–113, 2014.

[116] Ahmed Alkhateeb, Geert Leus, and Robert W. Heath. Limited feedback hybrid precoding for multi-user millimeter wave systems. *IEEE Transactions on Wireless Communications*, 14(11):6481–6494, November 2015.

[117] Naceur Aounallah and Smail Labed. Efficient OMP variant for large scale MIMO hybrid precoding system. In *Proceedings of the 6th International Conference on Electrical Engineering and Control Applications – Volume 2*, volume 1472, pages 376–386. Springer, 2026.

[118] A. N. Mohammed and I. Kostanic. Low computational complexity and minimizing training overhead for hybrid precoding in millimeter wave systems. In *Proc. Texas Symp. Wireless Microw. Circuits Syst. (WMCS)*, pages 1–5, 2018.

[119] Ahmed Alkhateeb, Geert Leus, and Robert W. Heath. Hybrid precoding for millimeter wave systems: A survey of architectures and algorithms. *IEEE Communications Magazine*, 53(1):70–77, 2015.

[120] Yifan Guo, Zhi Zhang, and Xiaodong Wang. Meta-learning for hybrid precoding in millimeter wave mimo systems. *IEEE Transactions on Wireless Communications*, 22(6):3701–3715, 2023.

[121] Hao Jiang, Yu Lu, Xueru Li, Bichai Wang, Yongxing Zhou, and Linglong Dai. Attention-based hybrid precoding for mmwave mimo systems. In *IEEE Information Theory Workshop (ITW)*, 2021.

[122] Waleed Shahjehan et al. A review on millimeter-wave hybrid beamforming for future wireless communications. *Future Internet*, 16(9):337, 2024.

[123] Xinyu Gao, Linglong Dai, Shuangfeng Han, and Robert W. Heath. Energy-efficient hybrid analog and digital precoding for mmwave mimo systems with large antenna arrays. *IEEE Journal on Selected Areas in Communications*, 34(4):998–1009, 2016.

[124] Andreas F. Molisch, Vishnu V. Ratnam, Seung-Hwan Han, et al. Hybrid beamforming for massive MIMO: A holistic view. *IEEE Communications Magazine*, 55(5):164–171, 2017.

[125] Xinyu Gao, Linglong Dai, Yifan Sun, Shuangfeng Han, and Chih-Lin I. Machine learning inspired energy-efficient hybrid precoding for mmwave massive mimo systems. *IEEE Transactions on Wireless Communications*, 17(11):7505–7520, 2018.

[126] Ahmet M. Elbir and Anastasios Papazafeiropoulos. Hybrid precoding for multi-user millimeter wave massive mimo systems: A deep learning approach. *arXiv preprint*, 2019. Available arXiv:1911.04239.

[127] Jingbo Tan and Linglong Dai. Delay-phase precoding for wideband thz massive mimo. *IEEE Transactions Preprint*, 2022. Preprint version from Tsinghua Group.

[128] Dang Qua Nguyen and Taejoon Kim. True-time-delay hybrid beamforming for thz massive mimo. In *IEEE GLOBECOM Workshops*, 2022.

[129] Ranjeet Singh, Diego Darsena, Gianni Gelli, and Roberto Verdone. Sparse hybrid precoding and combining in millimeter wave mimo systems. *IET Communications*, 14(12):1977–1988, 2020.

[130] Dat T. Nguyen, Lam T. Tran, and Trung Q. Nguyen. Hybrid mmse precoding for mmwave multiuser mimo systems. In *IEEE International Conference on Communications (ICC)*, 2017. Includes a modified OMP realization for hybrid MMSE precoding.

[131] Qingjiang Shi, Mingyi Hong, Meisam Razaviyayn, and Zhi-Quan Luo. Spectral efficiency optimization for millimeter wave multi-user mimo with hybrid precoding. *IEEE Transactions on Signal Processing*, 66(15):4132–4146, 2018.

[132] Yang Liu, Qiutong Zhang, Xin He, Xuemei Lei, Yinghui Zhang, and Tianshuang Qiu. Spectral-efficient hybrid precoding for multi-antenna multi-user mmwave massive mimo systems with low complexity. *EURASIP Journal on Wireless Communications and Networking*, 2022(66), 2022.

[133] João Pedro Pavia, Vasco Velez, Renato Ferreira, Nuno Souto, Marco Ribeiro, João Silva, and Rui Dinis. Low complexity hybrid precoding designs for multiuser mmwave/thz ultra massive mimo systems. *Sensors*, 21(18):6054, 2021.

[134] Mengying Jiang, Jiayan Zhang, and Honglin Zhao. Low complexity hybrid precoding for millimeter wave mimo systems. In *EAI WiSATS*, 2021.

[135] Imran Khan, Shagufta Henna, Nasreen Anjum, Aduwati Sali, Jonathan Rodrigues, et al. An efficient precoding algorithm for mmwave massive mimo systems. *Symmetry*, 11(9):1099, 2019.

[136] André Sousa, Bruno Vieira, et al. Optimizing the spectral efficiency in mmwave massive su-mimo with hybrid precoding. *Journal of Communication and Information Systems*, 36(1):1–10, 2021.

[137] Vu Nguyen Ha, Duy H. N. Nguyen, and Jean-François Frigon. Energy-efficient hybrid precoding for mmwave multi-user systems. *IEEE Systems Journal*, 13(4):3770–3781, 2019.

[138] Xinyu Gao, Linglong Dai, Yifan Sun, Shuangfeng Han, and Chih-Lin I. Machine learning inspired energy-efficient hybrid precoding for mmwave massive mimo systems. *IEEE Transactions on Wireless Communications*, 17(11):7505–7520, 2018.

[139] J. Zhang, X. Yu, and K. B. Letaief. Hybrid beamforming for 5g and beyond millimeter-wave systems: A holistic view. *IEEE Open Journal of the Communications Society*, 1:77–91, 2019.

[140] Ming Li, Zihuang Wang, Hongyu Li, Qian Liu, and Liang Zhou. Hardware-efficient hybrid beamforming solution for mmwave mimo systems. *IEEE Wireless Communications*, 2020. Preprint: arXiv:1812.07501.

[141] Wendi Wang, Huarui Yin, et al. Hybrid beamforming design and performance with imperfect phase shifters in multiuser mmwave systems. *arXiv preprint arXiv:1803.00317*, 2018.

[142] Zihuan Wang, Jianhua Zhang, et al. Hybrid beamforming with one-bit quantized phase shifters in mmwave mimo systems. *Preprint*, 2019.

[143] Yipu Yuan, Jung-Lang Yu, Yuzhuo Pan, and Shiyan Zheng. Low-resolution phase shifter-based hybrid precoding for mmwave massive mimo. *EURASIP Journal on Wireless Communications and Networking*, 2024(85):1–21, 2024.

[144] Xizixiang Wei, Yi Jiang, Qingwen Liu, and Xin Wang. Calibration of phase shifter network for hybrid beamforming in mmwave massive mimo systems. *IEEE Transactions on Signal Processing*, 68:2294–2309, 2020.

[145] Hamed Nosrati, Elias Aboutanios, Xiangrong Wang, and David Smith. Switch-based hybrid beamforming for massive MIMO communications in mmwave bands. *Signal Processing*, 200:108659, 2022.

[146] Chris Dick, Fred Harris, and Michael Rice. Implementing a real-time beamformer on an FPGA platform. *Xcell Journal*, pages 36–40, 2007.

[147] Rakesh Palisetty, Luis Manuel Garcés-Socarrás, Houssem Chaker, Vibhum Singh, Geoffrey Eappen, Wallace A. Martins, Vu Nguyen Ha, Juan A. Vasquez-Peralvo, Jorge

L. G. Rios, Juan C. Merlano-Duncan, Symeon Chatzinotas, Björn Ottersten, Ayhan Coskun, Steven King, Salvatore D'Addio, and Piero Angeletti. FPGA implementation of efficient beamformer for on-board processing in MEO satellites. In *Proc. IEEE 34th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, pages 1–7, 2023.

[148] Ming Li, Zihuan Wang, Hongyu Li, Qian Liu, and Liang Zhou. A hardware-efficient hybrid beamforming solution for mmwave mimo systems. *arXiv preprint arXiv:1812.07501*, 2018. Hardware-efficient POS+SW hybrid architecture; precoding/-combining and channel estimation design.

[149] ITU-R. Guidelines for evaluation of radio interface technologies for imt-2020. Technical Report Report ITU-R M.2412-0, International Telecommunication Union, 2017.

[150] Liang Dai et al. Energy-efficient hybrid precoding for mmwave massive mimo systems. *IEEE Transactions on Communications*, 67(3):1768–1781, 2019.

[151] Smail Labed and Naceur Aounallah. Two-level structured bayesian sparse recovery for cascaded uplink channel estimation in ris-assisted mmwave mimo systems. *Preprint*, 2025.

[152] R. Méndez-Rial et al. Hybrid mimo architectures for millimeter wave communications: Phase shifters or switches? *IEEE Access*, 4:247–267, 2016.

[153] Mengqian Tian, Shangang Fan, Jie Yang, Jian Xiong, and Guan Gui. Performance analysis of energy-efficient hybrid precoding for mmwave massive mimo system. In *Communications, Signal Processing, and Systems*, pages 92–99. Springer, 2019.

[154] Y. N. Samir, F. W. Zaki, and H. B. Nafea. Performance evaluation of mmwave–mimo based on hybrid precoding and combining. *Mansoura Engineering Journal*, 48(4), 2023.

[155] L. Zhao et al. Low-complexity hybrid precoding for sub-connected mmwave massive mimo. *Signal Processing*, 2024. Channel-hardening aided hybrid precoding.

[156] O. S. Faragallah, H. S. El-sayed, and M. G. El-Mashed. Hybrid precoding design for mmwave multi-user massive mimo under high mobility. *Transactions on Emerging Telecommunications Technologies*, 2023.

[157] Ahmed Alkhateeb, Geert Leus, and Robert W. Heath. Channel estimation and hybrid precoding for millimeter wave cellular systems. *IEEE Journal of Selected Topics in Signal Processing*, 8(5):831–846, 2014.

[158] J. Chen, Y. Li, X. Lin, F. Zhou, and Y. Wu. Double-structured sparsity-based channel estimation for ris-assisted massive mimo systems. *IEEE Wireless Communications Letters*, 10(4):856–860, 2021.

[159] T. Wang, S. Zhang, Y. Li, and J. Wei. Low-rank matrix completion based channel estimation for irs-assisted multiuser mimo systems. *IEEE Wireless Communications Letters*, 11(6):1167–1171, 2022.

[160] Y. Sun, Y. Liu, X. Gao, and X. Wei. Sparse bayesian learning for cascaded channel estimation in ris-aided mmwave mimo systems. *IEEE Transactions on Vehicular Technology*, 72(2):2208–2222, 2023.

[161] Y. Chen, L. Dai, et al. Channel estimation for ris-assisted wireless communications: Stationary or non-stationary? *IEEE Transactions on Signal Processing*, 2024. Early Access.

[162] Z. Albataineh, M. Amjad, and K. Qaraqe. Channel estimation for ris-assisted mmwave systems via entropy-regularized compressed sensing. *Scientific Reports*, 12(22018), 2022.

[163] Olivier Ledoit and Michael Wolf. The power of (non-)linear shrinking: A review and guide to covariance matrix estimation. *Journal of Financial Econometrics*, 20(1):187–218, 2022.

[164] Y. Liu, X. Sun, and S. Zhao. A covariance matrix shrinkage method with toeplitz rectified target for doa estimation. *AEU – International Journal of Electronics and Communications*, 81:50–55, 2017.

[165] X. Jing, S. Zhang, and W. Wang. Low-complexity soft-output signal detector for multiuser massive mimo. *Digital Signal Processing*, 111:102979, 2021.

[166] A. H. Neama et al. Enhanced hybrid detection technique for mmse using admm–cg with tridiagonal initialization. *Journal of Communications Software and Systems*, 20(4):287–296, 2024.

[167] Ahmed Alkhateeb, Geert Leus, and Robert W. Heath. Limited feedback hybrid precoding for multi-user millimeter wave systems. In *Proc. IEEE Information Theory and Applications Workshop (ITA)*, pages 1–10, 2014.

[168] Qianrui Li. Hybrid precoding for wideband multi-user MIMO millimeter wave system. In *Proc. IEEE Wireless Communications and Networking Conference (WCNC)*, pages 1–6, 2019.

[169] Hamed Radpour, Markus Hofer, and Thomas Zemen. Adaptive RIS control for mobile mmwave NLoS communication using single-bit feedback. *arXiv preprint arXiv:2504.16874*, 2025.

[170] Murat Bayraktar, Nuria González-Prelcic, and Hao Chen. Hybrid precoding and combining for mmwave full-duplex joint radar and communication systems under self-interference. In *Proc. IEEE International Conference on Communications (ICC)*, pages 4064–4069, 2024.

[171] Wagner V. F. Mauricio et al. Scheduling for massive mimo with hybrid precoding: Algorithms and system-level perspectives. *IEEE Transactions on Wireless Communications*, 2022.

[172] Y. T. Chen et al. Cross-layer stream allocation of mmimo-ofdm hybrid beamforming. *Open Physics*, 26(4), 2025.

[173] Xiaoming Yu, Jun-Chao Shen, Jun Zhang, and Khaled B. Letaief. Alternating minimization algorithms for hybrid precoding in millimeter wave mimo systems. *IEEE Journal on Selected Areas in Communications*, 10(3):485–500, 2016.

[174] Anastasios Papazafeiropoulos et al. Adaptive and sparse adaptive channel estimation for mmwave massive mimo. *Journal of Engineering Science*, 14(2), 2023.

[175] F. Samklang et al. Optimal hybrid precoding based qoe for partially connected mmwave massive mimo systems. *Journal of Communications and Networks*, 23(6):441–452, 2021.

[176] Divya Singh. Machine learning-based hybrid precoding for next-generation wireless communication systems: A comprehensive review. In J. Sen et al., editors, *Information and Communication Technology for Intelligent Systems. ICTIS 2024*, volume 1084 of *Lecture Notes in Networks and Systems*, pages 579–590, Singapore, 2025. Springer Nature Singapore.

[177] Chukwudi Nwufo et al. Reconfigurable intelligent surfaces (ris) and their role in millimeter-wave massive mimo networks: An overview. *European Transactions on Telecommunications*, 36(3):e70164, 2025.

[178] Javier Pérez Santacruz, Elmine Meyer, Roel X. F. Budé, Catalina Stan, Antonio Jurado-Navas, Ulf Johannsen, Idelfonso Tafur Monroy, and Simon Rommel. Outdoor mm-wave 5G/6G transmission with adaptive analog beamforming and IFoF fronthaul. *Scientific Reports*, 13(1):13945, 2023.

[179] Naceur Aounallah and Smail Labed. Low-complexity data detection based on combined SD and SOR algorithms for uplink massive MIMO systems. *Journal of Microwaves, Optoelectronics and Electromagnetic Applications*, 24(4), 2025. Editor: João C. W. A. Costa. Publication in collection: 17 Apr 2026.

[180] Naceur Aounallah and Labed Smail. Low-complexity Optimized Version of AOR Algorithm for Signal Precoding in Large-scale MIMO Systems. *Journal of Telecommunications and Information Technology*, 102(4):11–19, 2025.

Scientific Production

Authored paper- Published

- Smail Labed, Naceur Aounallah. "Efficient iterative detection based on conjugate gradient and successive over-relaxation methods for uplink massive MIMO systems." Journal of Telecommunications and Information Technology (2023). [29].

Co-Authored paper- Published

- Naceur Aounallah, Smail Labed. "Low-Complexity Data Detection Based on Combined SD and SOR Algorithms for Uplink Massive MIMO Systems." Journal of Microwaves, Optoelectronics and Electromagnetic Applications, 2025, vol. 24, p. e2025295878. [179].

- Naceur Aounallah, Smail Labed. "Low-complexity Optimized Version of AOR Algorithm for Signal Precoding in Large-scale MIMO Systems." Journal of Telecommunications and Information Technology, 2025, vol. 102, no 4, p. 11-19. [180].

Conference paper- Published

- Naceur Aounallah, Smail Labed. "Improved DOA estimation algorithms using modified covariance matrix." 2022 19th International Multi-Conference on Systems, Signals Devices (SSD). IEEE, 2022. p. 1737-1742. [100].

- Naceur Aounallah, Smail Labed. "Efficient OMP Variant for Large Scale MIMO Hybrid Precoding System." International Conference on Electrical Engineering and Control Applications. Singapore : Springer Nature Singapore, 2024. p. 376-384. [117].

Conference paper

- Smail Labed, Naceur Aounallah. "Low-complexity and Efficient Massive MIMO Signal Detection Based on Richardson Iteration Method with Various Initialization Techniques." ourth International Conference on Technological Advances in Electrical Engineering (ICTAEE'23), Skikda, Algeria, 2023. [28].

- Smail Labed, Naceur Aounallah. "Enhanced Direction of Arrival Estimation Based on Tuning Covariance Matrix Shrinkage Estimator." 2023 International Conference on Electrical Engineering and Advanced Technology (ICEEAT), held at Batna 2 University on November 5-7th, 2023.
- Naceur Aounallah, Smail Labed. "Fast Low-Complexity Data Detection Based on an Optimized SOR Algorithm for Uplink Massive MIMO Systems." 3rd International Conference on Computational Engineering and Intelligence Systems (ICCEIS2025) 21th-23th November 2025 At: University of Boumerdes, Algeria.