

A Study of Robust Debiasing Methods for Sparse Modeling: Moreau Enhancement and Beyond

July 2024

SUZUKI, Kyohei

A Thesis for the Degree of Ph.D. in Engineering

A Study of Robust Debiasing Methods for Sparse Modeling:
Moreau Enhancement and Beyond

July 2024

Graduate School of Science and Technology
Keio University

SUZUKI, Kyohei

Abstract

In recent years, methods to estimate sparse signals accurately even in the presence of severe outliers are gaining increasing attention. Existing outlier-robust approaches usually suffer from a severe tradeoff between robustness and global optimality. On the other hand, typical sparse estimation methods face the risk of missing some important groups of highly correlated features. This thesis addresses these issues by exploring an effective way of utilizing the so-called Moreau enhancement technique and by defining a new class of operators which extends this technique.

Chapter 1 introduces the background and motivation of the study.

Chapter 2 provides mathematical preliminaries which will be used throughout this thesis.

Chapter 3 presents a robust method for sparse signal estimation based on the minimax concave (MC) function (the Moreau enhancement of the ℓ_1 norm) to resolve the above tradeoff problem. The influence of outliers is reduced by the saturation property of the MC loss function. Moreover, our approach enjoys the global optimality by using the weak convexity of the MC function. Numerical examples show the remarkable robustness of the proposed method.

Chapter 4 extends the estimation method by distinguishing the statistical differences between Gaussian noise and outliers by introducing an auxiliary vector to model the noise. This improves estimation accuracy even in highly noisy environments. In addition, in analogy to the popular elastic net, the Tikhonov regularizer is used together with the MC function, yielding “the grouping effect” and resolving the issue of missing important groups mentioned above. In contrast to the elastic net, the grouping effect of the proposed method does not depend on the magnitudes of outliers. Numerical examples show the efficacy of the proposed method even in highly noisy environments.

Chapter 5 introduces a new notion of “the external division operator”, which extends the idea of the Moreau enhancement, and it presents a method to extract all correlated features accurately. The octagonal shrinkage and clustering algorithm for regression (OSCAR), which gives a better grouping effect than the elastic net, is known to underestimate the target signals. The

idea of the external division operator comes from the fact that the proximity operator of the MC function can be expressed as “an external division of two proximity operators of the ℓ_1 norm”. The external division operator for OSCAR turns out to be a generalization of its Moreau enhancement. Numerical examples demonstrate that the proposed method improves the performance dramatically by reducing the estimation bias.

Chapter 6 summarizes the results of this thesis and gives an outlook on the future research.

Acknowledgments

This thesis completes my studies in the undergraduate, master's, and doctoral courses at Keio University from 2019 to 2024. I would like to express my deepest gratitude to my supervisor Professor Masahiro Yukawa for his innumerable advice, discussion, and suggestions to my entire study in the graduate school. This work would not have been realized without his persistent help. Furthermore, it is my pleasure to thank all the current and former students of Yukawa Laboratory for the friendship. My special thanks go to Mrs. Misao Aso for her careful support on my overall activities related to my research. I would like also deeply thank to Professors Masaaki Ikehara, Toru Namerikawa, and Senior Assistant Professor Daichi Kitahara for serving as members of the examining committee of this thesis. My works related to this thesis were supported by JSPS Grants-in-Aid (18H01446, 22J22588, and 22H01492).

Contents

Abstract	i
Acknowledgments	iii
List of Figures	ix
List of Tables	xii
List of Symbols and Notation	xv
List of Abbreviations	xix
1 General Introduction	1
1.1 Background	1
1.1.1 Sparse Signal Estimation	1
1.1.2 Bias Reduction in Sparse Signal Estimation	3
1.1.3 Jointly Sparse Signal Recovery	4
1.1.4 Outlier-Robust Regression	6
1.1.4.1 Robust Jointly Sparse Signal Recovery	6
1.1.4.2 Issues Regarding Existing Robust Loss Functions	7
1.1.4.3 Sparse Outlier-Robust Regression	8
1.1.5 Feature Grouping	10
1.2 This Study	12
2 Preliminaries	15
2.1 Notation and Definitions	15
2.1.1 Bounded Linear Operators	16
2.1.2 Selected Elements of Convex Analysis	16
2.1.3 Nonexpansive and Firmly Nonexpansive Operators	17
2.1.4 Proximity Operator and Moreau Envelope	18
2.1.5 Monotone Operator	19
2.2 Bias Reduction Methods	20
2.2.1 Estimation Bias of Lasso	20

2.2.2	MC Function and Moreau Enhancement	21
2.2.3	Other Nonconvex Sparsity-Inducing Penalties for Bias Reduction Methods	24
2.3	Outlier-Robust Regression	25
2.4	Feature Grouping Methods	26
2.5	Firm-Shrinkage Operator	27
3	Robust Recovery of Jointly-Sparse Signals Based on Mini- max Concave Functions	29
3.1	Introduction	29
3.2	Proposed Approach	29
3.2.1	Problem Formulation	30
3.2.2	Motivation of Reformulation	31
3.2.3	Reformulation of (P_1)	31
3.2.4	Closed-Form Expressions for the Operators	34
3.2.5	Convergence Conditions and Complexity	34
3.2.6	Remarks	36
3.3	Numerical Examples	37
3.3.1	Experiment 3-A: Toy Data	37
3.3.1.1	Robustness	37
3.3.1.2	Support Recovery	41
3.3.2	Experiment 3-B: Application to MEGG Signal Recov- ery Problem	44
3.4	Conclusion	51
4	Sparse Stable Outlier-Robust Regression	53
4.1	Introduction	53
4.2	Problem Statement	54
4.3	Proposed Method	54
4.3.1	Reformulation of the Problem	56
4.3.2	Convexity Conditions	57
4.3.3	Convergence Conditions and Complexity	58
4.3.4	Grouping Effect	59
4.4	Convexity Conditions for a General Model	63
4.5	Numerical Examples	66
4.5.1	Experiment 4-A: Toy data	67
4.5.1.1	Performance with Different Outlier Magnitudes	68
4.5.1.2	Performance with Different Outlier Sparsity, m/n , and SNR	70
4.5.1.3	Comparison of Execution Time	74
4.5.1.4	Grouping Effect in the Absence of Outliers .	74
4.5.1.5	Grouping Effect in the Presence of Outliers .	78
4.5.2	Experiment 4-B: Fluctuations of Hyperparameters . .	80
4.5.2.1	Hyperparameters ρ , μ_1 , γ_1 , and γ_2	80

4.5.2.2	Hyperparameter α	82
4.5.3	Experiment 4-C: Real Data (Application to Speech De-noising)	84
4.6	Conclusion	87
5	External Division of Two Proximity Operators: An Application to Feature Grouping	89
5.1	Introduction	89
5.1.1	Operator Splitting Algorithms	89
5.1.2	Contributions	90
5.1.3	Relation to Previous Works	92
5.2	Approaches to Bias Reduction of OSCAR	92
5.2.1	Moreau Enhancement of the OSCAR Regularizer	92
5.2.2	Firm-Shrinkage Operator as External Division	96
5.2.3	Proposed Bias Reduction Approach for OSCAR: De-biased OSCAR	97
5.3	Proposed Method: External Division Operator	99
5.3.1	External Division of Two Nonexpansive Operators	101
5.3.2	External Division Operator of Two Proximity Operators	102
5.3.3	Gradient Algorithm With External Division Operator: Case of Strongly Convex Fidelity Term	103
5.3.4	Gradient Algorithm With External Division Operator: Case When Strong Convexity is Restricted to Subspace	104
5.3.5	Convexity Condition of ψ_ω Given in (5.29)	105
5.3.6	A Closed-Form Expression of φ_ω Defined in (5.31)	110
5.3.7	Relation of φ_ω and the Moreau Enhancement (the Case of $g_2 = \eta g_1$ for $\eta > 1$)	112
5.4	Numerical Examples	115
5.4.1	Experiment 5-A: Toy Data	115
5.4.2	Experiment 5-B: Influence of Hyperparameter η	116
5.5	Conclusion	120
6	General Conclusion	121
A	Contours of the Quadratic Loss Function	125
B	Weak Convexities of Some Nonconvex Penalties	127
C	Forward-Backward-Based Primal-Dual Splitting Method	129
D	Big Oh Notation	131
E	Matrix Computation Methods	133
E.1	Power Methods	133
E.2	QR Decomposition	134

F	LiMES Model	135
	F.1 Stable Outlier-Robust Regression	135
	F.2 LiMES Model: Convexity Conditions	136
G	Monotone Lipschitz-Gradient Denoiser	139
H	Decomposition of Proximity Operator	141
I	Proof for Chapter 3	143
	I.1 Proof of (3.11)	143
	I.2 Proof of Proposition 3.1	144
	I.3 Proof of Proposition 3.3	145
J	Proof for Chapter 4	149
	J.1 Proof of (4.15)	149
	J.2 Proof of Proposition 4.1	150
	J.3 An Alternative Proof of Proposition 4.1 (the Case of $\alpha = 1$)	153
	J.4 Proof of Proposition 4.3	156
	J.5 Proof of Proposition 4.4	157
	J.6 Proof of Proposition 4.5	160
	J.7 Proof of Proposition 4.6	162
	J.8 Proof of (4.51)	163
K	Proof for Chapter 5	165
	K.1 Proof of Proposition 5.1	165
	K.2 Proof of Proposition 5.7	173
	K.3 Proof of Proposition 5.8	174
	K.4 Proof of Proposition 5.9	175
	K.5 Proof of Proposition 5.10	179
	Bibliography	181
	Publications Related to the Dissertation	195
	Other Publications	197

List of Figures

1.1	Penalty functions for sparse signal estimation.	3
1.2	Comparison of the squared ℓ_2 norm and robust loss functions.	6
1.3	Derivative of robust loss functions.	9
1.4	An illustration of feature grouping.	10
1.5	Contours of the quadratic loss function centered at the least squares solution (blue) and the constraint region of the OSCAR regularizer (red) in the two-dimensional case. (a) The case when two features are uncorrelated, and (b) when two features are highly correlated.	11
1.6	The overview of this dissertation.	12
2.1	Contours of the ℓ_2 norm, the ℓ_1 norm, and the MC function.	22
3.1	Computational complexities of the proposed approach and the RFS.	36
3.2	Learning curves under $n = 512$, $m = 1024$, $n = 32$, outlier rate 30% and SNR 30 dB.	38
3.3	NMSE across outlier rate under $n = 160$, $m = 256$, and $r = 32$	40
3.4	Recovery probability across outlier rate for $n = 256$, $r = 32$, $m = 128$, $k = 16$, and SNR 10 dB.	42
3.5	Recovery probability as a function of (a) k/n and (b) r under SNR 10 dB, outlier rate 30%, and SOR -30 dB.	43
3.6	a) Clean MECG signals. (b) Amplitudes of wavelet coefficients of the MECG signals in each of the four channels.	45
3.7	(a) NMSE as a function of SOR under outlier rate 80%, (b) NMSE as a function of outlier rate under SOR -75 dB.	46
3.8	Measurements, original ECG signal, and recovered ECG signals from (a) the proposed method and (b) the RFS under outlier rate 80% and SOR -40 dB for Lead 1.	47
3.9	Measurements, original ECG signal, and recovered ECG signals from (a) the proposed method and (b) the RFS under outlier rate 80% and SOR -40 dB for Lead 2.	48

3.10	Measurements, original ECG signal, and recovered ECG signals from (a) the proposed method and (b) the RFS under outlier rate 80% and SOR -40 dB for V1.	49
3.11	Measurements, original ECG signal, and recovered ECG signals from (a) the proposed method and (b) the RFS under outlier rate 80% and SOR -40 dB for V2.	50
4.1	System mismatch across M_o for $\kappa_x = \lfloor 0.05n \rfloor$ and $\kappa_o = \lfloor 0.3m \rfloor$. (a) Overdetermined case ($m = 128, n = 64$) under SNR 5 dB. (b) Underdetermined case ($m = 64, n = 128$) under SNR 15 dB.	69
4.2	Learning curves for $m = 256, n = 512, \kappa_x = \lfloor 0.05n \rfloor, \kappa_o = \lfloor 0.3m \rfloor$, and $M_o = 100$ under SNR 15 dB.	70
4.3	System mismatch across (a) \bar{M}_o for outliers of model O_2 and (b) $\mu_{\bar{M}_o}$ for model O_3 , respectively, for $\kappa_x = \lfloor 0.05n \rfloor, \kappa_o = \lfloor 0.3m \rfloor, m = 128$, and $n = 64$ under SNR 15 dB.	71
4.4	System mismatch across outlier sparsity for $\kappa_x = \lfloor 0.05n \rfloor$ and $M_o = 100$. (a) Overdetermined case ($m = 128, n = 64$) under SNR 5 dB. (b) Underdetermined case ($m = 64, n = 128$) under SNR 15 dB.	72
4.5	System mismatch across m/n for $n = 128, \kappa_x = \lfloor 0.05n \rfloor, \kappa_o = \lfloor 0.3m \rfloor$, and $M_o = 100$ under SNR 15 dB.	73
4.6	System mismatch across SNR for $m = 64, n = 128, \kappa_x = \lfloor 0.05n \rfloor, \kappa_o = \lfloor 0.3m \rfloor$, and $M_o = 100$	73
4.7	CPU time across m for $n = 512, \kappa_x = \lfloor 0.05n \rfloor, \kappa_o = \lfloor 0.3m \rfloor$, and $M_o = 100$ under SNR 15 dB for outliers of model O_1	74
4.8	Solution paths for (a) lasso, (b) elastic net, (c) MC, (d) YALL1, (e) Huber FISTA, (f) extended lasso, (g) RPGG, and (h) S-SORR.	77
4.9	Solution paths for (a) lasso, (b) elastic net, (c) MC, (d) YALL1, (e) Huber FISTA, (f) extended lasso, (g) RPGG, and (h) S-SORR.	79
4.10	System mismatch across hyperparameters of S-SORR for $m = 64, n = 128, \kappa_x = \lfloor 0.05n \rfloor, \kappa_o = \lfloor 0.3m \rfloor$, and $\bar{M}_o = 100$ under SNR 15 dB.	81
4.11	Case with (a) $\gamma_1 = M_o$ and (b) $\gamma_1 > M_o$	82
4.12	The values of tuned hyperparameters across (a) \bar{M}_o for outliers of model O_2 and (b) $\mu_{\bar{M}_o}$ for model O_3 , respectively, for $m = 64, n = 128, \kappa_x = \lfloor 0.05n \rfloor, \kappa_o = \lfloor 0.3m \rfloor$, and $\bar{M}_o = 100$ under SNR 15 dB. The values of μ_2, \bar{M}_o , and $\mu_{\bar{M}_o}$ are plotted for reference.	83

4.13	(a) System mismatch and (b) sparseness measure across M_o for $m = 64$, $n = 128$, $\kappa_x = \lfloor 0.15m \rfloor$, and $\kappa_o = \lfloor 0.1m \rfloor$ under SNR 15 dB.	84
4.14	Application to speech denoising. (a) SNR 15 dB and SOR -30 dB, (b) SNR 15 dB and $\kappa_o = \lfloor 0.05m \rfloor$, and (c) SOR -40 dB and $\kappa_o = \lfloor 0.05m \rfloor$	86
5.1	Motivation of this study.	91
5.2	(a) Inclusion relation among the five cases of (5.6). (b) Visualization of the five cases in the three-dimensional case.	94
5.3	The function value of (5.12).	95
5.4	The contours of the Moreau enhancement of OSCAR (red) for $\lambda_1 = \lambda_2 := 0.5$ and $\gamma := 2$ and OSCAR (black) for the two-dimensional case.	96
5.5	A relation between the firm-shrinkage and two soft-shrinkage operators.	96
5.6	A geometric interpretation of $T_{\lambda_1, \lambda_2, \omega, \eta}^{\text{DO OSCAR}}(\mathbf{x})$ for the two-dimensional case. (a) x_1 and x_2 are in the same group, (b) x_2 is close to 0, and (c) x_1 and x_2 are in the different groups. The gray regions show $\mathbb{R}_{++}^2 \cap \mathcal{K}_{\geq 0}^2$	98
5.7	An overview of relations of the theoretical results in this chapter.	100
5.8	The relation among ψ_ω given in (5.29) and the Moreau envelopes ${}^1(g_1^*)$ and ${}^1(g_2^*)$ of the conjugate functions g_1^* and g_2^*	103
5.9	The function φ_ω in Proposition 5.9 for $g_1 := \cdot $, $g_2 := \eta \cdot $, and $\omega := 2$ with (a) $\eta := 2.0$, (b) $\eta := 1.5$, and (c) $\eta := 3.0$	111
5.10	The function φ_ω given in (5.59) with $\lambda_1 = \lambda_2 := 0.5$ and $\omega := 2.0$ for (a), $\eta := 2.0$ (the case when φ_ω coincides with the Moreau enhancement of the OSCAR regularizer), (b) $\eta := 1.5$, and (c) $\eta := 3.0$	114
5.11	System mismatch across σ_ϵ for dataset A with (a) SNR 20 dB and (b) SNR 15 dB.	117
5.12	System mismatch across r for dataset B with (a) $m := 400$ and (b) $m := 40$	118
5.13	System mismatch across hyperparameter η . (a) dataset A with SNR 20 dB and $\kappa := 0.1$, (b) the same as (a) except for SNR 15 dB, (c) dataset B with $m := 400$ and $r := 0.8$, and (d) the same as (c) except for $m := 40$. The blue lines indicate the value $\eta = \omega(\omega - 1)^{-1}$	119

List of Tables

1.1	Comparisons of robust loss functions	8
1.2	Comparisons of sparse outlier-robust methods.	9
3.1	Computational complexity (q : the number of iterations).	36
4.1	Total computational complexity (p : number of power iterations, q : number of algorithm iterations).	60
4.2	Formulations for sparse signal recovery.	67
4.3	Formulations for sparse outlier-robust signal regression.	67

List of Symbols and Notation

Real Line

\mathbb{R}	The set of real numbers
\mathbb{R}_{++}	The set of strictly positive real numbers
\mathbb{N}	The set of nonnegative integers
\mathbb{N}^*	The set of strictly positive integers
\min	Minimum
\max	Maximum
\sup	Supremum
$\xi \geq 0$	The real number ξ is nonnegative
$\xi > 0$	The real number ξ is strictly positive
$\xi \leq 0$	The real number ξ is nonpositive
$\xi < 0$	The real number ξ is strictly negative

Finite Dimensional Hilbert Spaces

$(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ and $(\mathcal{K}, \langle \cdot, \cdot \rangle_{\mathcal{K}})$	Finite-dimensional real Hilbert spaces
\mathbb{R}^n	The standard n -dimensional Euclidean space
$\mathbb{R}^{n \times m}$	Space of $m \times n$ real matrices
Id	Identity operator
$\{e_{d,i}\}_{i=1}^d$	The standard basis of \mathbb{R}^d for any dimension d
$(x_k)_{k=1}^{+\infty} \subset \mathcal{H}$	A sequence $(x_k \in \mathcal{H} \mid k = 1, 2, \dots)$
$\lim_{k \rightarrow +\infty} x_k = x$	A sequence $(x_k)_{k=1}^{+\infty}$ converges to x
$\limsup_{k \rightarrow +\infty} x_k$	Limit supremum of a sequence $(x_k)_{k=1}^n$
$\mathbb{E}[x]$	Expected value of a random variable x
$\mathcal{N}(\mu, v)$	Gaussian distribution with mean $\mu \in \mathbb{R}$ and variance $v > 0$
$\mathcal{U}(a, b)$	Uniform distribution over the closed interval $[a, b]$ for $a, b \in \mathbb{R}$ such that $a < b$
\mathcal{O}	Big Oh notation

Sets

$\{x \mid x \in P\}$	Set of all elements satisfying P
$C_1 \cup C_2$	Union of sets C_1 and C_2
$C_1 \cap C_2$	Intersection of sets C_1 and C_2
\exists	“there exists”
\forall	“for all”
2^C	Power set of a set C
$\text{span } C$	Span of a set C
$\text{int } C$	Interior of a set C
$B_{\mathcal{H}}(0, r)$	Closed ball with center $0 \in \mathcal{H}$ and radius r
d_C	Distance function to a set C
C^\perp	Orthogonal complement of a set C
$\text{card } C$	Cardinality of a set C
$C_1 \times C_2 \times \dots \times C_n$	Cartesian products of sets C_1, C_2, \dots, C_n
\emptyset	Empty set
$\mathcal{K}_{\geq 0}^n$	The set $\{\mathbf{x} \in \mathbb{R}^n \mid x_1 \geq x_2 \geq \dots \geq x_n \geq 0\}$
$\mathcal{K}_{>}^n$	The set $\{\mathbf{x} \in \mathbb{R}^n \mid x_1 > x_2 > \dots > x_n\}$

Vectors and Matrices

\mathbf{a}_i	The i th column of a matrix \mathbf{A}
$\mathbf{A}_{(i,:)}$	The i th row of a matrix \mathbf{A}
$A_{i,j}$	The (i, j) entry of a matrix \mathbf{A}
\mathbf{I}_n	$n \times n$ identity matrix
$\mathbf{0}_{n \times m}$	$n \times m$ zero matrix
$\mathbf{1}_n$	The vector $[1, 1, \dots, 1] \in \mathbb{R}^n$
$(\cdot)^\top$	Vector or matrix transpose
$\ \mathbf{x}\ _0$	The ℓ_0 pseudo-norm of \mathbf{x}
$\ \mathbf{x}\ _p$	The ℓ_p norm of \mathbf{x} for $p \geq 1$ and ℓ_p quasi-norm of \mathbf{x} for $0 < p < 1$
$\ \mathbf{x}\ _\infty$	The ℓ_∞ norm of \mathbf{x}
$\ \mathbf{A}\ _{2,0}$	The $\ell_{2,0}$ pseudo-norm of a matrix \mathbf{A}
$\ \mathbf{A}\ _{2,1}$	The $\ell_{2,1}$ norm of a matrix \mathbf{A}
$\ \mathbf{A}\ _{2,\infty}$	The $\ell_{2,\infty}$ norm of a matrix \mathbf{A}
$\ \mathbf{A}\ _F$	The Frobenius norm of a matrix \mathbf{A}
$\ \mathbf{A}\ _{\text{nuc}}$	The nuclear norm of a matrix \mathbf{A}
$\ \mathbf{A}\ _2$	The spectral norm of a square matrix \mathbf{A}
$\langle \mathbf{v}, \tilde{\mathbf{v}} \rangle_2$	The inner product of vectors \mathbf{v} and $\tilde{\mathbf{v}}$
$\langle \mathbf{A}, \mathbf{B} \rangle_F$	The Frobenius (trace) inner product of matrices \mathbf{A} and \mathbf{B}
$\lambda_{\max}(\mathbf{A})$	The largest eigenvalue of a symmetric matrix \mathbf{A} .

$\lambda_{\min}(\mathbf{A})$	The smallest eigenvalue of a symmetric matrix \mathbf{A} .
$\lambda_{\min}^{++}(\mathbf{A})$	The smallest strictly positive eigenvalue of a symmetric matrix \mathbf{A} .
\mathbf{A}^\dagger	The Moore-Penrose pseudo inverse of a matrix \mathbf{A}
diag	Creates a diagonal matrix
$\text{Tr}(\mathbf{A})$	Trace of a square matrix \mathbf{A}
rank \mathbf{A}	Rank of a matrix \mathbf{A}
range \mathbf{A}	Range space of a matrix \mathbf{A}
spark \mathbf{A}	Spark of \mathbf{A}
$(\cdot)_\diamond$	Sparse vector or matrix
$(\cdot)_*$	Gaussian random vector or matrix
supp(\mathbf{x})	Support set of a vector \mathbf{x}
$\mathbf{P}(\mathbf{x})$	Permutation matrix which sorts the components of \mathbf{x} in non-increasing order
$ \mathbf{x} $	The vector with the absolute values of the components of \mathbf{x}
$ \mathbf{x} _\downarrow$	The vector obtained by sorting the components of \mathbf{x} in non-increasing order of magnitude.
$ \mathbf{x} _{[i]}$	The i th largest component of $ \mathbf{x} $.

Operators

$\mathcal{B}(\mathcal{H}_1, \mathcal{H}_2)$	Set of bounded linear operators from \mathcal{H}_1 to \mathcal{H}_2
$L \succeq O$ ($L \succ O$)	Bounded linear operator L is positive semidefinite (positive definite)
L^*	The adjoint operator of a linear operator L
$L^{1/2}$	A square root of a positive semidefinite self-adjoint operator L
$\langle \cdot, \cdot \rangle_L$	L inner product for a positive definite operator
0	Zero operator (defined on an arbitrary space) which maps any point to the null vector
gra B	Graph of an operator B
dom B	Domain of an operator B
B^{-1}	Inverse of an operator B
range B	Range of an operator B
$T : \mathcal{H}_1 \rightarrow \mathcal{H}_2$	T is an operator from \mathcal{H}_1 to \mathcal{H}_2 defined everywhere on \mathcal{H}
$T_1 \circ T_2$	Composition of operators T_1 and T_2
Fix T	Set of fixed points of an operator T
soft	The soft-shrinkage operator defined on \mathbb{R}
Soft	The soft-shrinkage operator defined on \mathbb{R}^n
firm	The firm-shrinkage operator defined on \mathbb{R}

Firm	The firm-shrinkage operator defined on \mathbb{R}^n
sign	The sign operator defined on \mathbb{R}
Sign	The sign operator defined on \mathbb{R}^n
P_C	Projector onto a nonempty closed convex set C

Functions

$\text{lev}_{\leq a}$	Level set of a function f at height a
$\Gamma_0(\mathcal{H})$	Set of proper lower-semicontinuous convex functions from \mathcal{H} to $(-\infty, +\infty]$
f^*	Conjugate of a function f
σ_C	Support function of a set C
ι_C	Indicator function of a set C
$\ \cdot\ _*$	Dual norm of $\ \cdot\ $
$\text{argmin } f$	Set of global minimizers of a function f
$\text{Prox}_{\gamma f}$	Proximity operator of a function f
γf	Moreau envelope of index γ of a function f
f'	Derivative of a function f
$D_f(\mathbf{x}; \mathbf{d})$	Generalized directional derivative of a function f at a vector \mathbf{x} in the direction \mathbf{d}
∇f	Gradient of a function f
∂f	Subdifferential of a function f
$\lfloor \cdot \rfloor$	Floor function
$\Phi_{\gamma}^{\text{MC}}$	The MC function of index γ
$\Phi_{\gamma}^{\text{GMC}}$	The GMC function of index γ
Ψ_B	The GME function for a function Ψ and a linear operator B

List of Abbreviations

Methods

ADMM	alternating direction method of multipliers	p. 27
CoSaMP	compressive sensing matching pursuit	p. 2
CS-MUSIC	compressive MUSIC	p. 5
DOSCAR	debiased OSCAR	p. 86
FISTA	fast iterative shrinkage algorithm	p. 9
IHT	iterative hard thresholding	p. 2
ISTA	iterative shrinkage-thresholding algorithm	p. 85
LAD-lasso	least absolute deviation-lasso	p. 9
lasso	least absolute shrinkage and selection operator	p. 2
MP	matching pursuit	p. 2
MUSIC	multiple signal classification	p. 5
NIHT	normalized IHT	p. 2
OMP	orthogonal matching pursuit	p. 2
OSCAR	octagonal shrinkage and clustering algorithm for regression	p. 10
PAVA	pool adjacent violators algorithm	p. 26
RA-ORMP	rank aware order recursive matching pursuit	p. 5
RFS	robust feature selection	p. 6
RPGG	robust projected generalized gradient	p. 9
S-SORR	sparse stable outlier-robust regression	p. 52
SA-MUSIC	subspace augmented MUSIC	p. 5

SCoSaMP	simultaneous compressive sampling matching pursuit	p. 5
SLOPE	sorted ℓ_1 penalized estimation	p. 11
SNIHT	simultaneous normalized hard thresholding	p. 5
SOMP	simultaneous orthogonal matching pursuit	p. 5
SORR	stable outlier-robust regression	p. 7
YALL1	Your Algorithm for L1	p. 9

Functions

CEL0	continuous exact ℓ_0	p. 3
GMC	generalized MC	p. 21
GME	generalized-Moreau-enhanced	p. 21
MC	minimax concave	p. 3
SCAD	smoothly clipped absolute deviation	p. 3

Others

ECG	electrocardiography	p. 8
LiMES	linearly-involved Moreau-enhanced-over-subspace	p. 50
MECG	multi-lead electrocardiogram	p. 6
MMV	multiple measurement vectors	p. 4
MoL-Grad	monotone Lipschitz gradient	p. 85
NMSE	normalized mean squared errors	p. 36
SMV	single measurement vector	p. 4
SNR	signal-to-noise ratio	p. 36
SOR	signal-to-outlier ratio	p. 39

Chapter 1

General Introduction

The goal of this dissertation is to propose a mathematically rigorous framework for debiasing estimation methods of signals. To achieve this, we focus on the so-called Moreau enhancement technique, which has recently been developed extensively in signal processing (see Section 1.1.2). We will explore an effective way of utilizing this technique as well as investigating a new class of operators which extends it. This thesis presents (i) an outlier-robust approach to recover the jointly sparse signals in the presence of outliers, (ii) a sparse stable outlier-robust regression method in the presence of Gaussian noise, and (iii) an effective feature grouping method based on a new mathematical framework of the *external division operator*. This chapter first introduces the background briefly and then motivates this study.

1.1 Background

1.1.1 Sparse Signal Estimation

The analysis of high dimensional data has been in great demand in recent years, and sparseness has been exploited everywhere to deal with the challenge of the possible lack of data samples [1, 2, 3]. Here, a signal is called sparse if most of its components are zero. Although many real-world signals are not truly sparse, they can be well approximated by sparse signals in an appropriate basis [3, 4]. Such signals are referred to as compressible, approximately sparse, relatively sparse, or weakly sparse signals [3, 5]. Sparse signal estimation is formulated as the estimation problem of a sparse coefficient vector $\mathbf{x}_\diamond \in \mathbb{R}^n$ from the observation vector modeled as

$$\mathbf{y} = \mathbf{A}\mathbf{x}_\diamond + \boldsymbol{\varepsilon}_\star \in \mathbb{R}^m, \quad (1.1)$$

where $\mathbf{A} \in \mathbb{R}^{m \times n}$ is the input or measurement matrix and $\boldsymbol{\varepsilon}_\star \in \mathbb{R}^m$ is the white additive zero-mean Gaussian noise. Here, the subscripts \diamond and \star are used for sparse vectors and Gaussian random vectors, respectively.

The sparse signal estimation problem has been studied in many contexts including feature selection (also known as variable selection or support recovery in different literature) and compressive sensing. Feature selection [6, 7, 8] is a problem of selecting an important subset of features from input data to improve interpretability [7]. Once a small number of important features are selected, data analysis techniques such as regression or classification methods can be applied with improved performance and shortened training time [9, 6]. Feature selection plays an important role in dealing with high-dimensional data which frequently occur in many applications. Many existing feature selection methods are based on the estimation of sparse coefficient vectors such as [9, 10, 6, 7, 11]. When $m < n$ as in the case of high dimensional data, the estimation problem is ill-posed, and some prior knowledge about the solution such as sparsity is needed. Such an underdetermined case has been studied extensively in the context of compressive sensing [1, 2, 3]. The theoretical results of compressive sensing enable the recovery of sparse signals with less number of samples needed for the Nyquist-Shannon Theorem [3]. Compressive sensing has made significant impacts on many applications such as magnetic resonance imaging [12], image super-resolution [13], and compressive sensor networks [14].

To estimate the sparse vector \mathbf{x}_\diamond , most standard methods are based on the following problem formulation:

$$\min_{\mathbf{x} \in \mathbb{R}^n} \frac{1}{2} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \mu R(\mathbf{x}), \quad (1.2)$$

where $\|\cdot\|_2$ denotes the ℓ_2 norm (see Section 2.1), $R : \mathbb{R}^n \rightarrow \mathbb{R}$ is the regularizer, and $\mu > 0$ is the regularization parameter. Although the ℓ_0 pseudo-norm (which counts the number of nonzero components) is a natural choice for R , the corresponding problem is known to be NP-hard [15], *i.e.*, there is no efficient algorithm to solve it in polynomial time.

There are mainly two types of approaches to approximate this problem: the greedy and relaxation approaches. The greedy approaches include the matching pursuit (MP) [16], orthogonal matching pursuit (OMP) [17], iterative hard thresholding (IHT) [18], normalized IHT (NIHT) [19], and compressive sensing matching pursuit (CoSaMP) [20]. In general, although the greedy approaches are computationally efficient, there is no guarantee to find the global minimizer. The relaxation approach replaces the ℓ_0 pseudo-norm with another tractable penalty. This thesis mainly focuses on this approach. As a convex surrogate of the ℓ_0 pseudo-norm, ℓ_1 norm has been widely employed as R , in which case the problem in (1.2) has been studied as least absolute shrinkage and selection operator (lasso) [21]. Owing to its convexity, lasso achieves global optimality, *i.e.*, algorithms to solve the problem converge to a global minimizer of the cost function.

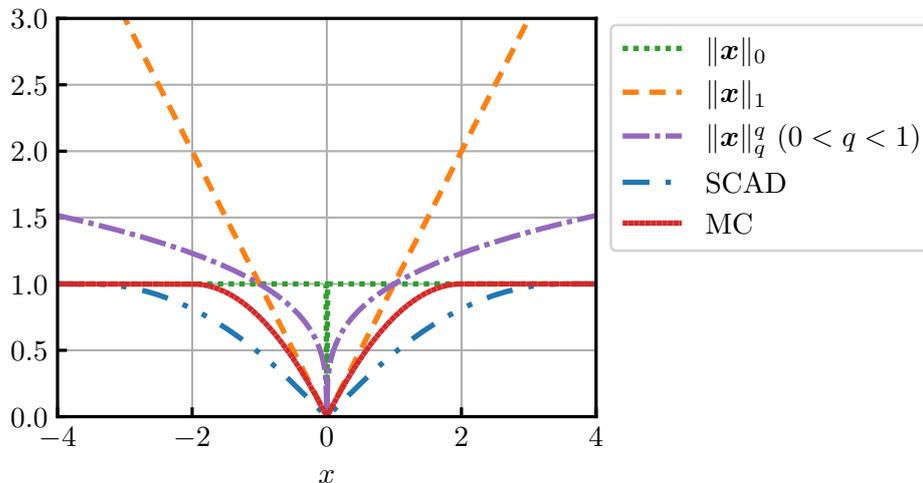


Figure 1.1: Penalty functions for sparse signal estimation.

1.1.2 Bias Reduction in Sparse Signal Estimation

Although the ℓ_1 norm induces the sparsity most effectively in the class of convex regularizers [22], it is known to lead to the estimation bias, *i.e.*, the expectation of the estimates does not meet the true signal [23]. See Section 2.2.1 for details.

To address this issue, numerous nonconvex penalties have been proposed, including the minimax concave (MC) penalty [24, 25], smoothly clipped absolute deviation (SCAD) [26], the ℓ_q quasi-norm for $q \in (0, 1)$ [27, 28, 29], continuous exact ℓ_0 (CEL0) [30], the capped ℓ_1 penalty [31], the logarithmic penalty [32], and Laplace exponential penalty [33] (see Section 2.2 for definitions). Some representative penalty functions for sparse signal estimation are illustrated in Figure 1.1. The outstanding properties of these penalties stem from the fact that they are better approximations of the ℓ_0 pseudo-norm than the ℓ_1 norm. Among those penalties, the MC, SCAD, CEL0, logarithmic, and Laplace exponential penalties are known to be weakly convex functions, *i.e.*, they become convex by adding the squared ℓ_2 norm multiplied by a constant (see also Sections 2.1.2 and 2.2). Owing to this property, the overall convexity of the cost function given in (1.2) can be preserved for the over-determined case ($m \geq n$). Particularly, this thesis focuses on the following MC-based formulation [24]:

$$\min_{\mathbf{x} \in \mathbb{R}^n} \frac{1}{2} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \mu_{\text{MC}} \Phi_\gamma^{\text{MC}}(\mathbf{x}), \quad (1.3)$$

where $\mu_{\text{MC}} > 0$, and $\Phi_\gamma^{\text{MC}} : \mathbb{R}^m \rightarrow [0, +\infty)$ is the MC function of index $\gamma > 0$ (see Section 2.2.2 for definition).

The MC penalty can be expressed as the ℓ_1 norm subtracted by its Moreau envelope (see Section 2.1.4 for definition). A generalization of this concept to any convex seed function instead of the ℓ_1 norm has been extensively developed in recent years as *the (generalized) Moreau enhancement* [25, 34, 35]. The Moreau enhancement has two major useful properties. Firstly, for some examples, the Moreau enhancement bridges the gap between the direct discrete measures and its convex envelope [34]. For example, the MC function parametrically bridges the gap between the ℓ_0 pseudo-norm and the ℓ_1 norm. Secondly, the Moreau enhancement is weakly convex. Owing to this property, the overall convexity of the cost function given in (1.2) can be preserved, and hence the global optimality is guaranteed. See Section 2.2.2 for details. It is worth mentioning here that numerous studies have been conducted on the so-called convex-nonconvex strategy which exploits non-convex penalties while maintaining the overall convexity of the cost function [36, 25, 37]. In this thesis, we will explore the effective use of the Moreau enhancement and its development.

1.1.3 Jointly Sparse Signal Recovery

In many applications, signals have pre-defined specific group structures of sparsity, and exploiting this information is important to obtain a desirable solution [38, 39]. A natural extension of the estimation problem given in (1.1) to address this case has been studied as the jointly sparse signal recovery problem, also known as the multiple measurement vectors (MMV) problem. This problem is formulated as the estimation of jointly sparse coefficient vectors $\{\mathbf{x}_{\diamond,i}\}_{i=1}^r \subset \mathbb{R}^n$, *i.e.*, having nonzero components at common locations, from the observation vectors modeled as

$$\mathbf{y}_i = \mathbf{A}\mathbf{x}_{\diamond,i} + \boldsymbol{\varepsilon}_\star \in \mathbb{R}^m, \quad \forall i = 1, 2, \dots, r, \quad (1.4)$$

i.e.,

$$\mathbf{Y} = \mathbf{A}\mathbf{X}_\diamond + \mathbf{E}_\star \in \mathbb{R}^{m \times r}, \quad (1.5)$$

where $\mathbf{Y} := [\mathbf{y}_1 \mathbf{y}_2 \dots \mathbf{y}_r] \in \mathbb{R}^{m \times r}$, $\mathbf{X}_\diamond := [\mathbf{x}_{\diamond,1} \mathbf{x}_{\diamond,2} \dots \mathbf{x}_{\diamond,r}] \in \mathbb{R}^{n \times r}$, and $\mathbf{E}_\star := [\boldsymbol{\varepsilon}_{\star,1} \boldsymbol{\varepsilon}_{\star,2} \dots \boldsymbol{\varepsilon}_{\star,r}] \in \mathbb{R}^{m \times r}$. Note that \mathbf{X}_\diamond is row sparse¹, *i.e.*, the vector $[\|\mathbf{X}_\diamond\|_{(1,:)}\|_2, \|\mathbf{X}_\diamond\|_{(2,:)}\|_2, \dots, \|\mathbf{X}_\diamond\|_{(n,:)}\|_2]^\top$ is sparse. The MMV problem has been extensively studied with a variety of applications such as array processing [40], spectrum analysis of time series [41], DNA microarrays [42], equalization of sparse communication channels, linear inverse problems [43], magnetoencephalography [44], and source localization in sensor networks [45]. In the particular case of $r = 1$, the MMV problem reduces to the setting of the single measurement vector (SMV) problem, which is an ordinary problem

¹Row sparsity is a special case of group sparsity, which considers an arbitrary group of sparse patterns of a vector of a matrix.

in compressive sensing. While the recovery results of MMV have no advantage to SMV in the worst-case scenario such as the case when the rank of \mathbf{X}_\diamond is one, MMV performs much better than recovering each channel individually in an average-case analysis, which assumes that \mathbf{X}_\diamond is generated at random from an appropriate distribution [46, 47, 48]. For the noiseless case, *i.e.*, $\mathbf{E}_\star = \mathbf{0}_{m \times n}$, a sufficient condition to uniquely determine \mathbf{X}_\diamond from \mathbf{Y} and \mathbf{A} is shown as [49]

$$\|\mathbf{X}_\diamond\|_{2,0} < \frac{\text{spark } \mathbf{A} - 1 + \text{rank } \mathbf{Y}}{2}. \quad (1.6)$$

Here, $\|\cdot\|_{2,0} : \mathbb{R}^{n \times r} \rightarrow \mathbb{R} : \mathbf{X} \mapsto \text{card}(\{i \in \{1, 2, \dots, n\} \mid \mathbf{X}_{(i,:)} \neq \mathbf{0}\})$ denotes the $\ell_{2,0}$ pseudo-norm, $\text{card}(\cdot)$ denotes the cardinality of a set, and the *spark* of \mathbf{A} is defined as the smallest number of columns of \mathbf{A} that are linearly dependent. Since $\text{rank } \mathbf{Y} = 1$ for the SMV problem, it can be seen that the condition for the recovery of \mathbf{X}_\diamond is relaxed in the MMV problem in general.

To estimate \mathbf{X}_\diamond , most standard methods are based on the following formulation:

$$\min_{\mathbf{X} \in \mathbb{R}^{n \times r}} \frac{1}{2} \|\mathbf{Y} - \mathbf{A}\mathbf{X}\|_{\text{F}}^2 + \mu R^{\text{MMV}}(\mathbf{X}), \quad (1.7)$$

where $R^{\text{MMV}} : \mathbb{R}^{n \times r} \rightarrow \mathbb{R}$ is a regularizer. Although a natural choice for R^{MMV} is the $\ell_{2,0}$ pseudo-norm, this problem is known to be NP-hard as well as the ℓ_0 pseudo-norm. Greedy methods for this problem include simultaneous orthogonal matching pursuit (SOMP) [50] (an extension of OMP [17]), simultaneous normalized hard thresholding (SNIHT) [51] (an extension of NIHT [19]), simultaneous compressive sampling matching pursuit (SCoSaMP) [51] (an extension of CoSaMP [20]), rank aware order recursive matching pursuit (RA-ORMP) [48], which extracts the rank information of \mathbf{X}_\diamond . Moreover, combinations of conventional multiple signal classification (MUSIC) algorithm [52] and some recent compressive sensing methods, subspace augmented MUSIC (SA-MUSIC) [53], compressive MUSIC (CS-MUSIC) [54], and semi-supervised MUSIC have been proposed.

When the $\ell_{2,1}$ norm (see Section 2), which is the convex surrogate of the $\ell_{2,0}$ norm, is employed as R^{MMV} , the problem has been studied as group lasso [55, 56]². Although group lasso has been applied successfully to several applications [57, 58, 9, 10, 59], it faces the same issues as lasso, raised in Section 1.1.1. To reduce the estimation bias of $\ell_{2,1}$ norm, nonconvex regularizers to promote group sparsity have been studied such as group SCAD [60], the group MC penalty [61, 62], and $\ell_{p,q}$ norm penalty for $p, q \in (0, 1)$ (see Section 2). In [63], a generalized Moreau enhancement of the $\ell_{2,1}$ norm is studied with the overall convexity.

²Note that group lasso covers the case when $m \geq n$, unlike the MMV problem.

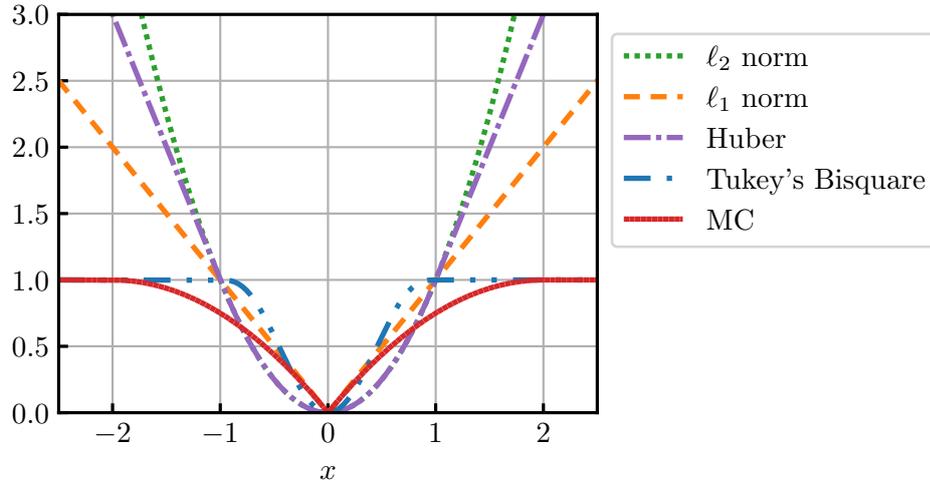


Figure 1.2: Comparison of the squared ℓ_2 norm and robust loss functions.

1.1.4 Outlier-Robust Regression

Outlier is ubiquitous, and robust methods in the presence of outliers have been studied widely [64, 65, 66]. While the formulation based on the quadratic loss function is known to be optimal for Gaussian noise, it is far from optimal when the residual contains non-Gaussian components such as outliers. Here, we assume that the outliers are sparse, and their nonzero values are significantly larger than the signals. Instead of the quadratic loss function, various robust loss functions have been studied in the context of robust statistics. Prominent examples of robust loss functions include Huber's, the least absolute deviation (LAD), and Tukey's bisquare (also called biweight) loss functions [64, 65, 66] (see Figure 1.2). See Section 2.3 for definitions. Here, the LAD is based on the ℓ_1 norm loss function, which is known to be more insensitive to large values than the squared ℓ_2 norm loss function. On the other hand, as explained in Section 2.2, the quadratic loss with the MC penalty can improve the performance of lasso by reducing the estimation bias. A naive question here is the following: *Can the robustness be improved by replacing the LAD with the MC loss function?*

1.1.4.1 Robust Jointly Sparse Signal Recovery

In some applications for jointly sparse signal recovery such as multi-lead electrocardiogram (MECG) [67], gene expression analysis [9], and computer vision [68, 69], the measurements may possibly be contaminated by jointly sparse outliers as

$$\mathbf{Y} = \mathbf{A}\mathbf{X}_\diamond + \mathbf{E}_\star + \mathbf{O}_\diamond \in \mathbb{R}^{m \times r} \quad (1.8)$$

where \mathbf{Y} , \mathbf{X}_\diamond , and \mathbf{E}_\star are defined in the same way as in Section 1.1.3, and $\mathbf{O}_\diamond := [\mathbf{o}_{\diamond,1} \mathbf{o}_{\diamond,2} \dots \mathbf{o}_{\diamond,r}] \in \mathbb{R}^{m \times r}$ is an outlier matrix. Here, \mathbf{X}_\diamond and \mathbf{O}_\diamond are assumed to be row sparse; this assumption is also used in [9, 10, 70]. The problem of robust jointly sparse signal recovery is stated as follows: recover \mathbf{X}_\diamond in (1.8) from the known/measurable matrices \mathbf{A} and \mathbf{Y} (with \mathbf{E}_\star and \mathbf{O}_\diamond unknown). Due to the presence of outliers, the classical regularized least square regression approach defined in (1.7) is known to fail.

To attain robustness against outliers, the robust feature selection (RFS) [9] has been proposed in the context of feature selection [9]:

$$(P_0) \quad \min_{\mathbf{X} \in \mathbb{R}^{n \times r}} \|\mathbf{Y} - \mathbf{A}\mathbf{X}\|_{2,1} + \mu \|\mathbf{X}\|_{2,1}.$$

This formulation can be seen as an extension of the LAD to the MMV problem in (1.9). The $\ell_{2,1}$ norm loss leads to outlier robustness compared to the classical approach. Despite its fast convergence, the computational cost for each iteration is high due to the presence of matrix inversion. Therefore, the scalability of the RFS to high dimensional data is rather limited. This gives rise to the first research question in this thesis:

(Q1) Can we construct an efficient method with remarkable robustness and global optimality for jointly sparse signal recovery based on the Moreau enhancement?

1.1.4.2 Issues Regarding Existing Robust Loss Functions

Although various robust loss functions have been studied, there are two major issues.

Issue (i): Existence of an inevitable tradeoff between robustness and global optimality

While Huber's and the LAD loss functions are convex and tractable, their robustness to large outliers is limited. In contrast, Tukey's loss suffers from the issue of local minima due to nonconvexity while it is highly robust to outliers since its gradient vanishes for large values due to the so-called strong re-descending property. Table 1.1 summarizes the comparisons of robust loss functions.

Let us consider which class of loss functions can break this tradeoff. First, it is known that no convex loss functions can be useful for this purpose. To obtain high robustness for huge outliers, vanishment of the derivative of the loss function above some threshold is important [64, 65]. However, the derivative of a convex function does not vanish because it is monotonically non-increasing. This encourages us to explore a nonconvex loss function. Specifically, the use of the MC function as a loss function is suitable since its derivative vanishes above a certain threshold, and it is mathematically tractable due to its weak convexity.

Issue (ii): Sensitivity against small perturbations

Table 1.1: Comparisons of robust loss functions

	Convexity	Robustness	Mathematical Tractability
LAD	convex	limited	Yes
Huber's loss	convex	limited	Yes
Tukey's loss	nonconvex	Yes	hard to ensure global optimality
MC loss	weakly convex	Yes	Yes

Let us consider the behavior of the loss functions in the vicinity of the origin. Figure 1.3 shows the derivative of representative robust loss functions. It can be seen that, for Huber's and Tukey's loss functions, the derivative vanishes at the origin. This property is important for the estimates to be insensitive to small perturbations, which implies that the statistical differences between outliers and noise are distinguished appropriately. On the other hand, unfortunately, the derivative for sparsity-inducing loss functions such as the LAD and the MC loss do not vanish as they approach the origin. Moreover, the residual, which is nonsparse due to the presence of Gaussian noise, is promoted to be sparse. Hence, the Gaussianity of noise is not reflected appropriately for these losses.

For the case when the coefficient vector to be estimated is assumed to obey i.i.d. zero-mean Gaussian distributions, these two issues have been resolved by the stable outlier-robust regression (SORR) [35], which is based on the MC loss function. The statistical difference between noise and outliers is explicitly distinguished by introducing the auxiliary vector to model the Gaussian noise. See Appendix F.1 for details.

1.1.4.3 Sparse Outlier-Robust Regression

Robustness in sparse signal estimation has been one of the most important aspects to be addressed in signal processing and machine learning over the years [71, 72, 73, 74]. Outliers can cause severe performance degradation in many applications including electrocardiography (ECG), which is always corrupted by electromyographic noise [71], image inpainting corrupted by salt-and-pepper noise [75], wireless sensor network [76], speech denoising [77], and direction of arrival estimation [78]. To obtain a reasonable solution in such a case, a use of an outlier-robust loss function, say $F : \mathbb{R}^m \rightarrow \mathbb{R}$, in place of the quadratic loss in (1.2), has been studied actively:

$$\min_{\mathbf{x} \in \mathbb{R}^n} F(\mathbf{y} - \mathbf{A}\mathbf{x}) + \mu R(\mathbf{x}). \quad (1.9)$$

Table 1.2 summarizes some existing sparse outlier-robust methods. **Issue**

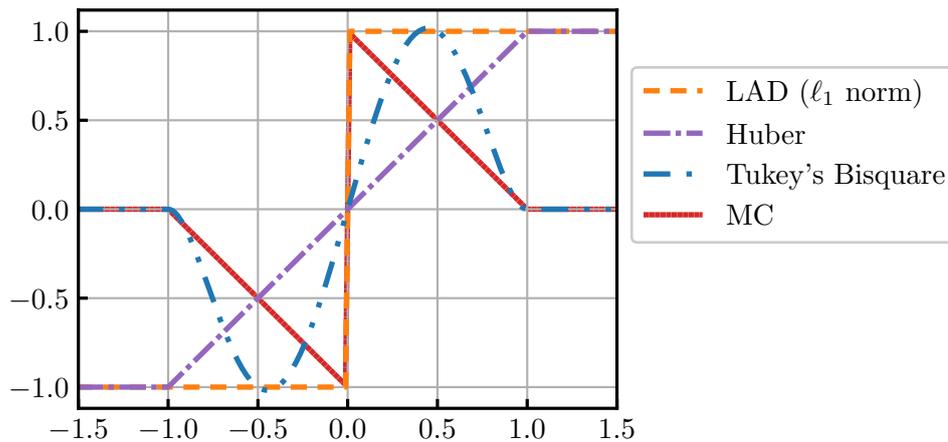


Figure 1.3: Derivative of robust loss functions.

Table 1.2: Comparisons of sparse outlier-robust methods.

	Overall Convexity	Robustness	Mathematical Tractability	Stability under Gaussian noise
YALL1	convex	limited	Yes	No
RPGG	nonconvex	Yes	hard to ensure global optimality	No
Huber FISTA	convex	limited	Yes	Yes
extended lasso	convex	limited	Yes	Yes

(i) is problematic in this case as well. For instance, setting $F = R := \|\cdot\|_1$ in (1.9) produces the formulation of least absolute deviation-lasso (LAD-lasso) [79], which is also studied as Your Algorithm for L1 (YALL1) [80]. Huber fast iterative shrinkage algorithm (FISTA) [81] corresponds to the case when F is Huber's function and R is the ℓ_1 norm. While YALL1 and Huber FISTA are mathematically tractable, robustness is limited. Letting $F := \Phi_{\gamma_1}^{\text{MC}}$ and $R := \Phi_{\gamma_2}^{\text{MC}}$ for $\gamma_1, \gamma_2 > 0$ produces the formulation of the robust projected generalized gradient (RPGG) method [72]. The MC function is far more insensitive to large values than the ℓ_1 norm because it is not affected by huge outliers, as mentioned above. Unfortunately, however, RPGG requires many iterations with a small step size to achieve a small reconstruction error. Moreover, it requires pseudoinverse computation in the initialization step, which may become a computational bottleneck on top of the issue of local minima [72].

As consistent with the previous section, existing approaches for sparse

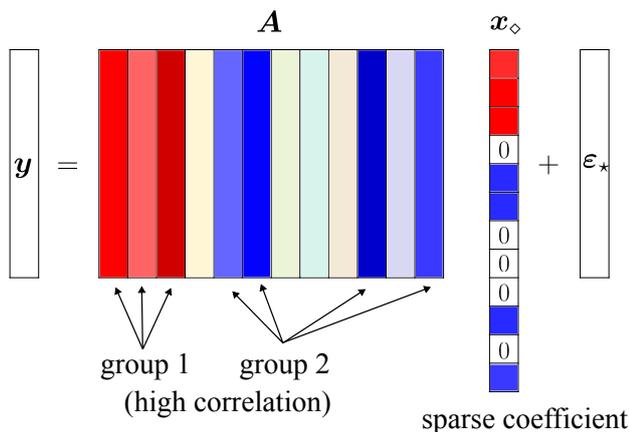


Figure 1.4: An illustration of feature grouping.

outlier-robust regression also suffer from **issue (ii)** in general. For example, YALL1 and RPKG are sensitive to small perturbations. Extended lasso [82] distinguishes the statistical differences between noise and outliers by introducing auxiliary variables to model the outliers into the lasso formulation. Unfortunately, its robustness is limited when the magnitude of outliers is unacceptably large (see Section 4.5). The primitive question here is the following:

(Q2) *Can we extend SORR to develop a sparse signal estimation method with stability in highly noisy environments?*

1.1.5 Feature Grouping

In many applications of sparse signal estimation, there are groups of highly correlated features, which are relevant to observations. In that case, it is known that lasso tends to select only one feature from each group [83]. This is problematic especially when one would like to select important groups of features relevant to observations, as illustrated in Figure 1.4. Such a situation happens in various fields such as gene expression analysis [84], brain imaging [85], and analysis of protein-protein interaction networks [86].

Many feature grouping methods³ have been proposed to yield *the grouping effect*, *i.e.*, select groups of highly correlated features, such as the elastic net [83], the fused lasso [87], the clustered lasso [88], and octagonal shrinkage and clustering algorithm for regression (OSCAR) [89, 90] (see Section 2.4 for the formulations). Unfortunately, the elastic net does not promote

³We note that the goal of feature grouping is different from the jointly sparse signal recovery problem. The former considers the case when there are groups of highly correlated features, while the latter does not assume the correlations of features in general. In addition, for the feature grouping, the group structure of the coefficient vector is not given in advance.

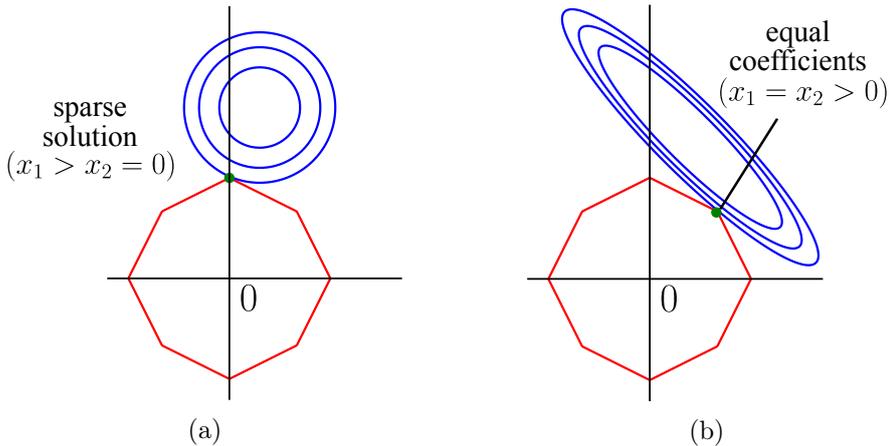


Figure 1.5: Contours of the quadratic loss function centered at the least squares solution (blue) and the constraint region of the OSCAR regularizer (red) in the two-dimensional case. (a) The case when two features are uncorrelated, and (b) when two features are highly correlated.

the equality of coefficients for the highly correlated features in general, which may lead to difficulty in the interpretation of the group structure [89, 90]. Besides, the fused lasso promotes the equality of coefficients only for the successive coefficients, and the clustered lasso does not group the negatively correlated features [90, 91].

In contrast to those methods, OSCAR is free from these limitations. Figure 1.5 shows the contours of the OSCAR regularizer in the two-dimensional case. Here, we assume that \mathbf{y} is centered and \mathbf{A} is standardized so that $\sum_{i=1}^m y_i = 0$, $\sum_{i=1}^m A_{i,j} = 0$, and $\|\mathbf{a}_j\|_2^2 = 1$ for $j = 1, 2, \dots, n$. In this case, the axes of the ellipses of the least squares are tilted 45 degrees from the coordinate axis (see Appendix A for its derivation). When two features are uncorrelated, as depicted in Figure 1.5(a), the contours of the quadratic loss function tend to intersect the constraint region on the vertical axis, yielding a sparse solution. On the other hand, when the two features are highly correlated, as depicted in Figure 1.5(b), the contours tend to intersect the constraint region at the point where $x_1 = x_2 > 0$, grouping the corresponding coefficient pairs. The properties of OSCAR have been extensively investigated with its generalization called the weighted ℓ_1 norm [92, 93] a.k.a. sorted ℓ_1 penalized estimation (SLOPE) in statistics [94] (see also [95]). OSCAR simultaneously encourages sparsity and equality of coefficients for highly correlated features.

Unfortunately, it is known that OSCAR may overpenalize the large pairwise coefficient differences [90, 86], which may cause underestimation. A possible approach would be to consider the Moreau enhancement of the OSCAR regularizer. However, unlike the ℓ_1 norm, no direct discrete measure

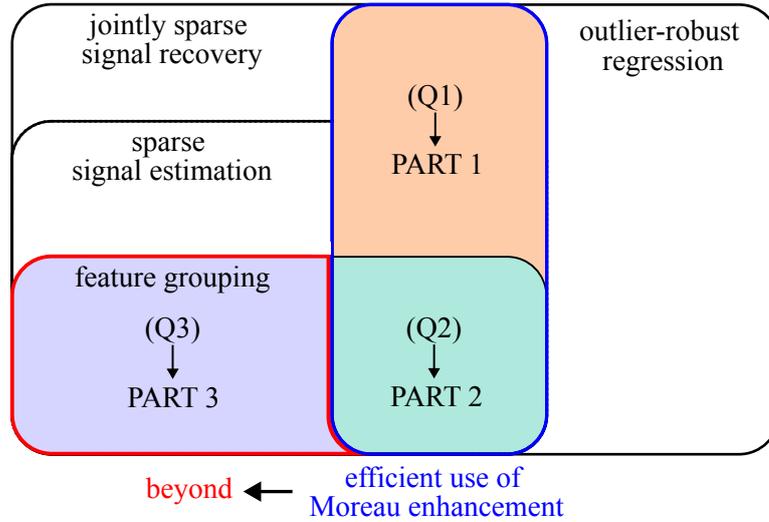


Figure 1.6: The overview of this dissertation.

corresponding to the OSCAR regularizer is known to the best of our knowledge. This implies that its bias reduction effect is unclear. This encourages us to consider the following question:

(Q3) Does the Moreau enhancement of the OSCAR regularizer bridges the OSCAR regularizer and a certain ideal nonconvex function? Moreover, is there any other approach to go beyond the Moreau enhancement?

1.2 This Study

This dissertation addresses the above three questions (Q1), (Q2), and (Q3). A key to answering the above questions must be an idea of the Moreau enhancement, which can reduce estimation bias while global optimality is guaranteed. This thesis explores how to effectively utilize the Moreau enhancement and defines a new class of operators which develops the idea of it. Figure 1.6 shows the overview of this dissertation. The main body is divided into three parts, each of which is devoted to each of the above questions, respectively.

Chapter 2 provides mathematical preliminaries which will be used throughout this thesis.

In Chapter 3, to answer (Q1), we propose a robust method of jointly-sparse signal estimation based on the MC functions (the Moreau enhancement of the ℓ_1 norm). The proposed formulation involves the MC loss, MC penalty, and Tikhonov penalty terms. Since the MC loss function is constant for large absolute errors, it can significantly reduce the influence of outliers. The key difference between the present study from those previ-

ous works presented in Section 1.1.2 is the use of the debiasing function to create a highly-outlier-insensitive loss function. Moreover, our approach enjoys global optimality by using the weak convexity of the MC function. This is the main difference from RPPG, which is based on the MC loss, and does not guarantee the global optimality. Numerical examples including the application of multi-lead electrocardiogram with real data demonstrate the remarkable robustness of the proposed method.

In Chapter 4, to answer (Q2), we integrate the sparse signal estimation method based on the MC-based sparse signal estimation method given in (1.3) and SORR. The statistical differences between Gaussian noise and outliers are appropriately distinguished by introducing an auxiliary vector to model the noise. This leads to accurate estimation even in highly noisy environments. In addition, in analogy to the popular elastic net, the Tikhonov regularizer is used together with the MC function, yielding the grouping effect. In contrast to the elastic net, the grouping effect of the proposed method does not depend on the magnitudes of outliers. Numerical examples show the efficacy of the proposed method even in highly noisy environments.

In Chapter 5, to answer (Q3), we introduce a new notion of “the external division operator”, which extends the idea of the Moreau enhancement, and we present a method to extract all correlated features accurately. The idea of the external division operator comes from the fact that the proximity operator of the MC function can be expressed as “an external division of two proximity operators of the ℓ_1 norm”. The external division operator for OSCAR turns out to be a generalization of its Moreau enhancement. Numerical examples demonstrate that the proposed method improves the performance dramatically by reducing the estimation bias.

Chapter 6 summarizes the results of this thesis and gives an outlook on future research.

Chapter 2

Preliminaries

2.1 Notation and Definitions

For any $\mathbf{z} \in \mathbb{R}^n$ and $p \in (0, +\infty)$, we define

$$\|\mathbf{z}\|_p := \left(\sum_{i=1}^n |z_i|^p \right)^{1/p}, \quad (2.1)$$

which is referred to as the ℓ_p norm for $p \geq 1$ and the ℓ_p quasi-norm for $0 < p < 1$. The ℓ_∞ norm of any $\mathbf{z} \in \mathbb{R}^n$ is defined by

$$\|\mathbf{z}\|_\infty := \max\{|z_1|, |z_2|, \dots, |z_n|\}. \quad (2.2)$$

The $\ell_{2,1}$, $\ell_{2,\infty}$, Frobenius, and nuclear norms of any $\mathbf{A} \in \mathbb{R}^{m \times n}$ are defined by

$$\|\mathbf{A}\|_{2,1} := \sum_{i=1}^m \|\mathbf{A}_{(i,:)}\|_2, \quad (2.3)$$

$$\|\mathbf{A}\|_{2,\infty} := \max\{\|\mathbf{A}_{(1,:)}\|_2, \dots, \|\mathbf{A}_{(m,:)}\|_2\}, \quad (2.4)$$

$$\|\mathbf{A}\|_{\text{F}} := \left(\sum_{i=1}^m \sum_{j=1}^n A_{i,j}^2 \right)^{1/2}, \quad (2.5)$$

$$\|\mathbf{A}\|_{\text{nuc}} := \sum_{i=1}^{\text{rank } \mathbf{A}} s_i(\mathbf{A}), \quad (2.6)$$

respectively. Here, $s_i(\mathbf{A})$ denotes the i th largest singular values of \mathbf{A} for $i = 1, 2, \dots, \text{rank } \mathbf{A}$. The Frobenius inner product of any $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m \times n}$ is defined by

$$\langle \mathbf{A}, \mathbf{B} \rangle_{\text{F}} := \text{Tr}(\mathbf{A}^{\text{T}} \mathbf{B}). \quad (2.7)$$

For any symmetric matrix $\mathbf{A} \in \mathbb{R}^{m \times m}$, the spectral norm of \mathbf{A} is defined by

$$\|\mathbf{A}\|_2 := \max_{\|\boldsymbol{\xi}\|_2=1} \|\mathbf{A}\boldsymbol{\xi}\|_2. \quad (2.8)$$

2.1.1 Bounded Linear Operators

Let $(\mathcal{H}, \langle \cdot, \cdot \rangle)$ be a finite-dimensional real Hilbert space. In this thesis, we consider only Euclidean spaces and spaces of real matrices as examples of finite-dimensional real Hilbert spaces. If a bounded linear operator $L \in \mathcal{B}(\mathcal{H}, \mathcal{H})$ satisfies $\langle Lx, x \rangle_{\mathcal{H}} \geq 0$ ($\langle Lx, x \rangle_{\mathcal{H}} > 0$) for any $x \in \mathcal{H}$, then L is positive semidefinite (positive definite), denoted by $L \succeq O$ ($L \succ O$). For any linear operator $L \in \mathcal{B}(\mathcal{H}, \mathcal{H})$, the adjoint operator $L^* \in \mathcal{B}(\mathcal{H}, \mathcal{H})$ is defined by the operator satisfying

$$\langle L\xi, x \rangle_{\mathcal{H}} = \langle \xi, L^*x \rangle_{\mathcal{H}}, \quad \forall x, \xi \in \mathcal{H}. \quad (2.9)$$

If $L^* = L$, it is called a self-adjoint operator. A square root of a positive semidefinite self-adjoint operator L is denoted by $L^{1/2}$ and is defined by a self-adjoint operator Λ satisfying $\Lambda^2 = L$. The inner product $\langle \cdot, \cdot \rangle_L$ is defined by $\langle x, \xi \rangle_L := \langle Lx, \xi \rangle_{\mathcal{H}}$ for any $x, \xi \in \mathcal{H}$.

2.1.2 Selected Elements of Convex Analysis

Definition 2.1 (Proper function). A function $f : \mathcal{H} \rightarrow (-\infty, +\infty] := \mathbb{R} \cup \{+\infty\}$ is *proper* if

$$\text{dom } f := \{x \in \mathcal{H} \mid f(x) < +\infty\} \neq \emptyset. \quad (2.10)$$

Definition 2.2 (Lower-semicontinuous function). A function $f : \mathcal{H} \rightarrow (-\infty, +\infty]$ is *lower-semicontinuous* on \mathcal{H} if the level set

$$\text{lev}_{\leq a} f := \{x \in \mathcal{H} \mid f(x) \leq a\} \quad (2.11)$$

is closed for any $a \in \mathbb{R}$.

Definition 2.3 (Convex function). A function $f : \mathcal{H} \rightarrow (-\infty, +\infty]$ is *convex* if

$$f(ax + (1-a)\xi) \leq af(x) + (1-a)f(\xi), \quad \forall x, \xi \in \mathcal{H}, \quad \forall a \in (0, 1). \quad (2.12)$$

Definition 2.4 (Strongly and weakly Convex functions). For any $\rho > 0$, $f : \mathcal{H} \rightarrow (-\infty, +\infty]$ is ρ -*strongly convex* if $f - \rho \|\cdot\|_{\mathcal{H}}^2/2$ is convex, and it is ρ -*weakly convex* if $f + \rho \|\cdot\|_{\mathcal{H}}^2/2$ is convex.

Definition 2.5 (Conjugate function). Given a proper function $f : \mathcal{H} \rightarrow (-\infty, +\infty]$, the *conjugate* (or *Fenchel conjugate*, *Legendre Transform*, or *Legendre-Fenchel Transform*) of f is defined by

$$f^* : \mathcal{H} \rightarrow \mathbb{R} \cup \{+\infty, -\infty\} : \xi \mapsto \sup_{x \in \mathcal{H}} (\langle x, \xi \rangle_{\mathcal{H}} - f(x)). \quad (2.13)$$

The following fact holds for a conjugate function.

Fact 2.1 ([96, Proposition 13.23(iv)]). Let $f : \mathcal{H} \rightarrow (-\infty, +\infty]$ and $L \in \mathcal{B}(\mathcal{H}, \mathcal{H})$ be bijective. Then, $(f \circ L)^* = f^* \circ (L^{-1})^*$.

Definition 2.6 (Metric projection). For any nonempty closed convex set $C \subset \mathcal{H}$ and any point $x \in \mathcal{H}$, there exists a unique point $P_C(x) \in C$ satisfying

$$d_C(x) := \min_{\xi \in C} \|x - \xi\|_{\mathcal{H}} = \|x - P_C(x)\|_{\mathcal{H}}. \quad (2.14)$$

The mapping $P_C : x \mapsto P_C(x) \in C$ is called the *metric projection* onto C .

Definition 2.7 (Support and indicator functions). For any nonempty closed convex set $C \subset \mathcal{H}$, the *support function* of C is defined by

$$\Gamma_0(\mathcal{H}) \ni \sigma_C : x \mapsto \sup_{\xi \in C} \langle x, \xi \rangle_{\mathcal{H}}, \quad (2.15)$$

and the *indicator function* is defined by

$$\Gamma_0(\mathcal{H}) \ni \iota_C : x \mapsto \begin{cases} 0, & \text{if } x \in C, \\ +\infty, & \text{if } x \notin C. \end{cases} \quad (2.16)$$

The conjugate of the support function is the indicator function, *i.e.*, $\sigma_C^* = \iota_C$ [96, Example 13.3]. Given any norm $\|\cdot\|$, the dual norm of $\|\cdot\|$ is defined by $\|\cdot\|_* := \sigma_C$ with $C := \{x \in \mathcal{H} \mid \|x\| \leq 1\}$ [97, 98].

Definition 2.8 (Coercive function). A function $f : \mathcal{H} \rightarrow (-\infty, +\infty]$ is *coercive* if

$$\lim_{\|x\|_{\mathcal{H}} \rightarrow +\infty} f(x) = +\infty. \quad (2.17)$$

2.1.3 Nonexpansive and Firmly Nonexpansive Operators

Definition 2.9 (Lipschitz continuous operator). An operator $T : \mathcal{H} \rightarrow \mathcal{H}$ is *β -Lipschitz continuous* with constant $\beta > 0$ if

$$\|Tx - T\xi\|_{\mathcal{H}} \leq \beta \|x - \xi\|_{\mathcal{H}}, \quad \forall (x, \xi) \in \mathcal{H} \times \mathcal{H}. \quad (2.18)$$

For $\beta = 1$, T is called a *nonexpansive operator*.

Definition 2.10 (Firmly nonexpansive operator). An operator $T : \mathcal{H} \rightarrow \mathcal{H}$ is *firmly nonexpansive* if there exists a nonexpansive operator $N : \mathcal{H} \rightarrow \mathcal{H}$ such that

$$T = \frac{1}{2} \text{Id} + \frac{1}{2} N. \quad (2.19)$$

For any $f \in \Gamma_0(\mathcal{H})$, Prox_f is firmly nonexpansive [96, Proposition 12.28]. Although nonexpansive operators are not necessarily firmly nonexpansive, this is true for the gradient of a convex function, as shown in the following fact.

Fact 2.2 (Baillon-Haddad Theorem [99], [96, Corollary 18.17]). Let $f : \mathcal{H} \rightarrow \mathbb{R}$ be a Fréchet differentiable convex function and let $\beta > 0$. Then, ∇f is β -Lipschitz continuous if and only if ∇f is β^{-1} -cocoercive. In particular, ∇f is nonexpansive if and only if ∇f is firmly nonexpansive.

Definition 2.11 (Cocoercive operator). An operator $T : \mathcal{H} \rightarrow \mathcal{H}$ is β -cocoercive for $\beta > 0$ if βT is firmly nonexpansive.

2.1.4 Proximity Operator and Moreau Envelope

Definition 2.12 (Proximity operator). Given any function $f : \mathcal{H} \rightarrow (-\infty, +\infty]$, the *proximity operator* of f of index $\gamma > 0$ (in an extended sense) is defined by

$$\text{Prox}_{\gamma f} : \mathcal{H} \rightarrow \mathcal{H} : x \mapsto \underset{\xi \in \mathcal{H}}{\text{argmin}} \left(f(\xi) + \frac{1}{2\gamma} \|x - \xi\|_{\mathcal{H}}^2 \right), \quad (2.20)$$

which is also denoted as s-Prox, if a unique minimizer exists [100]¹. A function f is called *proximable* (or *prox-friendly*) if Prox_f is easy to compute.

For any nonempty closed convex set $C \subset \mathcal{H}$, the proximity operator of ι_C corresponds to the projection operator P_C [96, Example 12.25].

Definition 2.13 (Moreau envelope). Given a function $f \in \Gamma_0(\mathcal{H})$, the *Moreau envelope* of f of index $\gamma > 0$ is defined by [96]

$$\begin{aligned} \gamma f : \mathcal{H} \rightarrow \mathbb{R} : x \mapsto &= \min_{\xi \in \mathcal{H}} \left(f(\xi) + \frac{1}{2\gamma} \|x - \xi\|_{\mathcal{H}}^2 \right) \\ &= f(\text{Prox}_{\gamma f}(x)) + \frac{1}{2\gamma} \|x - \text{Prox}_{\gamma f}(x)\|_{\mathcal{H}}^2. \end{aligned} \quad (2.21)$$

The following facts are used in this thesis.

Fact 2.3 (Proximity operator [96, Propositions 12.28 and 24.8]). Let $f \in \Gamma_0(\mathcal{H})$. Then, the following hold.

(a) For any $x, \xi \in \mathcal{H}$, it holds that

$$\text{Prox}_{f(\cdot - \xi)}(x) = \xi + \text{Prox}_f(x - \xi). \quad (2.22)$$

(b) The proximity operator Prox_f is firmly nonexpansive.

¹If f is nonconvex, the proximity operator is often defined as a set-valued operator. This paper focuses on the case when the proximity operator is a single-valued operator, and the notation s-Prox is also used to emphasize that.

Fact 2.4 (Moreau envelope [96] and [101, Fact 17.17]). Let $f \in \Gamma_0(\mathcal{H})$ and $\gamma > 0$. Then, the following hold.

- (a) (Convexity) The function γf is convex.
 (b) (Lipschitz continuity) The Moreau envelope is differentiable with the gradient

$$\nabla(\gamma f) = \gamma^{-1}(\text{Id} - \text{Prox}_{\gamma f}), \quad (2.23)$$

which is γ^{-1} -Lipschitz continuous.

- (c) (Moreau's decomposition) It holds that

$$\text{Prox}_{\gamma f} + \gamma \text{Prox}_{\gamma^{-1}f^*} \circ (\gamma^{-1} \text{Id}) = \text{Id}, \quad (2.24)$$

and

$$\frac{1}{2\gamma} \|\cdot\|_{\mathcal{H}}^2 = \gamma f + \gamma^{-1}(f^*) \circ (\gamma^{-1} \text{Id}). \quad (2.25)$$

- (d) (Lower bound)

$$f(x) \geq \gamma f(x), \quad \forall \gamma \in (0, +\infty), \quad \forall x \in \mathcal{H}. \quad (2.26)$$

- (e) (Convergence) (i) The function γf converges pointwise to f on $\text{dom } f$ as $\gamma \rightarrow 0$, *i.e.*,

$$\lim_{\gamma \rightarrow +0} \gamma f(x) = f(x), \quad \forall x \in \mathcal{H}. \quad (2.27)$$

Moreover, if f is uniformly continuous on a bounded set $S \subset \text{dom } f$, γf converges uniformly to f on S , *i.e.*, $\lim_{\gamma \rightarrow +0} \sup_{x \in S} |\gamma f(x) - f(x)| = 0$. In particular, if f is continuous on a compact set $S \subset \text{dom } f$, the *Heine's theorem* [102, Theorem 4.47] guarantees the uniform convergence of γf on S .

- (ii) It holds for any $x \in \mathcal{H}$ that

$$\lim_{\gamma \rightarrow +\infty} \gamma f(x) = \inf_{\xi \in \mathcal{H}} f(\xi). \quad (2.28)$$

2.1.5 Monotone Operator

Definition 2.14 (Graph, domain, and range). For any set-valued operator $B : \mathcal{H} \rightarrow 2^{\mathcal{H}}$, the *graph* of B is denoted by

$$\text{gra } B := \{(x, \xi) \in \mathcal{H} \times \mathcal{H} \mid \xi \in Bx\}, \quad (2.29)$$

the *domain* of B is denoted by

$$\text{dom } B := \{x \in \mathcal{H} \mid Bx \neq \emptyset\}, \quad (2.30)$$

and the *range* of B is denoted by

$$\text{range } B := \{\xi \in \mathcal{H} \mid \exists x \in \mathcal{H}, \text{ s.t. } \xi \in Bx\}. \quad (2.31)$$

Definition 2.15 (Monotone operator). A set-valued operator $B : \mathcal{H} \rightarrow 2^{\mathcal{H}}$ is *monotone* if

$$\langle x - \xi, u - v \rangle_{\mathcal{H}} \geq 0, \quad \forall (x, u) \in \text{gra } B, \quad \forall (\xi, v) \in \text{gra } B. \quad (2.32)$$

The β -cocoercive operator T for $\beta > 0$ satisfies the following inequality:

$$\langle Tx - T\xi, x - \xi \rangle_{\mathcal{H}} \geq \frac{\beta}{2} \|Tx - T\xi\|_{\mathcal{H}}^2, \quad (2.33)$$

and hence T is monotone.

Definition 2.16 (Subdifferential). Given a proper function $f : \mathcal{H} \rightarrow (-\infty, +\infty]$, the *subdifferential* of f is denoted by

$$\begin{aligned} \partial f : \mathcal{H} &\rightarrow 2^{\mathcal{H}} \\ : x &\mapsto \{z \in \mathcal{H} \mid \langle \xi - x, z \rangle_{\mathcal{H}} + f(z) \leq f(\xi), \quad \forall \xi \in \mathcal{H}\}, \end{aligned} \quad (2.34)$$

and, for any $x \in \mathcal{H}$, an element of $\partial f(x)$ is a *subgradient* of f at x .

For any $f \in \Gamma_0(\mathcal{H})$, it is well known that the proximity operator is expressed as

$$\text{Prox}_f = (\text{Id} + \partial f)^{-1}, \quad (2.35)$$

which is the *resolvent*² of ∂f [96, Proposition 16.44].

2.2 Bias Reduction Methods

2.2.1 Estimation Bias of Lasso

Unbiasedness for a large true coefficient is one of the principles proposed in [26] to select a good penalty function. Here, an estimator is *biased* if its expectation does not meet the true coefficient, *i.e.*, $\mathbb{E}[\hat{\mathbf{x}}] \neq \mathbf{x}_o$; otherwise it is called *unbiased* [23]. Assume that \mathbf{A} in (1.1) is orthonormal. In this case, Problem in (1.2) is equivalent to

$$\min_{\mathbf{x} \in \mathbb{R}^n} \frac{1}{2} \|\mathbf{A}^{\top} \mathbf{y} - \mathbf{x}\|_2^2 + \mu R(\mathbf{x}). \quad (2.36)$$

The solution to (2.36) for $R := \|\cdot\|_1$ is given by

$$\begin{aligned} \hat{\mathbf{x}} &= \text{Prox}_{\mu \|\cdot\|_1}(\mathbf{A}^{\top} \mathbf{y}) \\ &= \text{Soft}_{\mu}(\mathbf{A}^{\top} \mathbf{y}). \end{aligned} \quad (2.37)$$

²The resolvent of a set-valued operator $B : \mathcal{H} \rightarrow 2^{\mathcal{H}}$ is defined by $(\text{Id} + B)^{-1}$.

Here, for any $\gamma > 0$, the soft-shrinkage operator [103] is defined by

$$\begin{aligned} \text{Soft}_\gamma &:= \text{Prox}_{\gamma\|\cdot\|_1} : \mathbb{R}^n \rightarrow \mathbb{R}^n : \mathbf{x} := [x_1, x_2, \dots, x_n]^\top \\ &\mapsto [\text{soft}_\gamma(x_1), \text{soft}_\gamma(x_2), \dots, \text{soft}_\gamma(x_n)]^\top \end{aligned} \quad (2.38)$$

with

$$\text{soft}_\gamma : \mathbb{R} \rightarrow \mathbb{R} : x \mapsto \text{sign}(x) \max\{|x| - \gamma, 0\}, \quad (2.39)$$

where

$$\text{sign} : \mathbb{R} \rightarrow \{-1, 1\} : a \mapsto \begin{cases} 1, & \text{if } a \geq 0, \\ -1, & \text{if } a < 0. \end{cases} \quad (2.40)$$

Since $\text{soft}_\mu(\mathbf{a}_i^\top \mathbf{y}) = \mathbf{a}_i^\top \mathbf{y} - \mu$ if $\mathbf{a}_i^\top \mathbf{y} > \mu$, the lasso solution is biased, *i.e.*,

$$\mathbb{E}[\hat{x}_i] = x_{\diamond, i} - \mu \neq x_{\diamond, i}, \quad (2.41)$$

In general, when R is separable, *i.e.*, $R(\mathbf{x}) = \sum_{i=1}^n \rho_i(x_i)$ for any $\mathbf{x} \in \mathbb{R}^n$, \hat{x}_i satisfies the first order optimality conditions

$$\hat{x}_i - \mathbf{a}_i^\top \mathbf{y} + \mu \rho'_i(\hat{x}_i) = 0, \quad \forall i = 1, 2, \dots, n, \quad (2.42)$$

which are obtained by taking the first derivative of the cost function (2.36) with respect to x_i . Hence, if ρ'_i vanishes for large absolute values, it holds that

$$\mathbb{E}[\hat{x}_i] = \mathbf{a}_i^\top \mathbf{y} = x_{\diamond, i} \quad (2.43)$$

when $|\mathbf{a}_i^\top \mathbf{y}|$ is sufficiently large. This implies that the condition $\rho'_i(x_i) = 0$ for large $|x_i|$ is a sufficient condition for unbiasedness [26].

2.2.2 MC Function and Moreau Enhancement

Definition 2.17 (Minimax concave function). The MC function [24] with index $\gamma > 0$ is defined by

$$\Phi_\gamma^{\text{MC}} : \mathbb{R}^n \rightarrow [0, +\infty) : \mathbf{x} \mapsto \sum_{i=1}^n \phi_\gamma^{\text{MC}}(x_i), \quad (2.44)$$

where

$$\phi_\gamma^{\text{MC}} : \mathbb{R} \rightarrow [0, +\infty) : x \mapsto \begin{cases} |x| - \frac{x^2}{2\gamma}, & \text{if } |x| \leq \gamma, \\ \frac{\gamma}{2}, & \text{if } |x| > \gamma. \end{cases} \quad (2.45)$$

This function has been used as a penalty of least square problems for promoting sparsity without causing severe extra biases. It is actually a γ^{-1} -weakly convex function, and enjoys a better sparsity-seeking property than

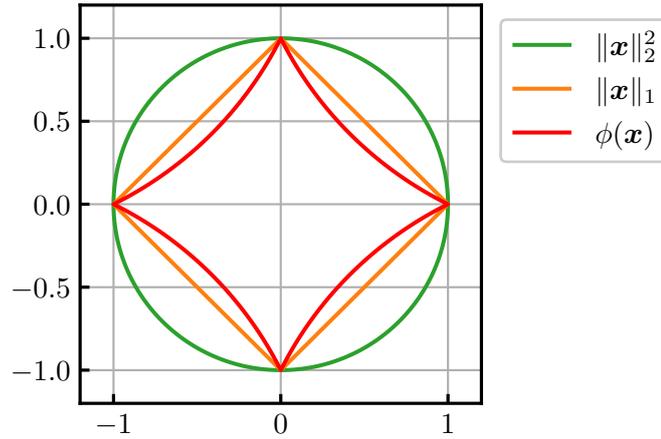


Figure 2.1: Contours of the ℓ_2 norm, the ℓ_1 norm, and the MC function.

the ℓ_1 norm (see Figure 2.1). Furthermore, it is less sensitive to large components than the ℓ_1 norm because the range of the function is bounded above by a certain level (see Figure 1.1).

The MC function can be expressed as the difference between the ℓ_1 norm and its Moreau envelope [25]:

$$\Phi_\gamma^{\text{MC}} = \|\cdot\|_1 - \gamma \|\cdot\|_1. \quad (2.46)$$

A generalization of the MC function to a nonseparable function is proposed as the generalized MC (GMC) function [25]. For any $\mathbf{B} \in \mathbb{R}^{m \times n}$, the GMC function is defined by

$$\begin{aligned} \Phi_{\mathbf{B}}^{\text{GMC}} : \mathbb{R}^n &\rightarrow [0, +\infty) \\ &: \mathbf{x} \mapsto \|\mathbf{x}\|_1 - \min_{\mathbf{z} \in \mathbb{R}^n} \left(\|\mathbf{x}\|_1 + \frac{1}{2} \|\mathbf{B}(\mathbf{x} - \mathbf{z})\|_2^2 \right). \end{aligned} \quad (2.47)$$

When the MC function is employed as R in (1.2), the convexity of the whole cost function cannot be preserved for the underdetermined case in general. In fact, it is known that a nonconvex penalty R needs to be nonseparable in general to preserve the convexity of the overall convexity [25]. Hence, the GMC function extends the applicability of the MC function owing to its nonseparability.

To extend the GMC function to a general convex function, the generalized-Moreau-enhanced (GME) penalty function has been proposed in [34]³.

³In [34], a more general penalty function than GME is proposed, which is applicable to broader scenarios.

Definition 2.18 (Generalized Moreau enhancement). Let $(\mathcal{H}, \langle \cdot, \cdot \rangle)$ and $(\mathcal{K}, \langle \cdot, \cdot \rangle)$ be the finite-dimensional real Hilbert spaces. For any $\Psi \in \Gamma_0(\mathcal{H})$ coercive with $\text{dom } \Psi = \mathcal{H}$, and an arbitrary bounded linear operator $B : \mathcal{H} \rightarrow \mathcal{K}$, the GME penalty function is defined by

$$\Psi_B : x \mapsto \Psi(x) - \min_{z \in \mathcal{H}} \left(\Psi(z) + \frac{1}{2} \|B(x - z)\|_{\mathcal{K}}^2 \right). \quad (2.48)$$

Letting $\mathcal{H} := \mathbb{R}^n$, $\Psi := \|\cdot\|_1$, and $B := \gamma^{-1/2} \mathbf{I}_n$ in (2.48) reduces to the MC function (2.45).

The Moreau enhancement has two major useful properties. Firstly, for some examples, the Moreau enhancement bridges the gap between the direct discrete measures and its convex envelope [34] as shown in the following fact.

Fact 2.5 ([34, Example 2]). 1. It holds for any $\mathbf{x} \in \mathbb{R}^n$ and $\gamma > 0$ that

$$\lim_{\gamma \rightarrow +0} \frac{2}{\gamma} (\|\cdot\|_1)_{\gamma^{-1/2} \text{Id}}(\mathbf{x}) = \|\mathbf{x}\|_0. \quad (2.49)$$

2. It holds for any $\mathbf{X} \in \mathbb{R}^{m \times n}$ and $\gamma > 0$ that

$$\lim_{\gamma \rightarrow +0} \frac{2}{\gamma} (\|\cdot\|_{\text{nuc}})_{\gamma^{-1/2} \text{Id}}(\mathbf{X}) = \text{rank } \mathbf{X}. \quad (2.50)$$

Secondly, the Moreau enhancement is weakly convex. More generally, the following fact shows the overall convexity conditions of the cost function based on the quadratic loss with the generalized Moreau enhancement.

Fact 2.6 ([34, Proposition 1]). Let \mathcal{X} , \mathcal{Y} , \mathcal{Z} and $B \in \mathcal{B}(\mathcal{Z}, \mathcal{Z})$ be the finite-dimensional real Hilbert spaces. Let $\Psi_B \in \Gamma_0(\mathcal{Z})$ be the GME function defined in Definition 2.18. For $\mathcal{L} \in \mathcal{B}(\mathcal{X}, \mathcal{Z})$, $(A, \mathcal{L}) \in \mathcal{B}(\mathcal{X}, \mathcal{Y}) \times \mathcal{B}(\mathcal{X}, \mathcal{Z})$, and $\mu > 0$, let

$$J_{\Psi_B \circ \mathcal{L}} : \mathcal{X} \rightarrow \mathbb{R} : x \mapsto \frac{1}{2} \|y - Ax\|_{\mathcal{Y}}^2 + \mu \Psi_B \circ \mathcal{L}(x). \quad (2.51)$$

Then, for the three conditions

$$(C1) \quad A^* A - \mu \mathcal{L}^* B^* B \mathcal{L} \succeq 0,$$

$$(C2) \quad J_{\Psi_B \circ \mathcal{L}} \in \Gamma_0(\mathcal{X}) \text{ for any } y \in \mathcal{Y},$$

$$(C3) \quad J_{\Psi_B \circ \mathcal{L}}^{(0)} := \frac{1}{2} \|A \cdot\|_{\mathcal{Y}}^2 + \mu \Psi_B \circ \mathcal{L} \in \Gamma_0(\mathcal{X}),$$

the relation $(C1) \Rightarrow (C2) \Leftrightarrow (C3)$ holds.

2.2.3 Other Nonconvex Sparsity-Inducing Penalties for Bias Reduction Methods

We list below some representative penalties for bias reduction introduced in Section 1.1.2. See also [104].

- **SCAD** [26]:

$$\phi_{\mu,\gamma}^{\text{SCAD}} : \mathbb{R} \rightarrow [0, +\infty) : x \mapsto \begin{cases} \mu|x|, & \text{if } |x| < \mu, \\ -\frac{x^2 - 2\gamma\mu|x| + \mu^2}{2(\gamma - 1)}, & \text{if } |x| \in [\mu, \gamma\mu), \\ \frac{(\gamma + 1)\mu^2}{2}, & \text{if } |x| \geq \gamma\mu, \end{cases} \quad (2.52)$$

where $\mu > 0$ and $\gamma > 2$. This penalty is $(\gamma - 1)$ -weakly convex function [105].

- ℓ_q **quasi-norm** for $q \in (0, 1)$ [27, 28, 29]:

$$\phi^{\ell_q} : \mathbb{R} \rightarrow [0, +\infty) : x \mapsto (1/q)|x|^q. \quad (2.53)$$

This penalty is not weakly convex (see Appendix B).

- **CELO** [30]:

$$\phi_{\gamma,\mu}^{\text{CELO}} : \mathbb{R} \rightarrow [0, +\infty) : x \mapsto \begin{cases} \mu - \frac{\gamma^2}{2} \left(|x| - \frac{\sqrt{2\mu}}{\gamma} \right)^2, & \text{if } |x| \leq \frac{\sqrt{2\mu}}{\gamma}, \\ \mu, & \text{if } |x| > \frac{\sqrt{2\mu}}{\gamma}, \end{cases} \quad (2.54)$$

where $\gamma, \mu > 0$. This penalty is γ^2 -weakly convex (see Appendix B).

- **Capped ℓ_1** [31]:

$$\phi_{\gamma}^{\text{cap}} : \mathbb{R} \rightarrow [0, +\infty) : x \mapsto \min\{\gamma, |x|\}, \quad (2.55)$$

where $\gamma > 0$. This penalty is not weakly convex (see Appendix B).

- **Logarithm penalty** [32]:

$$\phi_{\gamma}^{\text{log}} : \mathbb{R} \rightarrow [0, +\infty) : x \mapsto \log(1 + \gamma|x|), \quad (2.56)$$

where $\gamma > 0$. This penalty is $(\gamma/2)$ -weakly convex [106].

- **Laplace exponential penalty** [33]:

$$\phi_{\gamma}^{\text{exp}} : \mathbb{R} \rightarrow [0, +\infty) : x \mapsto (1 - \exp^{-\gamma|x|}), \quad (2.57)$$

where $\gamma > 0$. This penalty is $(\gamma/2)$ -weakly convex [106].

2.3 Outlier-Robust Regeression

We list below representative robust loss functions including those introduced in Section 1.1.4.

- **Huber's function** [64, 65, 66]:

$$\phi_\gamma^{\text{Huber}} : \mathbb{R} \rightarrow [0, +\infty) : x \mapsto \begin{cases} \frac{x^2}{2}, & \text{if } |x| \leq \gamma, \\ -\frac{\gamma^2}{2} + \gamma|x|, & \text{if } |x| > \gamma, \end{cases} \quad (2.58)$$

where $\gamma > 0$.

- **LAD** [64, 65, 66]:

$$\phi^{\text{LAD}} : \mathbb{R} \rightarrow [0, +\infty) : x \mapsto |x|. \quad (2.59)$$

- **Tukey's bisquare (also known as Tukey's biweight)** [64, 65, 66]:

$$\phi_\gamma^{\text{Tukey}} : \mathbb{R} \rightarrow [0, +\infty) : x \mapsto \begin{cases} \frac{1}{6} \left(1 - \left(1 - \frac{|x|^2}{\gamma^2} \right)^3 \right), & \text{if } |x| \leq \gamma, \\ \frac{1}{6}, & \text{if } |x| > \gamma, \end{cases} \quad (2.60)$$

where $\gamma > 0$.

- **Hampel's function** [107]:

$$\phi_{\gamma_1, \gamma_2}^{\text{Hampel}} : \mathbb{R} \rightarrow [0, +\infty) : x \mapsto \begin{cases} \frac{x^2}{2}, & \text{if } |x| \leq \gamma_1, \\ -\gamma_1|x| - \frac{\gamma_1^2}{2}, & \text{if } |x| \in (\gamma_1, \gamma_2], \\ \gamma_1\gamma_2 - \frac{\gamma_1^2}{2} + (\gamma_3 - \gamma_2)\frac{\gamma_1}{2} \left(1 - \left(\frac{\gamma_3 - |x|}{\gamma_3 - \gamma_2} \right)^2 \right), & \text{if } |x| \in (\gamma_2, \gamma_3], \\ \gamma_1\gamma_2 - \frac{\gamma_1^2}{2} + (\gamma_3 - \gamma_2), & \text{if } |x| > \gamma_3, \end{cases} \quad (2.61)$$

where $\gamma > 0$.

- **Lorentzian norm**⁴ [71]:

$$\phi_\gamma^{\text{Lorentzian}} : \mathbb{R} \rightarrow [0, +\infty) : x \mapsto \log \left(1 + \frac{x^2}{\gamma} \right), \quad (2.62)$$

where $\gamma > 0$.

⁴The Lorentzian norm is not a norm since it does not satisfy the positive homogeneity and triangle inequality.

2.4 Feature Grouping Methods

We list below the formulations of representative feature grouping methods introduced in Section 1.1.5.

- **Elastic net** [83]:

$$\min_{\mathbf{x} \in \mathbb{R}^n} \frac{1}{2} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \mu_1 \|\mathbf{x}\|_1 + \frac{\mu_2}{2} \|\mathbf{x}\|_2^2, \quad (2.63)$$

where $\mu_1, \mu_2 > 0$.

- **Fused lasso** [87]:

$$\min_{\mathbf{x} \in \mathbb{R}^n} \frac{1}{2} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \mu_1 \|\mathbf{x}\|_1 + \mu_2 \sum_{i=2}^n |x_i - x_{i-1}|, \quad (2.64)$$

where $\mu_1, \mu_2 > 0$.

- **Clustered lasso** [88]:

$$\min_{\mathbf{x} \in \mathbb{R}^n} \frac{1}{2} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \mu_1 \|\mathbf{x}\|_1 + \mu_2 \sum_{i < j} |x_i - x_j|, \quad (2.65)$$

where $\mu_1, \mu_2 > 0$.

- **OSCAR** [89, 90]:

$$\min_{\mathbf{x} \in \mathbb{R}^n} \frac{1}{2} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \Omega_{\lambda_1, \lambda_2}^{\text{OSCAR}}(\mathbf{x}), \quad (2.66)$$

where $\lambda_1, \lambda_2 > 0$, and

$$\Omega_{\lambda_1, \lambda_2}^{\text{OSCAR}} : \mathbb{R}^n \rightarrow [0, +\infty) : \mathbf{x} \mapsto \lambda_1 \|\mathbf{x}\|_1 + \lambda_2 \sum_{i < j} \max\{|x_i|, |x_j|\}. \quad (2.67)$$

The function $\Omega_{\lambda_1, \lambda_2}^{\text{OSCAR}}$ can also be expressed as

$$\Omega_{\lambda_1, \lambda_2}^{\text{OSCAR}} : \mathbb{R}^n \rightarrow [0, +\infty) : \mathbf{x} \mapsto \sum_{i=1}^n (\lambda_1 + (n-i)\lambda_2) |x|_{[i]}, \quad (2.68)$$

where $|x|_{[i]}$ is the i th largest component of $[|x_1|, |x_2|, \dots, |x_n|]^\top$ so that $|x|_{[1]} \geq |x|_{[2]} \geq \dots \geq |x|_{[n]}$ is satisfied. Let $\mathbf{w} := [w_1, w_2, \dots, w_n]^\top$ such that

$$w_i = \lambda_1 + \lambda_2(n-i), \quad \forall i = 1, 2, \dots, n. \quad (2.69)$$

Then, OSCAR is expressed as the ordered weighted ℓ_1 norm [92], *i.e.*,

$$\Omega_{\lambda_1, \lambda_2}^{\text{OSCAR}}(\mathbf{x}) = \langle \mathbf{w}, |\mathbf{x}|_{\downarrow} \rangle_2, \quad (2.70)$$

where

$$|\mathbf{x}|_{\downarrow} := \mathbf{P}(|\mathbf{x}|)|\mathbf{x}| \in \mathbb{R}^n, \quad (2.71)$$

and $\mathbf{P}(|\mathbf{x}|) \in \mathbb{R}^{n \times n}$ denotes a permutation matrix which sorts the components of $|\mathbf{x}| := [|x_1|, |x_2|, \dots, |x_n|]^T \in \mathbb{R}^n$ in non-increasing order [92]. The proximity operator of OSCAR for $\gamma > 0$ is given by [92]:

$$\text{Prox}_{\gamma \Omega_{\lambda_1, \lambda_2}^{\text{OSCAR}}} : \mathbf{x} \mapsto \text{Sign}(\mathbf{x}) \odot \mathbf{P}(|\mathbf{x}|)^T P_{\mathcal{K}_{\geq 0}^n} (|\mathbf{x}|_{\downarrow} - \gamma \mathbf{w}), \quad (2.72)$$

where

$$\text{Sign} : \mathbb{R}^n \rightarrow \mathbb{R}^n : \mathbf{x} \mapsto [\text{sign}(x_1), \text{sign}(x_2), \dots, \text{sign}(x_n)]^T, \quad (2.73)$$

and

$$\mathcal{K}_{\geq 0}^n := \{\mathbf{x} \in \mathbb{R}^n \mid x_1 \geq x_2 \geq \dots \geq x_n \geq 0\} \quad (2.74)$$

is the monotone nonnegative cone [98].

The projection of any $\mathbf{x} \in \mathbb{R}^n$ onto $\mathcal{K}_{\geq 0}^n$ can be effectively computed by pool adjacent violators algorithm (PAVA) [108, 109, 110].

2.5 Firm-Shrinkage Operator

For any γ, τ satisfying $\gamma > \tau > 0$, the firm-shrinkage operator [111] is defined by

$$\begin{aligned} \text{Firm}_{\tau, \gamma} : \mathbb{R}^n \rightarrow \mathbb{R}^n : \mathbf{x} &:= [x_1, x_2, \dots, x_n]^T \\ &\mapsto [\text{firm}_{\tau, \gamma}(x_1), \text{firm}_{\tau, \gamma}(x_2), \dots, \text{firm}_{\tau, \gamma}(x_n)]^T \end{aligned} \quad (2.75)$$

with

$$\begin{aligned} \text{firm}_{\tau, \gamma} : \mathbb{R} &\rightarrow \mathbb{R} \\ : x &\mapsto \begin{cases} 0, & \text{if } |x| < \tau, \\ \frac{\gamma \text{sign}(x)}{\gamma - \tau} (|x| - \tau), & \text{if } |x| \in [\tau, \gamma), \\ x, & \text{if } |x| \geq \gamma. \end{cases} \end{aligned} \quad (2.76)$$

The firm-shrinkage operator is known as the proximity operator of the MC function as $\text{Firm}_{\tau, \gamma} = \text{Prox}_{\tau \Phi_{\gamma}^{\text{MC}}}$ [112]. The relation between the firm-shrinkage and soft-shrinkage operators will be investigated in Chapter 5.

Chapter 3

Robust Recovery of Jointly-Sparse Signals Based on Minimax Concave Functions

3.1 Introduction

In this chapter, to answer (Q1) raised in Chapter 1.2, we cast the task of robust recovery of jointly sparse signals as a minimization problem of a sum of a weakly convex loss function and a strongly convex regularizer (which ensures the convexity of the whole cost). The use of nonconvex loss function actually makes the optimization problem difficult to solve directly by the convex optimization methods such as the alternating direction method of multipliers (ADMM) [113] and the primal-dual splitting methods [114, 115, 116], even with the well-established firm-shrinkage operator [112] (see Section 3.2.2). To circumvent this difficulty, we reformulate the problem with the Moreau decomposition so that the problem can be solved by the primal-dual splitting method [116]. We derive the parameter designs/ranges to ensure convergence. Numerical examples show that the proposed approach is far more robust against strong outliers compared to the existing $\ell_{2,1}$ -based RFS approach. It is also shown that the proposed approach is effective in an application to the recovery of MECG signals, which is known to be typically corrupted by impulsive electromyographic noise [71].

3.2 Proposed Approach

We first show the proposed problem formulation for robust recovery of jointly-sparse signals, which can be solved by the primal-dual splitting method under an appropriate reformulation, and then present closed-form expressions of the operators used in the algorithm. We finally present the convergence analysis

and discuss the computational complexity.

3.2.1 Problem Formulation

We formulate the robust jointly-sparse signal recovery problem as follows:

$$(P_1) \quad \min_{\mathbf{X} \in \mathbb{R}^{n \times r}} J(\mathbf{X}) := \left(\Phi_{2,1,\mathbf{L}}^{\text{MC}}(\mathbf{Y} - \mathbf{A}\mathbf{X}) + \mu_1 \Phi_{2,1,\mathbf{M}}^{\text{MC}}(\mathbf{X}) + \frac{\mu_2}{2} \|\mathbf{X}\|_{\text{F}}^2 \right),$$

where $\mu_1 \geq 0$, $\mu_2 \geq 0$, and $\Phi_{2,1,\mathbf{L}}^{\text{MC}} : \mathbb{R}^{m \times r} \rightarrow [0, +\infty)$ and $\Phi_{2,1,\mathbf{M}}^{\text{MC}} : \mathbb{R}^{n \times r} \rightarrow [0, +\infty)$ are the Moreau enhancement (see (2.48)) of the $\ell_{2,1}$ norms for the diagonal operators $\mathbf{L} := \text{diag}(l_1, l_2, \dots, l_m) \in \mathbb{R}^{m \times m}$ and $\mathbf{M} := \text{diag}(m_1, m_2, \dots, m_n) \in \mathbb{R}^{n \times n}$ for $l_1, l_2, \dots, l_m, m_1, m_2, \dots, m_n > 0$, respectively, defined as follows¹:

$$\Phi_{2,1,\mathbf{L}}^{\text{MC}}(\mathbf{Z}) := (\|\cdot\|_{2,1})_{\mathbf{L}}(\mathbf{Z}) = \sum_{i=1}^m \Phi_{l_i^{-1}}^{\text{MC}}(\mathbf{Z}_{(i,:)}), \quad \forall \mathbf{Z} \in \mathbb{R}^{m \times r}, \quad (3.1)$$

$$\Phi_{2,1,\mathbf{M}}^{\text{MC}}(\mathbf{\Xi}) := (\|\cdot\|_{2,1})_{\mathbf{M}}(\mathbf{\Xi}) = \sum_{i=1}^n \Phi_{m_i^{-1}}^{\text{MC}}(\mathbf{\Xi}_{(i,:)}), \quad \forall \mathbf{\Xi} \in \mathbb{R}^{n \times r}. \quad (3.2)$$

For $\mu_2 = 0$, (P₁) with $l_1, l_2, \dots, l_m, m_1, m_2, \dots, m_n \rightarrow +0$ reduces to (P₀). Since the MC functions used in both loss and penalty terms in (P₁) are nonconvex, the third term is necessary to obtain the convexity of the whole cost function (see Proposition 3.1 below and its following discussions). Indeed, the use of $\Phi_{2,1,\mathbf{L}}^{\text{MC}}(\mathbf{Y} - \mathbf{A}\mathbf{X})$ makes the outliers less important than using the $\ell_{2,1}$ norm as shown in Figure 1.1.

Some readers may consider that the third term would annihilate the benefits of the MC terms. Fortunately, however, such annihilation does not happen to the first term $\Phi_{2,1,\mathbf{L}}^{\text{MC}}(\mathbf{Y} - \mathbf{A}\mathbf{X})$ when μ_2 is set to an appropriate value according to Proposition 3.1 presented below. This is because the minimization of $\Phi_{2,1,\mathbf{L}}^{\text{MC}}(\mathbf{Y} - \mathbf{A}\mathbf{X}) + \mu_1 \Phi_{2,1,\mathbf{M}}^{\text{MC}}(\mathbf{X}) + \mu_2 \|\mathbf{X}\|_{\text{F}}^2/2$ is closely related to the minimization of $\Phi_{2,1,\mathbf{L}}^{\text{MC}}(\mathbf{B}) + \mu_1 \Phi_{2,1,\mathbf{M}}^{\text{MC}}(\mathbf{A}^\dagger \mathbf{Y} - \mathbf{A}^\dagger \mathbf{B}) + \mu_2 \|\mathbf{A}^\dagger \mathbf{Y} - \mathbf{A}^\dagger \mathbf{B}\|_{\text{F}}^2/2$ by relating \mathbf{X} and \mathbf{Y} by $\mathbf{B} = \mathbf{Y} - \mathbf{A}\mathbf{X}$ if $\mathbf{A}^\dagger \mathbf{A} = \mathbf{I}_n$. Here, the latter problem for $\mu_1 = 0$ is the ordinary least square cost penalized by the MC penalty, and this suggests that the former problem benefits from the MC function as well as the latter one. The other MC term $\Phi_{2,1,\mathbf{M}}^{\text{MC}}(\mathbf{X})$ together with the quadratic term can be considered to be a nonconvex (and block-structured) analog of the popular elastic net regularizer [83], which will be studied further in Chapter 4. We show in Section 3.2.3 that (P₁) is solved by using the primal-dual splitting method [117, 118].

¹Although \mathbf{L} and \mathbf{M} are set to scaled versions of the identity matrix in all the examples, the present definition as a diagonal operator is advantageous when some prior knowledge is available about which columns are more likely to contain outliers.

3.2.2 Motivation of Reformulation

We discuss the possibility of applying the primal-dual splitting method or ADMM directly to Problem (P₁). We first mention that, in order to consider the first and third terms as a single function of $\mathbf{A}\mathbf{X}$, one needs the additional condition $\text{rank } \mathbf{A} = n$, which strictly limits the applicability. A possible approach would therefore be to consider the second and third terms as a single function of \mathbf{X} and then use the firm-shrinkage operators studied in [112]. Let us first consider the primal-dual splitting methods [114, 115, 116]. These methods require the convexity of each term of the cost function as well as that of the entire cost. Moreover, these primal-dual methods use the proximity operator of the conjugate function of the first term $\Phi_{2,1,\mathbf{L}}^{\text{MC}}(\mathbf{Y} - \cdot)$ of (P₁). Due to the fact that the Fenchel conjugate of a given function coincides with that of its lower-semicontinuous convex envelope [96, Proposition 13.16], one can verify that $\text{Prox}_{(\Phi_{2,1,\mathbf{L}}^{\text{MC}})^*} = 0$ since the lower-semicontinuous convex envelope of $\Phi_{2,1,\mathbf{L}}^{\text{MC}}$ is a constant function, and hence $\text{Prox}_{(\Phi_{2,1,\mathbf{L}}^{\text{MC}})^{**}} = \text{Id}$. As a result, $\text{Prox}_{[\Phi_{2,1,\mathbf{L}}^{\text{MC}} \circ (\mathbf{Y} - \cdot)]^*} = 0$ due to the basic properties of conjugate function and proximity operator [96]. This means that the first term gives no impact on the algorithm output, and thus there is no hope to obtain a solution of (P₁). In contrast, the proximity operator of the conjugate function of $\Phi_{2,1,\mathbf{L}}^{\text{MC}}$ does not appear explicitly in the ADMM iterate. However, the convergence analysis is nontrivial in this case. For instance, the widely known approach of Eckstein and Bertsekas [119] applies the Douglas-Rachford splitting method to the dual problem. The conjugate function appearing in the dual problem is replaced by the original function due essentially to the Moreau decomposition, which cannot be applied to the current nonconvex case. Another equivalent form of ADMM iterate (which will be referred to as ADMM') is given in [120], bypassing the dual formulation. To apply this algorithm formally to Problem (P₁), consider using ADMM' to solve the following problem. minimize $f(\mathbf{X}) + g(\mathbf{Z})$ subject to $\mathbf{A}\mathbf{X} + \mathbf{Z} = \mathbf{Y}$, where $f : \mathbb{R}^{n \times r} \rightarrow (-\infty, +\infty]$ is a ρ -weakly convex function for $\rho > 0$, and $g \in \Gamma_0(\mathbb{R}^{m \times r})$. To ensure the convergence, firm nonexpansivity of $\text{Prox}_{\gamma p_1}$ is used in [120], where $p_1(\mathbf{U}) := \min_{\mathbf{X} \in \mathbb{R}^{n \times r}} \{f(\mathbf{X}) \mid \mathbf{A}\mathbf{X} = \mathbf{U} + \mathbf{Y}\}$. However, $\text{Prox}_{\gamma p_1}$ can be shown to be α -cocoercive for $\alpha := 1 - \gamma\rho\|\mathbf{A}^\dagger\|_2^2$ for $\gamma > 0$ such that $\alpha \in (0, 1)$. Since $\alpha < 1$, $\text{Prox}_{\gamma p_1}$ is not even nonexpansive.

3.2.3 Reformulation of (P₁)

The following lemma is used for the reformulation of Problem (P₁).

Lemma 3.1. For any $\mathbf{X} \in \mathbb{R}^{n \times d}$,

$$\Phi_{2,1,\mathbf{M}}^{\text{MC}}(\mathbf{X}) = \|\mathbf{X}\|_{2,1} - \frac{1}{2}\|\mathbf{X}\|_{\mathbf{M}}^2 + {}^1(q_C \circ \mathbf{M}^{1/2})(\mathbf{M}^{1/2}\mathbf{X}), \quad (3.3)$$

and

$$\begin{aligned}\Phi_{2,1,L}^{\text{MC}}(\mathbf{Y} - \mathbf{A}\mathbf{X}) &= \|\mathbf{Y} - \mathbf{A}\mathbf{X}\|_{2,1} - \frac{1}{2}\|\mathbf{Y} - \mathbf{A}\mathbf{X}\|_L^2 \\ &\quad + {}^1(i_C \circ \mathbf{L}^{1/2})(\mathbf{L}^{1/2}(\mathbf{Y} - \mathbf{A}\mathbf{X})),\end{aligned}\quad (3.4)$$

where

$$C := \text{lev}_{\leq 1} \|\cdot\|_{2,\infty} := \{\mathbf{X} \in \mathbb{R}^{n \times r} \mid \|\mathbf{X}\|_{2,\infty} \leq 1\}.\quad (3.5)$$

Proof. It is known that the conjugate function of any norm is defined as the indicator function of the unit ball of its dual norm [98, Example 3.26]. Since the dual norm of the $\ell_{2,1}$ norm is the $\ell_{2,\infty}$ norm [121, Section 3.3], we obtain $(\|\cdot\|_{2,1})^* = i_C$. It therefore follows, for any $\mathbf{X} \in \mathbb{R}^{n \times r}$, that

$$\begin{aligned}\Phi_{2,1,M}^{\text{MC}}(\mathbf{X}) &= \|\mathbf{X}\|_{2,1} - {}^1(\|\cdot\|_{2,1} \circ \mathbf{M}^{-1/2})(\mathbf{M}^{1/2}\mathbf{X}) \\ &= \|\mathbf{X}\|_{2,1} - \left(\frac{1}{2}\|\cdot\|_{\text{F}}^2 - {}^1(i_C \circ (\mathbf{M}^{1/2})^\top)\right)(\mathbf{M}^{1/2}\mathbf{X}) \\ &= \|\mathbf{X}\|_{2,1} - \frac{1}{2}\|\mathbf{X}\|_M^2 + {}^1(i_C \circ \mathbf{M}^{1/2})(\mathbf{M}^{1/2}\mathbf{X}).\end{aligned}\quad (3.6)$$

Here, the first equality is due to [36, Lemma 1], and the second equality is due to Facts 2.1 and 2.4(c). The latter case is proven in the same way. \square

By Lemma 3.1, Problem (P₁) is reformulated as follows.

$$\begin{aligned}&\min_{\mathbf{X} \in \mathbb{R}^{n \times r}} \left(\Phi_{2,1,L}^{\text{MC}}(\mathbf{Y} - \mathbf{A}\mathbf{X}) + \mu_1 \Phi_{2,1,M}^{\text{MC}}(\mathbf{X}) + \frac{\mu_2}{2}\|\mathbf{X}\|_{\text{F}}^2 \right) \\ \Leftrightarrow &\min_{\mathbf{X} \in \mathbb{R}^{n \times r}} \left(\frac{\mu_2}{2}\|\mathbf{X}\|_{\text{F}}^2 - \frac{1}{2}\|\mathbf{A}\mathbf{X}\|_L^2 - \frac{\mu_1}{2}\|\mathbf{X}\|_M^2 + \langle \mathbf{A}\mathbf{X}, \mathbf{Y} \rangle_L \right. \\ &\quad \left. + {}^1(i_C \circ \mathbf{L}^{1/2})(\mathbf{L}^{1/2}(\mathbf{Y} - \mathbf{A}\mathbf{X})) + \mu_1 {}^1(i_C \circ \mathbf{M}^{1/2})(\mathbf{M}^{1/2}\mathbf{X}) \right. \\ &\quad \left. + \mu_1 \|\mathbf{X}\|_{2,1} + \|\mathbf{Y} - \mathbf{A}\mathbf{X}\|_{2,1} \right) \\ \Leftrightarrow &(\text{P}'_1) \min_{\mathbf{X} \in \mathbb{R}^{n \times r}} (F(\mathbf{X}) + G(\Theta\mathbf{X})),\end{aligned}\quad (3.7)$$

where

$$\Theta := \begin{bmatrix} \mathbf{A} \\ \mu_1 \mathbf{I}_n \end{bmatrix} \in \mathbb{R}^{(m+n) \times n}\quad (3.8)$$

Here, the functions F and G are defined as follows:

$$\begin{aligned}F : \mathbb{R}^{n \times r} \rightarrow \mathbb{R} : \mathbf{X} \mapsto &\frac{\mu_2}{2}\|\mathbf{X}\|_{\text{F}}^2 - \frac{1}{2}\|\mathbf{A}\mathbf{X}\|_L^2 - \frac{\mu_1}{2}\|\mathbf{X}\|_M^2 + \langle \mathbf{A}\mathbf{X}, \mathbf{Y} \rangle_L \\ &+ {}^1(i_C \circ \mathbf{L}^{1/2})(\mathbf{L}^{1/2}(\mathbf{Y} - \mathbf{A}\mathbf{X})) + \mu_1 {}^1(i_C \circ \mathbf{M}^{1/2})(\mathbf{M}^{1/2}\mathbf{X}),\end{aligned}\quad (3.9)$$

$$G : \mathbb{R}^{(n+m) \times r} \rightarrow \mathbb{R} : \mathbf{Z} \mapsto \left\| \mathbf{Z} - \begin{bmatrix} \mathbf{Y} \\ \mathbf{0}_{n \times r} \end{bmatrix} \right\|_{2,1}.\quad (3.10)$$

The function G is clearly convex and nonsmooth. On the other hand, the function F is smooth, *i.e.*, F is differentiable and its gradient ∇F is Lipschitz continuous with constant (see Appendix I.1 for its derivation)

$$\begin{aligned} \beta &:= \lambda_{\max}(\mu_2 \mathbf{I}_n - \mathbf{A}^\top \text{diag}(l_1, l_2, \dots, l_m) \mathbf{A} - \mu_1 \text{diag}(m_1, m_2, \dots, m_n)) \\ &\quad + \lambda_{\max}(\mathbf{A}^\top \text{diag}(l_1, l_2, \dots, l_m) \mathbf{A}) + \mu_1 \max\{m_1, m_2, \dots, m_n\}. \end{aligned} \quad (3.11)$$

The function F is also strictly convex under a certain condition as shown by Proposition 3.1 below.

Proposition 3.1. The function F is convex if

$$\mu_2 \geq \lambda_{\max}\{\mathbf{A}^\top \text{diag}(l_1, l_2, \dots, l_m) \mathbf{A} + \mu_1 \text{diag}(m_1, m_2, \dots, m_n)\}. \quad (3.12)$$

The condition (3.12) is also necessary when

$$\begin{aligned} K &:= \left\{ \mathbf{X} \in \mathbb{R}^{n \times r} \mid {}^1(\iota_C \circ \mathbf{L}^{1/2})(\mathbf{L}^{1/2}(\mathbf{Y} - \mathbf{A}\mathbf{X})) = 0 \right\} \\ &\quad \cap \left\{ \mathbf{X} \in \mathbb{R}^{n \times r} \mid {}^1(\iota_C \circ \mathbf{M}^{1/2})(\mathbf{M}^{1/2}\mathbf{X}) = 0 \right\} \end{aligned} \quad (3.13)$$

has a nonempty interior. In particular, F is strongly convex if (3.12) holds with strict inequality.

Proof. See Appendix I.2. \square

Strict convexity of the entire function in (P'_1) is verified by Proposition 3.1 with the convexity of G . The coercivity of the entire function in (P'_1) is verified with the following proposition.

Proposition 3.2. The cost function J in (3.2.1) is coercive.

Proof. It holds from (3.2) that, for any $\mathbf{B} \in \mathbb{R}^{m \times r}$,

$$\begin{aligned} \Phi_{2,1,\mathbf{L}}^{\text{MC}}(\mathbf{B}) &:= \|\mathbf{B}\|_{2,1} - \min_{\mathbf{Z} \in \mathbb{R}^{m \times r}} \left(\|\mathbf{Z}\|_{2,1} + \frac{1}{2} \|\mathbf{B} - \mathbf{Z}\|_{\mathbf{L}}^2 \right) \\ &\geq \|\mathbf{B}\|_{2,1} - \left[\|\mathbf{Z}\|_{2,1} + \frac{1}{2} \|\mathbf{B} - \mathbf{Z}\|_{\mathbf{L}}^2 \right]_{\mathbf{Z}=\mathbf{B}} \\ &= 0. \end{aligned} \quad (3.14)$$

In the same way, we obtain $\Phi_{2,1,\mathbf{M}}^{\text{MC}}(\mathbf{X}) \geq 0$ for any $\mathbf{X} \in \mathbb{R}^{n \times r}$. Therefore, J is coercive since the sum of a coercive function and a function bounded below is coercive [96, Corollary 11.16]. \square

By Propositions 3.1 and 3.2, the cost function of (P'_1) is strictly convex and coercive if $\mu_2 > \lambda_{\max}\{\mathbf{A}^\top \text{diag}(l_1, l_2, \dots, l_m) \mathbf{A} + \mu_1 \text{diag}(m_1, m_2, \dots, m_n)\}$, and hence the uniqueness and existence of the solution are guaranteed in this case [96, Corollary 11.16].

Applying the forward-backward-based primal-dual algorithm [117, 118] (see Appendix C) to solve (P'_1) produces Algorithm 3.1.

Algorithm 3.1 : Forward-backward primal-dual splitting method for solving (P'_1)

Set $\mathbf{X}_0 \in \mathbb{R}^{n \times r}$, $\mathbf{V}_0 \in \mathbb{R}^{(n+m) \times r}$, and $\tau, \varsigma > 0$.

For $k = 0, 1, 2, \dots$

$$\mathbf{X}_{k+\frac{1}{2}} := \mathbf{X}_k - \tau \nabla F(\mathbf{X}_k),$$

$$\mathbf{V}_{k+1} := (\text{Id} - \text{Prox}_{(\tau/\varsigma)G}) \left\{ \begin{bmatrix} \mathbf{A}\mathbf{X}_{k+\frac{1}{2}} \\ \mu_1 \mathbf{X}_{k+\frac{1}{2}} \end{bmatrix} + \left(\mathbf{I}_{n+m} - \varsigma \begin{bmatrix} \mathbf{A}\mathbf{A}^\top & \mu_1 \mathbf{A} \\ \mu_1 \mathbf{A}^\top & \mu_1^2 \mathbf{I}_n \end{bmatrix} \right) \mathbf{V}_k \right\}$$

$$\mathbf{X}_{k+1} := \mathbf{X}_{k+\frac{1}{2}} - \varsigma [\mathbf{A}^\top \ \mu_1 \mathbf{I}_n] \mathbf{V}_{k+1}$$

3.2.4 Closed-Form Expressions for the Operators

We present below closed-form expressions of the operators used in Algorithm 3.1 below.

Proposition 3.3. Closed-form expressions for ∇F and $\text{Prox}_{(\tau/\varsigma)G}$ are given as follows:

1. $\nabla F : \mathbf{X} \mapsto \mu_2 \mathbf{X} - \mu_1 \sum_{i=1}^n \min \left\{ \frac{1}{\|\mathbf{X}_{(i,:)}\|_2}, m_i \right\} \mathbf{e}_{n,i} \mathbf{X}_{(i,:)}^\top$
 $+ \mathbf{A}^\top \sum_{i=1}^m \min \left\{ \frac{1}{\|[\mathbf{Y} - \mathbf{A}\mathbf{X}]_{(i,:)}\|_2}, l_i \right\} \mathbf{e}_{m,i} [\mathbf{Y} - \mathbf{A}\mathbf{X}]_{(i,:)}^\top,$
2. $\text{Prox}_{(\tau/\varsigma)G} : \mathbf{Z} \mapsto \begin{bmatrix} \mathbf{Y} \\ \mathbf{0}_{n \times r} \end{bmatrix}$
 $+ \sum_{i=1}^m \max \left\{ 1 - \frac{\tau/\varsigma}{\|[\mathbf{Z} - \mathbf{Y}]_{(i,:)}\|_2}, 0 \right\} \mathbf{e}_{m+n,i} [\mathbf{Z} - \mathbf{Y}]_{(i,:)}^\top$
 $+ \sum_{i=1}^n \max \left\{ 1 - \frac{\tau/\varsigma}{\|\mathbf{Z}_{(m+i,:)}\|_2}, 0 \right\} \mathbf{e}_{m+n,m+i} \mathbf{Z}_{(m+i,:)}^\top,$

Proof. See Appendix I.3. □

3.2.5 Convergence Conditions and Complexity

The following proposition gives a guarantee of convergence of Algorithm 3.1 to solve (P'_1) .

Proposition 3.4. Assume that convexity condition (3.12) is satisfied. Let $(\mathbf{X}_k)_{k \in \mathbb{N}}$ and $(\mathbf{V}_k)_{k \in \mathbb{N}}$ be the sequences generated by Algorithm 3.1, respectively. Suppose that

$$0 < \tau < \frac{2}{\beta}, \quad 0 < \varsigma \leq \frac{1}{\lambda_{\max}(\mathbf{A}^\top \mathbf{A}) + \mu_1^2}. \quad (3.15)$$

Then, the sequence $(\mathbf{X}_k)_{k \in \mathbb{N}}$ converges to a solution of (P'_1) , and $((\tau/\varsigma)\mathbf{V}_k)_{k \in \mathbb{N}}$ converges to a solution of the associated dual problem

$$\min_{\zeta \in \mathbb{R}^{m+n}} F^*(-\Theta^\top \zeta) + G^*(\zeta). \quad (3.16)$$

Proof. The proof is due to Fact C.1 in light of Proposition 4.1. \square

We now analyze the computational complexities of our approach and the RFS. The complexity of our approach can be divided into two parts: the number of multiplications to compute τ and ς (step 1) and the one for the iterations (step 2).

In step 1, we consider the number of multiplications to attain β and $\lambda_{\max}(\Theta^\top \Theta)$, from which the proximal parameters τ and ς can be obtained due to (3.15). Using the power method, the computational complexity for β by (3.11) is given by $\mathcal{O}(\min\{n^2, m^2\} \max\{n, m, p_\beta\})$ where p_β is the number of power iterations. See appendix D for the definition of the big Oh notation \mathcal{O} . The above arguments are based on the following equations:

$$\begin{aligned} & \lambda_{\max}(\mu_2 \mathbf{I}_n - \mathbf{A}^\top \text{diag}(l_1, l_2, \dots, l_m) \mathbf{A} - \mu_1 \text{diag}(m_1, m_2, \dots, m_n)) \\ &= \lambda_{\max}(\mu_2 \mathbf{I}_m - \text{diag}(l_1^{1/2}, l_2^{1/2}, \dots, l_m^{1/2}) \mathbf{A} \mathbf{A}^\top \text{diag}(l_1^{1/2}, l_2^{1/2}, \dots, l_m^{1/2}) \\ & \quad - \mu_1 \text{diag}(m_1, m_2, \dots, m_n)), \end{aligned} \quad (3.17)$$

$$\begin{aligned} & \lambda_{\max}(\mathbf{A}^\top \text{diag}(l_1, l_2, \dots, l_m) \mathbf{A}) \\ &= \lambda_{\max}(\text{diag}(l_1^{1/2}, l_2^{1/2}, \dots, l_m^{1/2}) \mathbf{A} \mathbf{A}^\top \text{diag}(l_1^{1/2}, l_2^{1/2}, \dots, l_m^{1/2})). \end{aligned} \quad (3.18)$$

The complexity of $\lambda_{\max}(\Theta^\top \Theta)$ scales in $\mathcal{O}(p_\varsigma \min\{mnr, n^2(m+r)\})$, where p_ς is the number of power iterations to obtain ς . This is verified as follows: it holds for any $\mathbf{X} \in \mathbb{R}^{n \times r}$ that

$$\Theta^\top \Theta \mathbf{X} = \mathbf{A}^\top \mathbf{A} \mathbf{X} + \mu_1^2 \mathbf{X} \quad (3.19)$$

and the complexity of $\mathbf{A}^\top \mathbf{A} \mathbf{X}$ is $\mathcal{O}(\min\{mnr, n^2(m+r)\})$. The overall complexity of τ and ς amounts to $\mathcal{O}(\max\{\min\{n^2, m^2\} \max\{n, m, p_\beta\}, p_\varsigma \min\{mnr, n^2(m+r)\}\})$.

In step 2, the complexity for updating $\mathbf{X}_{k+\frac{1}{2}}$ and \mathbf{V}_{k+1} of Algorithm 3.1 are $\mathcal{O}(nmr)$ and $\mathcal{O}(nmr)$, respectively. Hence, the overall complexity per iteration scales in $\mathcal{O}(nmr)$. The complexity of the RFS is $\mathcal{O}(m(m+n) \max(m+n, r))$, which is larger than that of the proposed approach for a sufficiently large m (see Table 3.1) Figure 3.1 compares the CPU time for the convergence of the proposed method and the RFS. This experiment is performed in Python 3.8.13 on a 64-bit PC with 12th gen Intel(R) Core(TM) i9-12900 CPU (5.1 GHz). The experimental setting is the same as Experiment 3-A in Section 3.3 below for SNR 30 dB except that $n := 128$.

We finally mention that the RFS problem (P_0) can be solved by the well-established convex optimization techniques such as the primal-dual splitting

Table 3.1: Computational complexity (q : the number of iterations).

Algorithm	Time complexity
Proposed	$\mathcal{O}(\max\{qnmr, \max\{\min\{n^2, m^2\} \max\{n, m, p_\beta\}, p_\zeta \min\{mnr, n^2(m+r)\}\}\})$
RFS	$\mathcal{O}(qm(m+n) \max\{m+n, r\})$

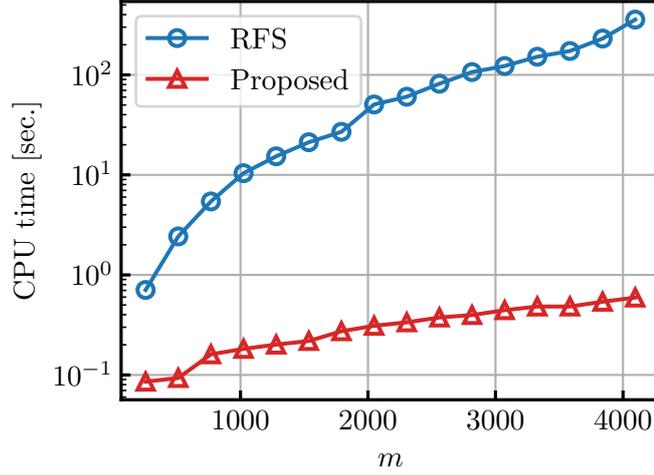


Figure 3.1: Computational complexities of the proposed approach and the RFS.

methods [114, 116, 115] and ADMM [113], all of which are more efficient than the method in [9].

3.2.6 Remarks

Remark 3.1. One may consider solving Problem (P'_1) by other algorithms such as the primal-dual splitting method of Chambolle and Pock and ADMM. It is difficult to compute the proximity operator of F in this case; see [122, 59] for some cases in which the sum of functions is jointly proximable. A possible way to circumvent this difficulty is to apply Chambolle and Pock's algorithm through the Pierra-type product-space reformulation [123, 124] as follows. We first decompose F into a sum of multiple proximable functions $\sum_{i=1}^p f_i$ for $p > 2$ and $f_i \in \Gamma_0(\mathbb{R}^{n \times r})$ ($i = 1, 2, \dots, p$), and then let the arguments of the functions be $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_p$, respectively. We finally add the indicator function to enforce the linear constraint $\mathbf{X}_1 = \dots = \mathbf{X}_p$. This strategy results in an undesirable increase in the number of variables, causing significant increases in the computational costs and memory requirements. An application of ADMM to (P'_1) , on the other hand, essentially requires the proximity operator of F . This implies the necessity of the inner loop to

find the proximity operator numerically by some iterative methods. To avoid such increases in complexity and memory requirements and the inner loop, we leverage the primal-dual splitting method of Condat [116] to solve (P'_1) in the present study.

Remark 3.2. Problem (P'_1) is equivalent to the following problem:

$$\min_{\mathbf{X} \in \mathbb{R}^{n \times r}} (F(\mathbf{X}) + \tilde{G}(\mathbf{X}) + \tilde{H}(\Theta \mathbf{X})), \quad (3.20)$$

where

$$\Gamma_0(\mathbb{R}^{n \times r}) \ni \tilde{G} : \mathbf{X} \mapsto \mu_1 \|\mathbf{X}\|_{2,1}, \quad (3.21)$$

$$\Gamma_0(\mathbb{R}^{m \times r}) \ni \tilde{H} : \mathbf{Z} \mapsto \|\mathbf{Y} - \mathbf{Z}\|_{2,1}. \quad (3.22)$$

The problem in (3.20) can be solved by the primal-dual splitting method of Condat [116]. However, the conditions on the step sizes for this algorithm are more restrictive than Algorithm 3.1. We note that there may be better optimization algorithms to solve (P'_1) , and the best selection of the algorithm is an open issue.

Remark 3.3. Some readers may consider that the use of the GMC functions for the loss and penalty terms in (P_1) instead of the MC functions would be more advantageous. Although the GMC penalty with appropriate parameters allows the global optimality for the problem involving the quadratic loss (see (1.1)) even in the underdetermined case, it is not necessary in the present study. The quadratic loss cannot be strongly convex over the whole space in the underdetermined case, and hence it does not annihilate the weak convexity of the MC penalty. However, in the present study, the quadratic term in (P_1) does not involve any linear operator, and hence it is strongly convex even in the underdetermined case. Hence, the convexity of (P_1) can be guaranteed without employing the GMC functions.

3.3 Numerical Examples

We first show the robustness of the proposed approach and its performance in the support recovery task. We then evaluate the efficacy of the proposed approach in an application to the recovery of MEG signals.

3.3.1 Experiment 3-A: Toy Data

3.3.1.1 Robustness

Matrices $\mathbf{X}_\diamond \in \mathbb{R}^{n \times r}$ and $\mathbf{A} \in \mathbb{R}^{m \times n}$ are generated from the i.i.d. normal distribution $\mathcal{N}(0, 1)$. Here, we consider dense \mathbf{X}_\diamond to show the pure effects

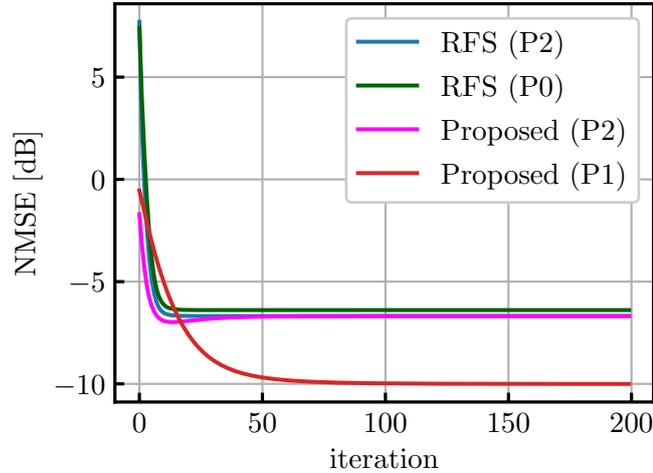


Figure 3.2: Learning curves under $n = 512$, $m = 1024$, $r = 32$, outlier rate 30% and signal-to-noise ratio (SNR) 30 dB.

of the robustification (excluding the sparsification effect). The matrix \mathbf{E}_\star is generated from the i.i.d. normal distribution with SNR, where

$$\text{SNR} := \frac{\mathbb{E}[\|\mathbf{A}\mathbf{X}_\diamond\|_{\text{F}}^2]}{\mathbb{E}[\|\mathbf{E}_\star\|_{\text{F}}^2]}. \quad (3.23)$$

The outlier matrix \mathbf{O}_\diamond is column sparse, and its non-zero elements are drawn from $\mathcal{N}(0, 1)$ and are then multiplied by the factor 100. To measure the accuracy of the recovered signals, the normalized mean squared errors (NMSE) given by

$$\text{NMSE} := \frac{\|\mathbf{X}_\diamond - \hat{\mathbf{X}}\|_{\text{F}}^2}{\|\mathbf{X}_\diamond\|_{\text{F}}^2} \quad (3.24)$$

are used. The RFS [9] is considered for comparison. In [9], it is mentioned that one can easily extend the RFS to those problems with different penalties, such as

$$(\text{P}_2) \min_{\mathbf{X} \in \mathbb{R}^{n \times r}} \|\mathbf{Y} - \mathbf{A}\mathbf{X}\|_{2,1} + \frac{\mu_2}{2} \|\mathbf{X}\|_{\text{F}}^2, \quad (3.25)$$

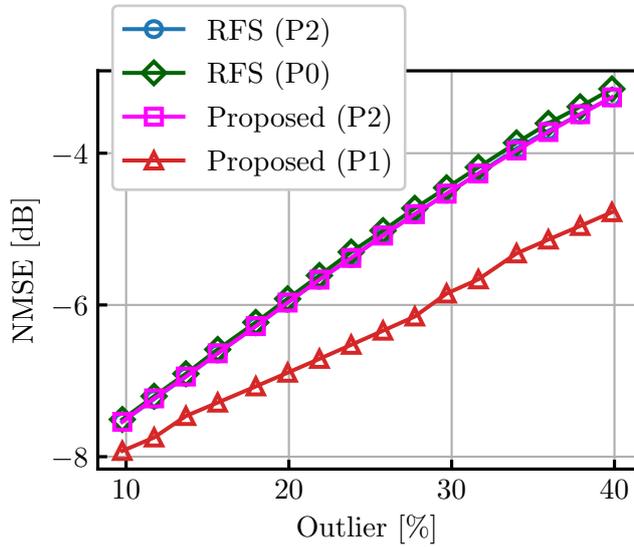
which is a special case of (P₁) for $\mathbf{L} = \mathbf{0}_{m \times m}$ and $\mu_1 = 0$.

Since \mathbf{X}_\diamond is dense, we let $\mu_1 = 0$ and $\mathbf{M} = \mathbf{0}_{n \times n}$. On the basis of Proposition 3.1, μ_2 is set to $\mu_2 := \lambda_{\max}\{\mathbf{A}^\top \text{diag}(l_1, l_2, \dots, l_m) \mathbf{A}\}$ which gave the best performance. We let $l_1 = \dots = l_m = \alpha_l > 0$, and α_l and λ in (P₀) are optimized via grid search using the clean data which are free from noise and outliers.

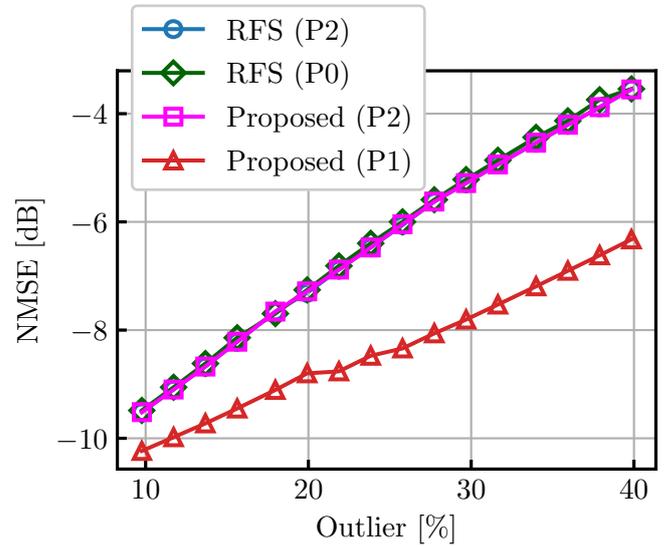
Figure 3.2 depicts the learning curves for $n = 512$, $m = 1024$, $r = 32$, and the SNR is set to 30 dB. Although the proposed approach for Problem (P₁) requires a larger number of iterations than the

RFS for Problem (P₀), the proposed approach is indeed more beneficial than the RFS in terms of the computational time. To be specific, $\mathcal{O}(\max\{qnmr, \max\{\min\{n^2, m^2\} \max\{n, m, p_\beta\}, p_\zeta \min\{mnr, n^2(m+r)\}\}\}) \approx 2.8 \times 10^8$ for the proposed approach, and $qm(m+n) \max(m+n, r) \approx 1.1 \times 10^9$ for the RFS, where $q = 131$, $p_\beta = 1167$, and $p_\zeta = 219$ for the proposed approach, and $q = 27$ for the RFS.

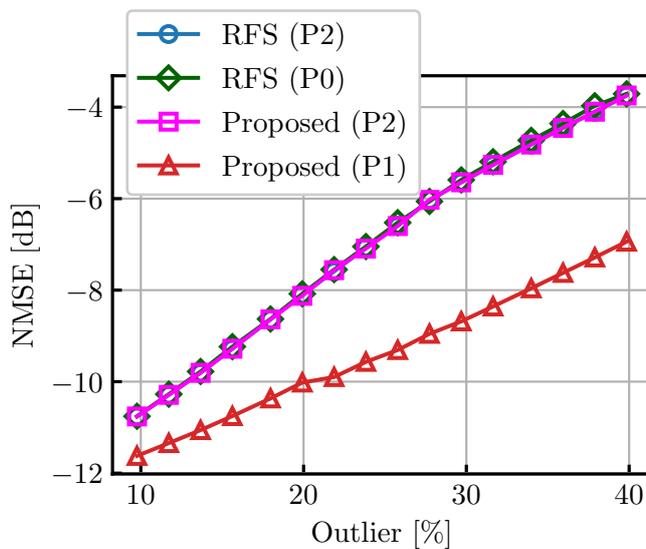
Figure 3.3 depicts the performance for $n = 160$, $m = 256$, and $r = 32$ under 300 trials when the outlier rate (the ratio between m and the number of nonzero columns of the outlier matrix \mathbf{O}_\diamond) changes between 10% and 40%. The SNR is set to 10, 20, 30, and $+\infty$ dB (*i.e.*, the noiseless case). It is seen that the proposed approach is more effective for a denser outlier matrix. We mention that, once recovering the support, one can remove those data corresponding to the off-support components and solve another (smaller size) regression problem that involves a smaller number of variables than that of the original problem. In that respect, the present setting of $n < m$ is a reasonable choice. Note that the performance of the RFS (P₂) and Proposed (P₂) are particularly different in Fig. 3.3(d) possibly because the matrix inversion produces numerical errors for the RFS (P₂).



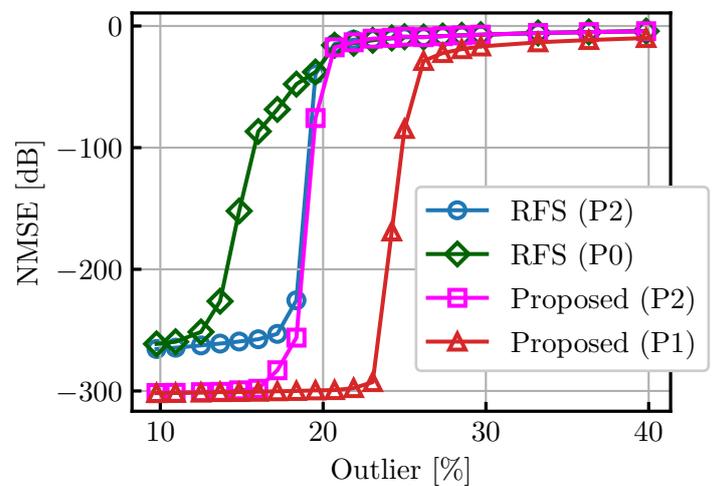
(a) SNR 10 dB



(b) SNR 20 dB



(c) SNR 30 dB



(d) noiseless

Figure 3.3: NMSE across outlier rate under $n = 160$, $m = 256$, and $r = 32$.

3.3.1.2 Support Recovery

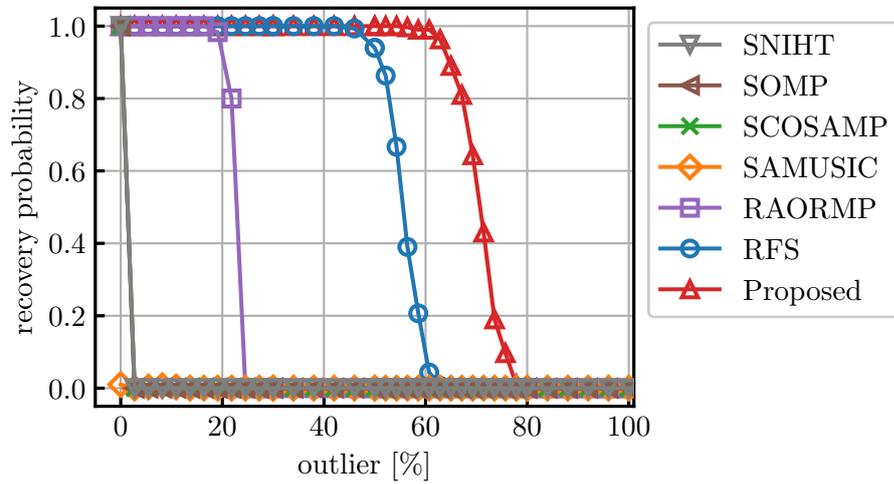
We demonstrate the remarkable robustness of the proposed approach for support recovery. We compare the proposed approach to the RFS and the state-of-the-art MMV algorithms: SNIHT [51], SOMP [50], SCoSaMP [51], SA-MUSIC [53], and RA-ORMP [48].

First, we investigate the robustness of the proposed approach for sparse signals under SNR 10 dB for $n = 256$, $m = 128$, and $r = 32$. We generate matrices $\mathbf{X}_\diamond \in \mathbb{R}^{n \times r}$ and $\mathbf{A} \in \mathbb{R}^{m \times n}$, both of which obey the i.i.d. normal distribution $\mathcal{N}(0, 1)$, and set $n - k$ column vectors of \mathbf{X}_\diamond to the zero vectors, where k is called block sparsity. We define the signal-to-outlier ratio (SOR) as

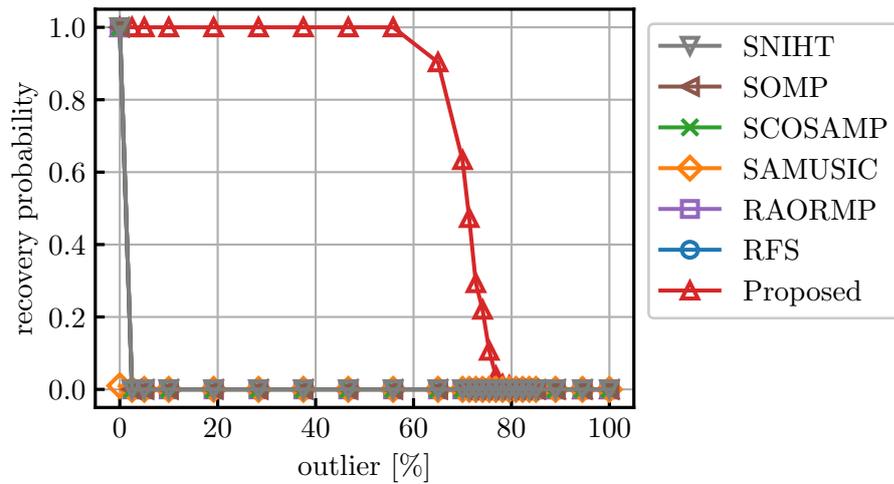
$$\text{SOR} := \frac{\mathbb{E}[\|\mathbf{A}\mathbf{X}_\diamond\|_{\text{F}}^2]/m}{\mathbb{E}[\|\mathbf{O}_\diamond\|_{\text{F}}^2]/\kappa_o}, \quad (3.26)$$

where κ_o denotes the number of non-zero column vectors in \mathbf{O}_\diamond . We let SOR -30 and -3000 dB. We let $\mu_2 := \lambda_{\max}\{\mathbf{A}^\top \text{diag}(l_1, l_2, \dots, l_m)\mathbf{A} + \mu_1 \text{diag}(m_1, m_2, \dots, m_n)\}$, $l_1 = \dots = l_m = \alpha_l > 0$, $m_1 = \dots = m_n = \alpha_m > 0$, and α_l , α_m , μ_1 , and λ in (P₀) are optimized via grid search.

Figure 3.4 shows the support recovery probability under 300 trials for different rates of outliers. It is seen that the proposed approach recovers the support successfully under higher outlier-rate situations than the RFS. We stress here that the proposed approach succeeds even under SOR -3000 dB up to the outlier rate 55%, while the RFS fails for the outlier rate less than 1%. Figure 3.5 plots the recovery probability as a function of k/n and r under SNR 10 dB, outlier 30%, SOR -30 dB, and 300 trials. The proposed approach achieves significantly higher recovery probabilities than the RFS due to its remarkable robustness coming from the use of the MC loss function.

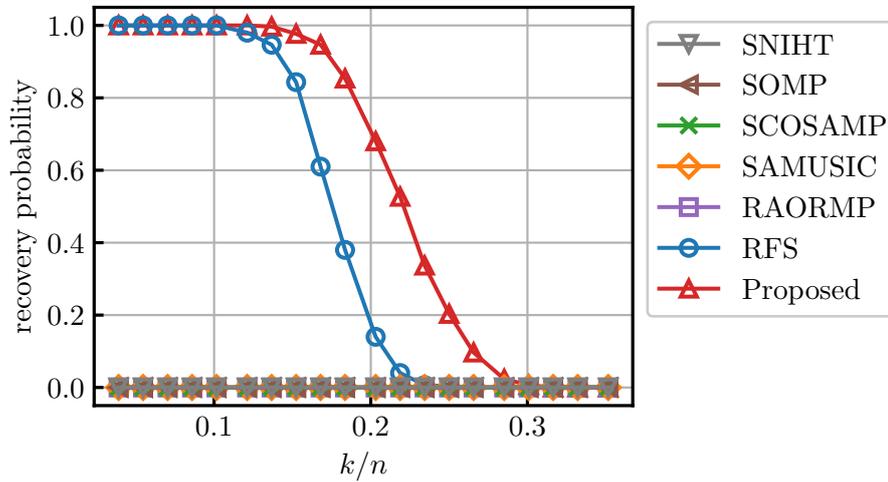
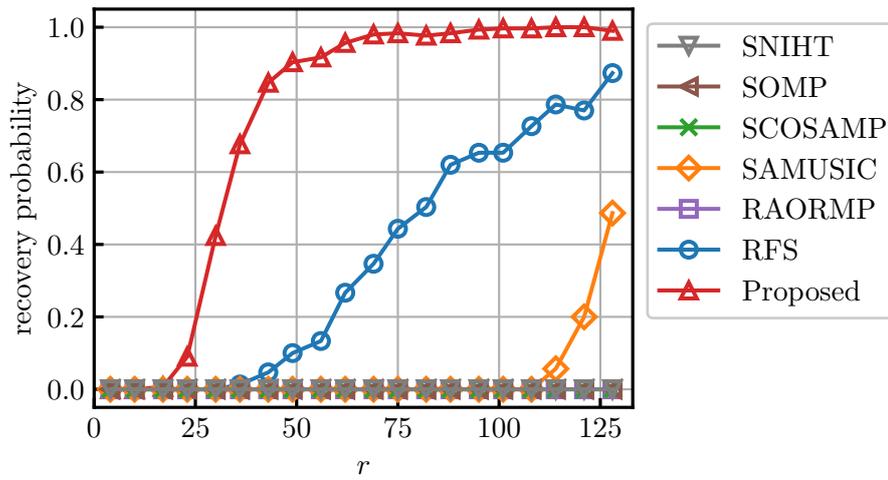


(a) SOR -30 dB



(b) SOR -3000 dB

Figure 3.4: Recovery probability across outlier rate for $n = 256$, $r = 32$, $m = 128$, $k = 16$, and SNR 10 dB.

(a) $m = 128, n/m = 2, n = 32$.(b) $m = 128, n/m = 1.5, k/n = 0.35$.Figure 3.5: Recovery probability as a function of (a) k/n and (b) r under SNR 10 dB, outlier rate 30%, and SOR -30 dB.

3.3.2 Experiment 3-B: Application to MEGC Signal Recovery Problem

We consider the MEGC signal recovery problem using the publicly available database PTB [125] [126]. The MMV approach has been considered for MEGC signals because the signals of each channel tend to have common sparse patterns in a wavelet domain [127] as shown in Figure 3.6. While the dataset contains 12 channels, we process $r := 4$ independent channels (Lead 1, Lead 2, V1, V2). From each channel, we extract signals of length $m := 512$ and construct the data matrix $\mathbf{Z}^{\text{MEGC}} \in \mathbb{R}^{m \times r}$. The data matrix is sparsified by Daubechies-4 (db4) wavelet basis² as

$$\mathbf{Z}^{\text{MEGC}} := \mathbf{W}\mathbf{X}_\diamond, \quad (3.27)$$

where $\mathbf{W} \in \mathbb{R}^{m \times m}$ is an orthonormal wavelet basis, and $\mathbf{X}_\diamond \in \mathbb{R}^{m \times r}$ is the matrix composed of the wavelet coefficient vectors. The decomposition level of the wavelets is set to 7 based on the relation between the sampling frequency and the decomposition level [129]. The measurement matrix $\mathbf{Y} \in \mathbb{R}^{m \times r}$ is corrupted by noise and outliers as follows:

$$\mathbf{Y} = \mathbf{W}\mathbf{X}_\diamond + \mathbf{E}_\star + \mathbf{O}_\diamond. \quad (3.28)$$

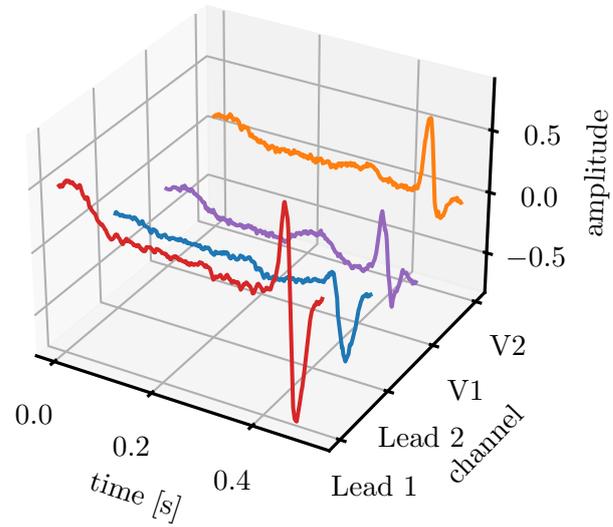
Each component of the noise matrix \mathbf{E}_\star follows i.i.d. $\mathcal{N}(0, \sigma_{\varepsilon_\star}^2)$. The SNR is set to 15 dB. The outlier matrix \mathbf{O}_\diamond is jointly sparse, and the number of non-zero column vectors is κ_o . The nonzero components of \mathbf{O}_\diamond follow i.i.d. $\mathcal{N}(0, \sigma_{o_\diamond}^2)$ with a given SOR.

We let $\mu_2 := \lambda_{\max}\{\mathbf{W}^\top \text{diag}(l_1, l_2, \dots, l_m)\mathbf{W} + \mu_1 \text{diag}(m_1, m_2, \dots, m_n)\}$, $l_1 = l_2 = \dots = l_m = \alpha_l > 0$, $m_1 = m_2 = \dots = m_n = \alpha_m > 0$, and α_l , α_m , and μ_1 in (P_0) are optimized via grid search. The evaluation metric is NMSE defined as

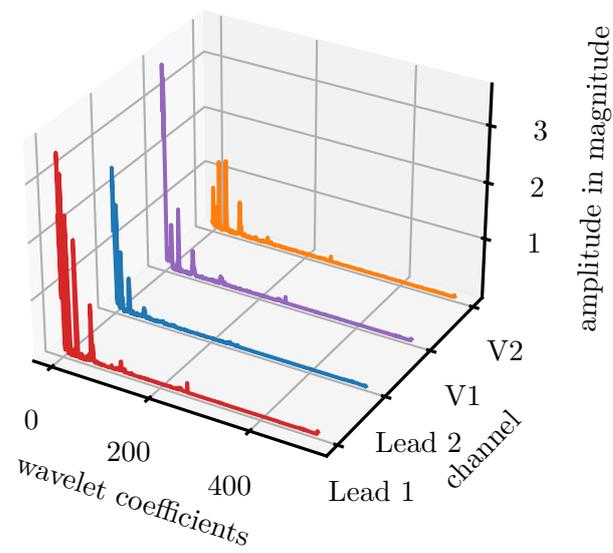
$$\text{NMSE} := \frac{\|\mathbf{Z}^{\text{MEGC}} - \mathbf{W}\hat{\mathbf{X}}\|_{\text{F}}^2}{\|\mathbf{Z}^{\text{MEGC}}\|_{\text{F}}^2} \quad (3.29)$$

Figures 3.7(a) and 3.7(b) depict NMSE averaged over 300 trials across different outlier rates and SOR, respectively. The proposed approach outperforms the RFS in the presence of noise and outliers over a wide range of outlier rates and SOR. Figures 3.8-3.11 show the original MEGC signals and recovered signals by the proposed approach and the RFS in the presence of severe outliers. Specifically, Figure 3.8 shows that some part of the signal is underestimated by the RFS since its robustness is limited. In contrast, the proposed method recovers the MEGC signal with high precision.

²It is known that Daubechies and Symlet families yield more sparse representation for ECG data than other representative types of wavelet such as Meyer, biorthogonal, and reverse biorthogonal wavelets [128].

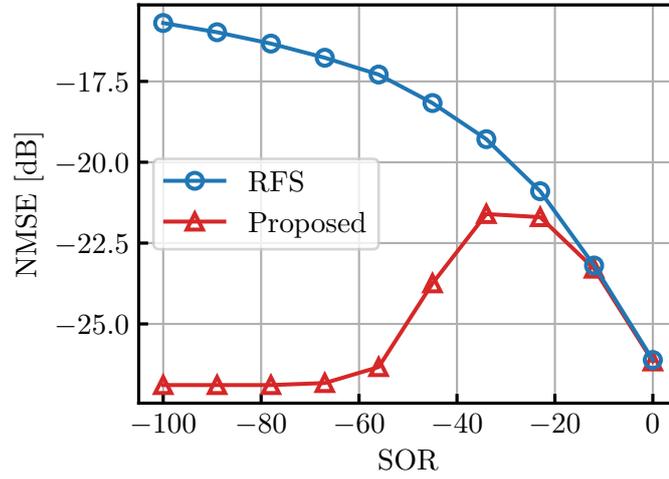


(a)

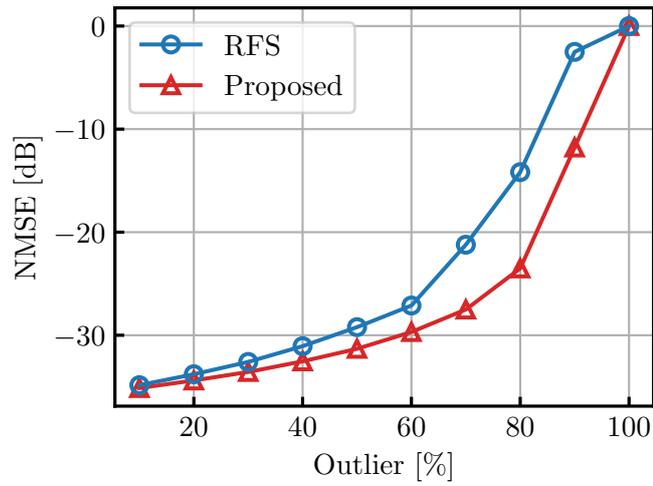


(b)

Figure 3.6: a) Clean MEGC signals. (b) Amplitudes of wavelet coefficients of the MEGC signals in each of the four channels.

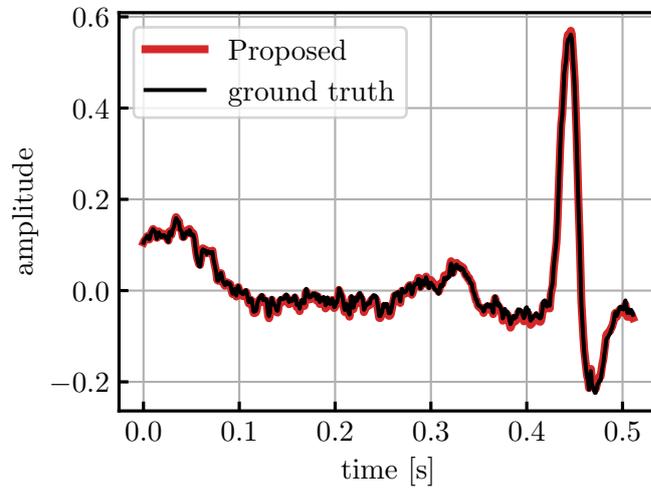


(a)

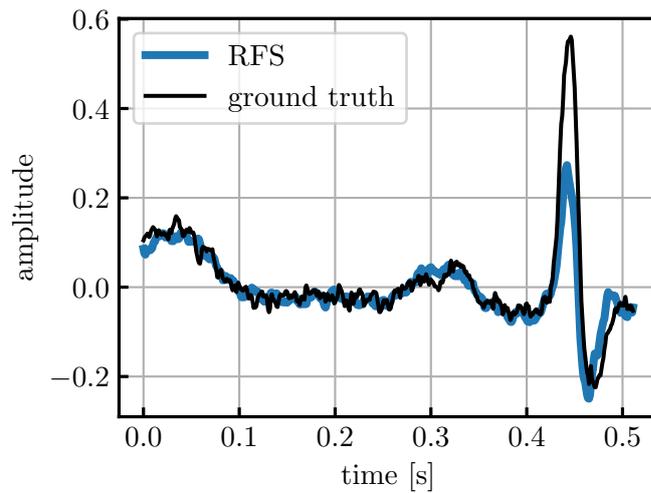


(b)

Figure 3.7: (a) NMSE as a function of SOR under outlier rate 80%, (b) NMSE as a function of outlier rate under SOR -75 dB.

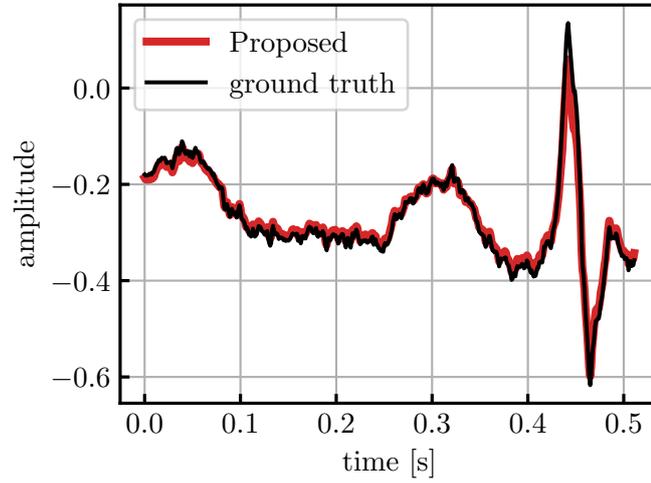


(a)

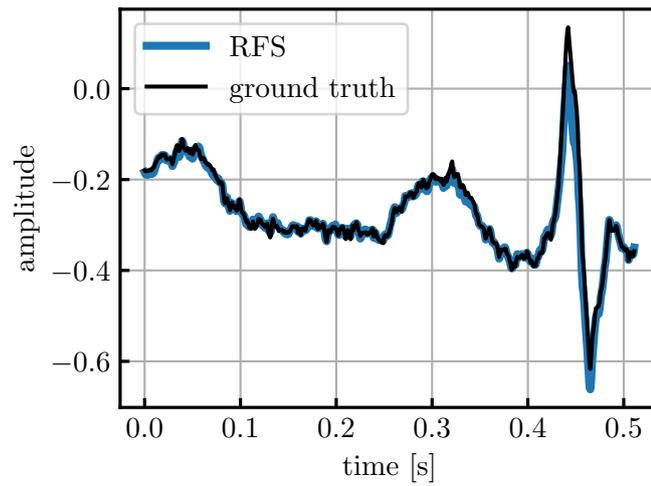


(b)

Figure 3.8: Measurements, original ECG signal, and recovered ECG signals from (a) the proposed method and (b) the RFS under outlier rate 80% and SOR -40 dB for Lead 1.

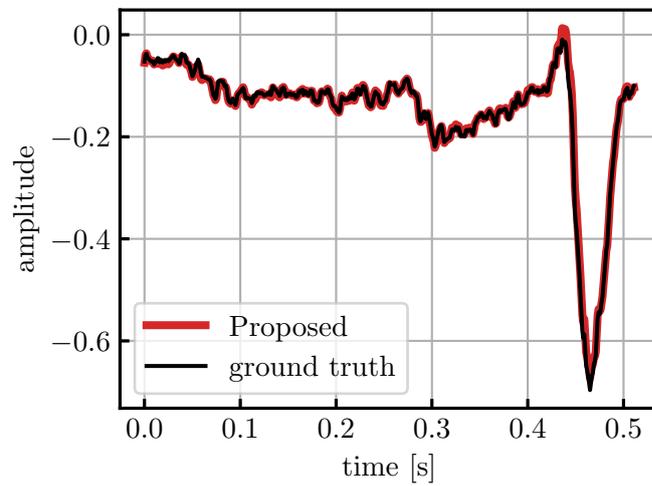


(a)

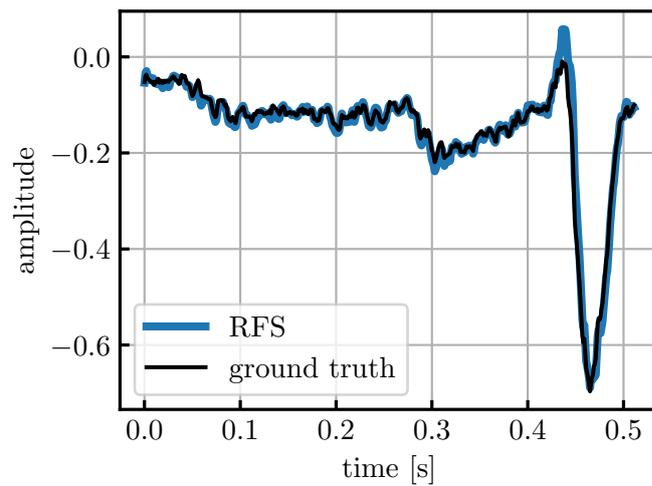


(b)

Figure 3.9: Measurements, original ECG signal, and recovered ECG signals from (a) the proposed method and (b) the RFS under outlier rate 80% and SOR -40 dB for Lead 2.

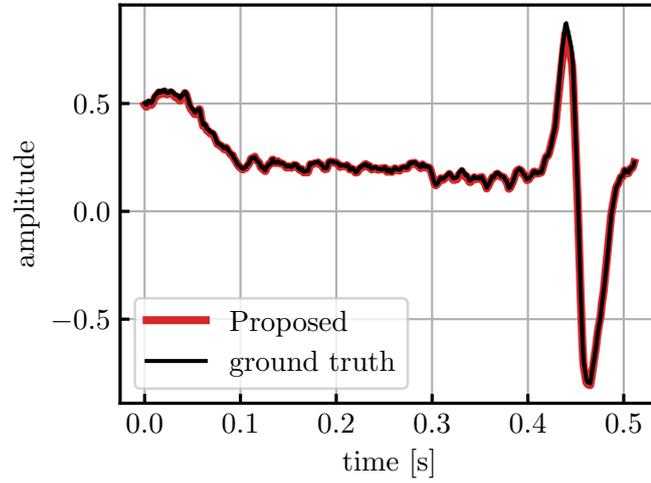


(a)

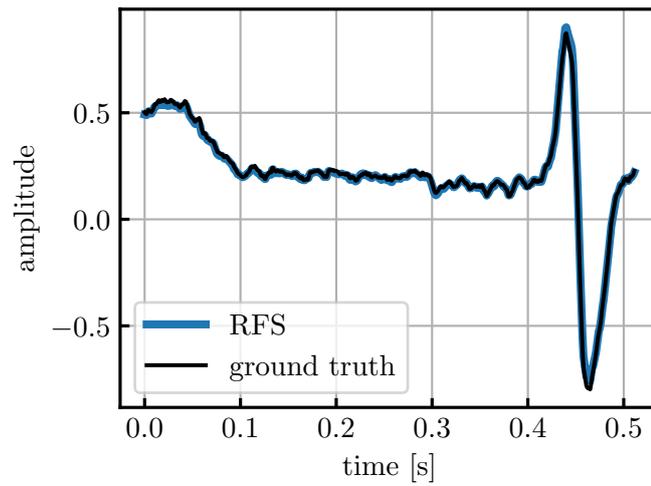


(b)

Figure 3.10: Measurements, original ECG signal, and recovered ECG signals from (a) the proposed method and (b) the RFS under outlier rate 80% and SOR -40 dB for V1.



(a)



(b)

Figure 3.11: Measurements, original ECG signal, and recovered ECG signals from (a) the proposed method and (b) the RFS under outlier rate 80% and SOR -40 dB for V2.

3.4 Conclusion

We proposed a robust approach to recovering jointly-sparse signals in the presence of outliers to fundamentally solve the tradeoff between robustness and global optimality. This addresses (Q1), which was raised in Chapter 1.2. The main result was that the MC loss function led to the remarkable robustness to outliers. The overall convexity of the cost function was guaranteed by exploiting the weak convexity of the MC functions. The problem was solved, via reformulation based on the Moreau decomposition, by the primal-dual splitting method, for which the convergence condition for the current specific case was derived with a Lipschitz constant of the gradient used in the method. The numerical results showed that the proposed approach outperformed the RFS in terms of robustness against outliers. We finally remark that, although this chapter concerns the jointly-sparse signals for generality, the applicability of the proposed approach is never limited to such signals and it will be useful in a wide range of applications including robust regression (such as the single measurement vector case in which the system may or may not be sparse and in which the outliers need to be managed).

Chapter 4

Sparse Stable Outlier-Robust Regression

4.1 Introduction

This chapter presents a robust method to estimate sparse signals which appropriately distinguishes the statistical differences between Gaussian noise and outliers to answer (Q2) raised in Chapter 1.2. The major contributions of this chapter are summarized below.

1. We propose a sparse outlier-robust signal regression method which integrates the formulations for nearly unbiased sparse estimation based on the MC penalty in (1.3) and SORR which will be presented in (F.1.3) (see Appendix F.1) in the framework of convex optimization. The proposed method enjoys remarkable advantages including (i) notable outlier robustness, (ii) stability under heavy Gaussian noise, (iii) high accuracy for both overdetermined and underdetermined cases, and (iv) theoretical convergence guarantees to a global minimizer by employing the operator splitting method. To the best of our knowledge, none of the existing methods achieves those desirable properties simultaneously.
2. We formulate the linearly-involved Moreau-enhanced-over-subspace (LiMES) model [35] in a certain product space to allow the use of multiple (generalized) MC functions (see Section 4.4). This allows us to derive a sufficient condition for convexity of the cost function involving two MC functions and to identify when it is also a necessary condition (Proposition 4.1).
3. We use the MC and the squared ℓ_2 penalties simultaneously, inspired by the elastic net penalty [83]. This formulation leads to a certain remarkable grouping effect under an appropriate parameter choice. Specifi-

cally, we show that, in contrast to the elastic net, when a pair of features are highly correlated, the upper bound of the discrepancy between the corresponding coefficients is independent of the norm of the observation vector which could be large owing to huge outliers.

4. Numerical examples show that the proposed approach is more robust against strong outliers than the existing sparse outlier-robust signal regression methods. The efficacy of the proposed method is also shown in application to speech denoising even when the speech signal is corrupted by huge outliers.

The remainder of the chapter is organized as follows. Section 4.2 states the problem addressed, and introduces selected elements of convex analysis. Section 4.3 presents the proposed approach with convergence analysis. Section 4.4 presents the convexity condition by introducing the LIMES-based general model mentioned above. Section 4.5 demonstrates the simulation results, followed by the conclusion in Section 4.6.

4.2 Problem Statement

Let $\mathbf{y} \in \mathbb{R}^m$ be the measurement vector contaminated by noise and outliers:

$$\mathbf{y} = \mathbf{A}\mathbf{x}_\diamond + \boldsymbol{\varepsilon}_\star + \mathbf{o}_\diamond. \quad (4.1)$$

Here, $\mathbf{o}_\diamond \in \mathbb{R}^m$ is the sparse outlier vector. We assume that $\boldsymbol{\varepsilon}_\star$ follows a multivariate Gaussian distribution $\mathcal{N}(\mathbf{0}, \sigma_{\boldsymbol{\varepsilon}_\star}^2 \mathbf{I}_m)$. The present study concerns the task of recovering \mathbf{x}_\diamond from given \mathbf{y} and \mathbf{A} . Most of the loss functions proposed in the literature do not distinguish the statistical difference between noise and outliers appropriately, as mentioned in Section 1.1.4.2. Specifically, sparsity-inducing loss functions such as the MC loss are sensitive to small perturbations since the derivative does not vanish as it approaches the origin. To overcome this issue, we propose the formulation integrating the MC-based sparse signal recovery and SORR to reflect the sparsity of outliers with stability in highly noisy situations.

4.3 Proposed Method

To attain both benefits of (1.3) and (F.1.3), we blend those two formulations as follows:

$$\begin{aligned} \min_{\mathbf{x} \in \mathbb{R}^n, \boldsymbol{\varepsilon} \in \mathbb{R}^m} J_{\text{S-SORR}}(\mathbf{x}, \boldsymbol{\varepsilon}) := & \\ & \alpha \left(\mu_{\text{SORR}} \Phi_{\gamma_1}^{\text{MC}}(\mathbf{y} - (\mathbf{A}\mathbf{x} + \boldsymbol{\varepsilon})) + \frac{1}{2} \|\mathbf{x}\|_2^2 + \frac{\rho_{\text{SORR}}}{2} \|\boldsymbol{\varepsilon}\|_2^2 \right) \\ & + (1 - \alpha) \left(\frac{1}{2} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \mu_{\text{MC}} \Phi_{\gamma_2}(\mathbf{x}) \right), \end{aligned} \quad (4.2)$$

where $\alpha \in (0, 1]$ controls the balance between sparseness and robustness, and $\gamma_1, \gamma_2 > 0$. For convenience, let $\mu_1 := \alpha\mu_{\text{SORR}} > 0$, $\mu_2 := (1 - \alpha)\mu_{\text{MC}} \geq 0$, and $\rho := \alpha\rho_{\text{SORR}} > 0$, respectively. Then, (4.2) can be rewritten as follows:

$$\begin{aligned} \min_{\mathbf{x} \in \mathbb{R}^n, \boldsymbol{\varepsilon} \in \mathbb{R}^m} & \frac{1}{2} (\alpha \|\mathbf{x}\|_2^2 + (1 - \alpha) \|\mathbf{y} - \mathbf{Ax}\|_2^2) + \mu_1 \Phi_{\gamma_1}^{\text{MC}}(\mathbf{y} - (\mathbf{Ax} + \boldsymbol{\varepsilon})) \\ & + \mu_2 \Phi_{\gamma_2}^{\text{MC}}(\mathbf{x}) + \frac{\rho}{2} \|\boldsymbol{\varepsilon}\|_2^2. \end{aligned} \quad (4.3)$$

This formulation is referred to as *sparse stable outlier-robust regression (SSORR)*. Here, we regard μ_1 , μ_2 , and ρ_1 as parameters to be tuned independently from α .

One may consider that the quadratic terms in (4.3) would annihilate the benefits of the MC terms. However, it is well known that the quadratic loss in the problem

$$\min_{\mathbf{x} \in \mathbb{R}^n} \frac{1}{2} \|\mathbf{y} - \mathbf{Ax}\|_2^2 + \mu_{\text{MC}} \Phi_{\gamma_2}(\mathbf{x}) \quad (4.4)$$

does not annihilate the benefit of the MC penalty [24, 25]. In a similar argument to Section 3.2.1, it can be seen that the quadratic penalties in the problem

$$\min_{\mathbf{x} \in \mathbb{R}^n, \boldsymbol{\varepsilon} \in \mathbb{R}^m} \mu_{\text{SORR}} \Phi_{\gamma_1}^{\text{MC}}(\mathbf{y} - (\mathbf{Ax} + \boldsymbol{\varepsilon})) + \frac{1}{2} \|\mathbf{x}\|_2^2 + \frac{\rho_{\text{SORR}}}{2} \|\boldsymbol{\varepsilon}\|_2^2 \quad (4.5)$$

would not annihilate the benefit of the MC loss. Hence, such annihilation would not occur in (4.3) as well.

In the particular case of $\alpha = 1$, the proposed formulation (4.3) reduces to the formulation in [130], providing remarkably robust estimates. In this case, however, μ_1 and μ_2 need to be sufficiently small for the overall cost to be convex. This means that $\|\mathbf{x}\|_2^2/2$ is relatively large compared to the other terms, and it may thus cause a large estimation bias which might cancel the nearly-unbiased-estimation property of $\mu_2 \Phi_{\gamma_2}^{\text{MC}}(\mathbf{x})$. Therefore, for a smaller α , one may choose μ_1 and μ_2 in such a way that the cost function of

$$\mu_1 \Phi_{\gamma_1}^{\text{MC}}(\mathbf{y} - (\mathbf{Ax} + \boldsymbol{\varepsilon})) + \frac{\alpha}{2} \|\mathbf{x}\|_2^2 + \frac{\rho}{2} \|\boldsymbol{\varepsilon}\|_2^2 \quad (4.6)$$

and

$$\mu_2 \Phi_{\gamma_2}^{\text{MC}}(\mathbf{x}) + \frac{1 - \alpha}{2} \|\mathbf{y} - \mathbf{Ax}\|_2^2 \quad (4.7)$$

are convex, respectively. By doing so, a sparse and less biased estimate could be attained while keeping the remarkable outlier-robustness (see Section 4.5.2.2). One can also see that (4.3) involves the two penalty terms $\mu_2 \Phi_{\gamma_2}^{\text{MC}}(\mathbf{x})$ and $\alpha \|\mathbf{x}\|_2^2/2$, and this actually induces a notable grouping effect which essentially differs from that of the popular elastic net, as will be discussed in Section 4.3.4.

4.3.1 Reformulation of the Problem

Problem (4.3) can be rewritten in a simpler form with $\boldsymbol{\xi} := [\mathbf{x}^\top \boldsymbol{\varepsilon}^\top]^\top \in \mathbb{R}^{n+m}$ as follows:

$$\begin{aligned} \min_{\boldsymbol{\xi} \in \mathbb{R}^{n+m}} \frac{1}{2} \left\| \begin{bmatrix} \boldsymbol{\Theta}_3^{1/2} \\ (1-\alpha)^{1/2} \mathbf{A} \boldsymbol{\Theta}_2 \end{bmatrix} \boldsymbol{\xi} - \begin{bmatrix} \mathbf{0}_{n+m} \\ (1-\alpha)^{1/2} \mathbf{y} \end{bmatrix} \right\|_2^2 &+ \mu_1 \Phi_{\gamma_1}^{\text{MC}}(\boldsymbol{\Theta}_1 \boldsymbol{\xi} - \mathbf{y}) \\ &+ \mu_2 \Phi_{\gamma_2}^{\text{MC}}(\boldsymbol{\Theta}_2 \boldsymbol{\xi}), \end{aligned} \quad (4.8)$$

where¹

$$\boldsymbol{\Theta}_1 := [\mathbf{A} \ \mathbf{I}_m] \in \mathbb{R}^{m \times (n+m)}, \quad (4.9)$$

$$\boldsymbol{\Theta}_2 := [\mathbf{I}_n \ \mathbf{0}_{n \times m}] \in \mathbb{R}^{n \times (n+m)}, \quad (4.10)$$

$$\boldsymbol{\Theta}_3 := \begin{bmatrix} \alpha \mathbf{I}_n & \mathbf{0}_{n \times m} \\ \mathbf{0}_{m \times n} & \rho \mathbf{I}_m \end{bmatrix} \in \mathbb{R}^{(n+m) \times (n+m)}. \quad (4.11)$$

Since the MC function satisfies (2.46), the cost function in Problem (4.8) can be split into smooth and nonsmooth proximable terms as follows:

$$\begin{aligned} &\frac{1}{2} \left\| \begin{bmatrix} \boldsymbol{\Theta}_3^{1/2} \\ (1-\alpha)^{1/2} \mathbf{A} \boldsymbol{\Theta}_2 \end{bmatrix} \boldsymbol{\xi} - \begin{bmatrix} \mathbf{0}_{n+m} \\ (1-\alpha)^{1/2} \mathbf{y} \end{bmatrix} \right\|_2^2 \\ &\quad + \mu_1 (\|\boldsymbol{\Theta}_1 \boldsymbol{\xi} - \mathbf{y}\|_1 - \gamma_1 \|\cdot\|_1(\boldsymbol{\Theta}_1 \boldsymbol{\xi} - \mathbf{y})) + \mu_2 (\|\boldsymbol{\Theta}_2 \boldsymbol{\xi}\|_1 - \gamma_2 \|\cdot\|_1(\boldsymbol{\Theta}_2 \boldsymbol{\xi})) \\ &= F(\boldsymbol{\xi}) + G(\boldsymbol{\Theta} \boldsymbol{\xi}), \end{aligned} \quad (4.12)$$

where $\boldsymbol{\Theta} := \begin{bmatrix} \mu_1 \boldsymbol{\Theta}_1 \\ \mu_2 \boldsymbol{\Theta}_2 \end{bmatrix} \in \mathbb{R}^{(n+m) \times (n+m)}$ and

$$\begin{aligned} F : \mathbb{R}^{(n+m) \times (n+m)} : \boldsymbol{\xi} \mapsto &\frac{1}{2} \left\| \begin{bmatrix} \boldsymbol{\Theta}_3^{1/2} \\ (1-\alpha)^{1/2} \mathbf{A} \boldsymbol{\Theta}_2 \end{bmatrix} \boldsymbol{\xi} - \begin{bmatrix} \mathbf{0}_{n+m} \\ (1-\alpha)^{1/2} \mathbf{y} \end{bmatrix} \right\|_2^2 \\ &- \mu_1 \gamma_1 \|\cdot\|_1(\boldsymbol{\Theta}_1 \boldsymbol{\xi} - \mathbf{y}) - \mu_2 \gamma_2 \|\cdot\|_1(\boldsymbol{\Theta}_2 \boldsymbol{\xi}), \end{aligned} \quad (4.13)$$

$$G : \mathbb{R}^{(n+m) \times (n+m)} : \boldsymbol{\xi} \mapsto \left\| \boldsymbol{\xi} - \begin{bmatrix} \mu_1 \mathbf{y} \\ \mathbf{0}_n \end{bmatrix} \right\|_1. \quad (4.14)$$

The function G is clearly convex and nonsmooth, and its proximity operator can be expressed in a closed form. On the other hand, F is smooth, *i.e.*, F is differentiable and its gradient ∇F is Lipschitz continuous with constant (see Appendix J.1 for its derivation)

$$\beta := \max\{\alpha, \rho\} + (1 - \alpha + \mu_1 \gamma_1^{-1}) \lambda_{\max}(\mathbf{A}^\top \mathbf{A}) + \mu_1 \gamma_1^{-1} + \mu_2 \gamma_2^{-1}. \quad (4.15)$$

¹Note that $\boldsymbol{\Theta}_3$ corresponds to $\boldsymbol{\Sigma}_\xi$ in [35], which is the covariance matrix of $\boldsymbol{\xi}$, following a Gaussian distribution. The notation has been changed, because $\boldsymbol{\xi}$ is non-Gaussian in this study due to the sparseness assumption of \mathbf{x} , and because $\boldsymbol{\Theta}_3$ is no longer the covariance matrix of $\boldsymbol{\xi}$.

The convexity conditions of the smooth part F in (4.13) are shown below so that the forward-backward-based primal-dual algorithm [118] (see Appendix C) can be applied to solve Problem (4.3), which produces Algorithm 4.1.

4.3.2 Convexity Conditions

The convexity conditions of F are given as follows.

Proposition 4.1. The function F in (4.13) is convex if

$$0 \leq \mu_2 \leq \gamma_2 \left[\alpha - \left(\frac{\mu_1 \rho}{\rho \gamma_1 - \mu_1} - 1 + \alpha \right) \lambda_{\max}(\mathbf{A}^\top \mathbf{A}) \right] \quad (4.16)$$

This condition is necessary when

$$K := \{[\mathbf{x}^\top \ \boldsymbol{\varepsilon}^\top]^\top \in \mathbb{R}^{n+m} \mid \|\mathbf{y} - \mathbf{A}\mathbf{x} - \boldsymbol{\varepsilon}\|_\infty \leq \gamma_1, \|\mathbf{x}\|_\infty \leq \gamma_2\} \neq \emptyset \quad (4.17)$$

has a nonempty interior.

Proof. The proof is given in Appendix J.2 based on the general results to be presented in Section 4.4. An alternative proof of this proposition is given in Appendix J.3. \square

In practice, an equivalent convexity condition stated in the following lemma is useful to set the hyperparameters μ_1 and μ_2 :

Corollary 4.1. The convexity condition in (4.16) holds if and only if the following two conditions are jointly satisfied:

$$(K-I) \quad \mu_1 \leq \frac{\gamma_1 \rho (\alpha + (1 - \alpha) \lambda_{\max}(\mathbf{A}^\top \mathbf{A}))}{\alpha + (\rho + 1 - \alpha) \lambda_{\max}(\mathbf{A}^\top \mathbf{A})},$$

$$(K-II) \quad \mu_2 \leq \gamma_2 \left[\alpha - \left(\frac{\mu_1 \rho}{\rho \gamma_1 - \mu_1} - 1 + \alpha \right) \lambda_{\max}(\mathbf{A}^\top \mathbf{A}) \right].$$

We mention that, when F is convex, the whole cost function is also convex since G is convex. Suppose that (K-I) and (K-II) are jointly satisfied with strict inequalities. In this case, F , and hence the whole cost function, are strictly convex, and thus the existence and uniqueness of a solution of the whole cost function is guaranteed by [96, Corollary 11.16].² Note here that the existence follows from the fact that (i) the first term in (4.8) is coercive, *i.e.*,

$$\lim_{\|\boldsymbol{\xi}\|_2 \rightarrow +\infty} \frac{1}{2} \left\| \begin{bmatrix} \boldsymbol{\Theta}_3^{1/2} \\ (1 - \alpha)^{1/2} \mathbf{A} \boldsymbol{\Theta}_2 \end{bmatrix} \boldsymbol{\xi} - \begin{bmatrix} \mathbf{0}_{n+m} \\ (1 - \alpha)^{1/2} \mathbf{y} \end{bmatrix} \right\|_2^2 = +\infty, \quad (4.18)$$

²Let $f, g : \mathcal{H} \rightarrow (-\infty, +\infty]$ be defined over a given Hilbert space \mathcal{H} . To show the existence of a minimizer of $f + g \in \Gamma_0(\mathcal{H})$ for $f, g \in \Gamma_0(\mathcal{H})$, it suffices that f is coercive and g is bounded below [96, Corollary 11.16].

Algorithm 4.1 : Sparse Stable Outlier-Robust Regression (S-SORR)

Set $\mathbf{x}_0, \mathbf{w}_0 \in \mathbb{R}^n$, $\boldsymbol{\varepsilon}_0, \mathbf{v}_0 \in \mathbb{R}^m$, and $\tau, \varsigma > 0$.

For $k = 0, 1, 2, \dots$

$$\begin{aligned} \begin{bmatrix} \mathbf{x}_{k+\frac{1}{2}} \\ \boldsymbol{\varepsilon}_{k+\frac{1}{2}} \end{bmatrix} &:= \begin{bmatrix} \mathbf{x}_k \\ \boldsymbol{\varepsilon}_k \end{bmatrix} - \tau \nabla F \left(\begin{bmatrix} \mathbf{x}_k \\ \boldsymbol{\varepsilon}_k \end{bmatrix} \right), \\ \begin{bmatrix} \mathbf{v}_{k+1} \\ \mathbf{w}_{k+1} \end{bmatrix} &:= (\text{Id} - \text{Prox}_{(\tau/\varsigma)G}) \left\{ \begin{bmatrix} \mu_1 \left(\mathbf{A}\mathbf{x}_{k+\frac{1}{2}} + \boldsymbol{\varepsilon}_{k+\frac{1}{2}} \right) \\ \mu_2 \mathbf{x}_{k+\frac{1}{2}} \end{bmatrix} \right. \\ &\quad \left. + \left(\mathbf{I}_{n+m} - \varsigma \begin{bmatrix} \mu_1^2 (\mathbf{A}\mathbf{A}^\top + \mathbf{I}_m) & \mu_1 \mu_2 \mathbf{A} \\ \mu_1 \mu_2 \mathbf{A}^\top & \mu_2^2 \mathbf{I}_n \end{bmatrix} \right) \begin{bmatrix} \mathbf{v}_k \\ \mathbf{w}_k \end{bmatrix} \right\}, \\ \begin{bmatrix} \mathbf{x}_{k+1} \\ \boldsymbol{\varepsilon}_{k+1} \end{bmatrix} &:= \begin{bmatrix} \mathbf{x}_{k+\frac{1}{2}} \\ \boldsymbol{\varepsilon}_{k+\frac{1}{2}} \end{bmatrix} - \varsigma \begin{bmatrix} \mu_1 \mathbf{A}^\top \mathbf{v}_{k+1} + \mu_2 \mathbf{w}_{k+1} \\ \mu_1 \mathbf{v}_{k+1} \end{bmatrix}. \end{aligned}$$

and (ii) the remaining MC functions in (4.8) are bounded below.

With the parameters μ_1 and μ_2 satisfying (K-I) and (K-II), Problem (4.3) can be solved by the following forward-backward-based primal-dual algorithm [117, 118, 131].

By Fact 2.4(b), it holds, for any $\boldsymbol{\xi} \in \mathbb{R}^{n+m}$, that

$$\begin{aligned} \nabla F(\boldsymbol{\xi}) &= (\boldsymbol{\Theta}_3^{1/2})^\top \boldsymbol{\Theta}_3^{1/2} \boldsymbol{\xi} + (1 - \alpha)(\mathbf{A}\boldsymbol{\Theta}_2)^\top (\mathbf{A}\boldsymbol{\Theta}_2 \boldsymbol{\xi} - \mathbf{y}) \\ &\quad - \mu_1 \gamma_1^{-1} \boldsymbol{\Theta}_1^\top (\text{Id} - \text{Soft}_{\gamma_1})(\boldsymbol{\Theta}_1 \boldsymbol{\xi} - \mathbf{y}) - \mu_2 \gamma_2^{-1} \boldsymbol{\Theta}_2^\top (\text{Id} - \text{Soft}_{\gamma_2})(\boldsymbol{\Theta}_2 \boldsymbol{\xi}) \\ &= \begin{bmatrix} \nabla_{\mathbf{x}} F(\boldsymbol{\xi}) \\ \nabla_{\boldsymbol{\varepsilon}} F(\boldsymbol{\xi}) \end{bmatrix}, \end{aligned} \quad (4.19)$$

where

$$\begin{aligned} \nabla_{\mathbf{x}} F : \mathbb{R}^{n+m} \rightarrow \mathbb{R}^n : \boldsymbol{\xi} \mapsto &\alpha \mathbf{x} + (1 - \alpha) \mathbf{A}^\top (\mathbf{A}\mathbf{x} - \mathbf{y}) \\ &- \mu_1 \gamma_1^{-1} \mathbf{A}^\top (\text{Id} - \text{Soft}_{\gamma_1})(\mathbf{A}\mathbf{x} + \boldsymbol{\varepsilon} - \mathbf{y}) \\ &- \mu_2 \gamma_2^{-1} (\text{Id} - \text{Soft}_{\gamma_2})(\mathbf{x}), \end{aligned} \quad (4.20)$$

$$\nabla_{\boldsymbol{\varepsilon}} F : \mathbb{R}^{n+m} \rightarrow \mathbb{R}^m : \boldsymbol{\xi} \mapsto \rho \boldsymbol{\varepsilon} - \mu_1 \gamma_1^{-1} (\text{Id} - \text{Soft}_{\gamma_1})(\mathbf{A}\mathbf{x} + \boldsymbol{\varepsilon} - \mathbf{y}). \quad (4.21)$$

By Fact 2.3(a), the proximity operator appearing in Algorithm 4.1 can be expressed, for any $\boldsymbol{\zeta} \in \mathbb{R}^{m+n}$, by

$$\text{Prox}_{(\tau/\varsigma)G} : \mathbb{R}^{n+m} \rightarrow \mathbb{R}^{n+m} : \boldsymbol{\zeta} \mapsto \begin{bmatrix} \mu_1 \mathbf{y} \\ \mathbf{0}_n \end{bmatrix} + \text{Soft}_{\tau/\varsigma} \left(\boldsymbol{\zeta} - \begin{bmatrix} \mu_1 \mathbf{y} \\ \mathbf{0}_n \end{bmatrix} \right). \quad (4.22)$$

4.3.3 Convergence Conditions and Complexity

The following proposition gives a guarantee of convergence of Algorithm 1 to a solution of (4.3).

Proposition 4.2. Assume that convexity conditions (K-I) and (K-II) are satisfied. Let $(\mathbf{x}_k, \boldsymbol{\varepsilon}_k)_{k \in \mathbb{N}}$ and $(\mathbf{v}_k, \mathbf{w}_k)_{k \in \mathbb{N}}$ be the sequences generated by Algorithm 4.1, respectively. Suppose that

$$0 < \tau < \frac{2}{\beta}, \quad 0 < \varsigma \leq \frac{1}{\lambda_{\max}(\boldsymbol{\Theta}^\top \boldsymbol{\Theta})}. \quad (4.23)$$

Then, the sequence $(\mathbf{x}_k, \boldsymbol{\varepsilon}_k)_{k \in \mathbb{N}}$ converges to a solution of the problem in (4.3), and $((\tau/\varsigma)\mathbf{v}_k, (\tau/\varsigma)\mathbf{w}_k)_{k \in \mathbb{N}}$ converges to a solution of the associated dual problem

$$\min_{\boldsymbol{\zeta} \in \mathbb{R}^{m+n}} F^*(-\boldsymbol{\Theta}^\top \boldsymbol{\zeta}) + G^*(\boldsymbol{\zeta}). \quad (4.24)$$

Proof. The proof is due to Fact C.1 in light of Proposition 4.1. \square

We remark that the computational complexity to obtain $\lambda_{\max}(\boldsymbol{\Theta}^\top \boldsymbol{\Theta})$ is $\mathcal{O}(nmp)$ with the number p of power iterations. This is verified as follows: for any $\mathbf{x} \in \mathbb{R}^n$ and $\boldsymbol{\varepsilon} \in \mathbb{R}^m$, it holds that

$$\boldsymbol{\Theta}^\top \boldsymbol{\Theta} \begin{bmatrix} \mathbf{x} \\ \boldsymbol{\varepsilon} \end{bmatrix} = \begin{bmatrix} (\mu_1^2 \mathbf{A}^\top \mathbf{A} + \mu_2^2 \mathbf{I}_n) \mathbf{x} + \mu_1^2 \mathbf{A}^\top \boldsymbol{\varepsilon} \\ \mu_1^2 \mathbf{A} \mathbf{x} + \mu_1^2 \boldsymbol{\varepsilon} \end{bmatrix}, \quad (4.25)$$

and the complexity of $\mathbf{A}^\top \mathbf{A} \mathbf{x}$ and $\mathbf{A}^\top \boldsymbol{\varepsilon}$ are $\mathcal{O}(nm)$. Since the normalization scales in $\mathcal{O}(n+m)$, the complexity for $\lambda_{\max}(\boldsymbol{\Theta}^\top \boldsymbol{\Theta})$ amounts to $\mathcal{O}(nmp)$. The total complexity of Algorithm 4.1 is, on the other hand, $\mathcal{O}(nmq)$ with the number q of algorithm iterations since each iteration of Algorithm 4.1 requires $\mathcal{O}(nm)$ complexity. The computational complexities of the other sparse outlier-robust recovery methods are summarized in Table 4.1. We consider the computational complexities when YALL1 and extended lasso are solved by the ADMM [132] and the primal-dual splitting method. RPGG [72] requires $\mathcal{O}(m^2(m+n))$ complexity to obtain the pseudoinverse of $[\mathbf{A} \ \mathbf{I}_m]$ in the initialization step and $\mathcal{O}(m(n+m))$ complexity per iteration. Hence, the total complexity for RPGG is $\mathcal{O}(m(m+q)(n+m))$, which is larger than that of S-SORR for a sufficiently large m . The complexity of S-SORR is comparable to the other methods based on convex optimization (extended lasso, Huber FISTA, and YALL1).

4.3.4 Grouping Effect

It is well known that, when some groups of features are highly correlated, lasso tends to choose only one feature in the group (see **issue (iv)** in Section 1.1.1), and the elastic net penalty [83] overcomes this issue by adding a quadratic function to the lasso formulation (see Section 1.1.5). We show below that the MC penalty by itself shares the same issue as lasso, and the grouping effect can be attained by adding an extra penalty of the quadratic function in analogy to the elastic net penalty. While we assume that two

Table 4.1: Total computational complexity (p : number of power iterations, q : number of algorithm iterations).

	computational complexity
extended lasso	$\mathcal{O}(nm(p+q))$
Huber FISTA	$\mathcal{O}(nm(p+q))$
YALL1	$\mathcal{O}(nmq)$
RPGG	$\mathcal{O}(m(m+q)(n+m))$
S-SORR	$\mathcal{O}(nm(p+q))$

features are highly correlated in the following discussions, the case when more than two vectors are highly correlated can also be treated straightforwardly by applying the results to each pair of such features.

Let $\hat{\mathbf{x}} \in \mathbb{R}^n$ be a minimizer of the convex cost function

$$J(\mathbf{x}) := H(\mathbf{A}\mathbf{x}) + R(\mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^n, \quad (4.26)$$

where $H : \mathbb{R}^m \rightarrow (-\infty, +\infty]$ is a (possibly nonconvex) loss function and $R : \mathbb{R}^n \rightarrow (-\infty, +\infty]$ is a (possibly nonconvex) regularizer, respectively.

Proposition 4.3. Assume that $J(\mathbf{x})$ has a minimizer $\hat{\mathbf{x}}$, there exist some indices $i, j \in \{1, 2, \dots, n\}$ such that $\mathbf{a}_i = \mathbf{a}_j$, $|\hat{x}_i|, |\hat{x}_j| > \gamma$, and $\hat{x}_i \hat{x}_j > 0$ for a given $\gamma > 0$. Then, the following statements hold.

- (a) Let $R(\mathbf{x}) := \mu \Phi_\gamma^{\text{MC}}(\mathbf{x})$. Then, $\hat{\mathbf{x}}^*$ is another minimizer of $J(\mathbf{x})$ in (4.26) other than $\hat{\mathbf{x}}$, where

$$\hat{x}_k^* := \begin{cases} \hat{x}_k, & k \neq i \text{ and } k \neq j, \\ (\hat{x}_i + \hat{x}_j)\alpha, & k = i, \\ (\hat{x}_i + \hat{x}_j)(1 - \alpha), & k = j, \end{cases} \quad (4.27)$$

for any

$$\alpha \in \left[\frac{\gamma}{|\hat{x}_i| + |\hat{x}_j|}, \frac{1}{2} \right]. \quad (4.28)$$

- (b) Let $R(\mathbf{x}) := \mu \Phi_\gamma^{\text{MC}}(\mathbf{x}) + \|\mathbf{x}\|_2^2/2$. Then, $\hat{x}_i = \hat{x}_j$ for any $\mu > 0$.

Proof. See Appendix J.4. □

Proposition 4.3(a) states that, while two column vectors \mathbf{a}_i and \mathbf{a}_j are identical, the corresponding components of a minimizer of (4.26) are different excluding the case of $\alpha = 1/2$. Since the normalized MC penalty gives a parametric bridge between $\|\cdot\|_0$ (defined as the number of nonzero components of a vector) and $\|\cdot\|_1$ [34], γ is desired to be small to attain a sparse estimate with small bias. However, a small γ allows α to be nearly zero,

making $\hat{x}_i^* = (\hat{x}_i + \hat{x}_j)\alpha$ nearly zero as well. This is undesirable in such a situation when the two variables need to be identified as important components [83]. In contrast, Proposition 4.3(b) states that the MC penalty regularized by the quadratic function yields the same coefficients for the two features, *i.e.*, it exhibits the grouping effect. In the specific case of the MC loss, the result of Proposition 4.3(b) can be generalized to the case when two features are highly correlated as below.

Proposition 4.4. Suppose that each column of \mathbf{A} is normalized as $\|\mathbf{a}_i\|_2 = 1$ for $i = 1, 2, \dots, n$. Let

$$H := \mu_1 \Phi_{\gamma_1}^{\text{MC}}(\cdot - \mathbf{y} - \boldsymbol{\varepsilon}) + \frac{1 - \alpha}{2} \|\cdot - \mathbf{y}\|_2^2 \quad (4.29)$$

for given $\gamma_1, \mu_1 > 0$ and for given $\mathbf{y}, \boldsymbol{\varepsilon} \in \mathbb{R}^m$, and let

$$R := \mu_2 \Phi_{\gamma_2}^{\text{MC}} + \frac{\alpha}{2} \|\cdot\|_2^2 \quad (4.30)$$

for given $\gamma_2 > 0$, $\mu_2 \geq 0$, and $\alpha \in (0, 1]$. Assume that the following conditions hold:

- (a) $\tau_1 + \lambda_{\min}(\tau_2 \mathbf{A}^\top \mathbf{A}) \geq 0$ (\Leftarrow (K-II)), where $\tau_1 := \mu_1^{-1} \gamma_1 (\alpha - \mu_2 \gamma_2^{-1})$ and $\tau_2 := (1 - \alpha) \mu_1^{-1} \gamma_1 - 1$.
- (b) $\hat{x}_i \hat{x}_j > 0$.

Then, it holds that

$$\begin{aligned} |\hat{x}_i - \hat{x}_j| \leq \alpha^{-1} & \left[\sqrt{2(1 - \mathbf{a}_i^\top \mathbf{a}_j)(\mu_1^2 m + (1 - \alpha)\mu_1 \gamma_1 + (1 - \alpha)^2 \|\mathbf{y}\|_2^2)} \right. \\ & \left. + \mu_2 |(\phi_{\gamma_2}^{\text{MC}})'(\hat{x}_i) - (\phi_{\gamma_2}^{\text{MC}})'(\hat{x}_j)| \right], \end{aligned} \quad (4.31)$$

where $(\phi_{\gamma_2}^{\text{MC}})' : \mathbb{R} \setminus \{0\} \rightarrow \mathbb{R}$ is the derivative of the MC function defined in (2.45). Note that $\mathbf{a}_i^\top \mathbf{a}_j \in [-1, 1]$ is the sample correlation.

Proof. The proof is given in Appendix J.5. \square

Proposition 4.4 states that the difference between the corresponding coefficients \hat{x}_i and \hat{x}_j becomes smaller when the correlation between \mathbf{a}_i and \mathbf{a}_j becomes larger. The upper bound in (4.31) provides a quantitative measure of the grouping effect. Specifically, the following corollary provides a contrasting result to the naïve elastic net by considering the case when $\alpha = 1$ and $\gamma_2 \rightarrow +\infty$.

Corollary 4.2. Suppose that each column of \mathbf{A} is normalized as $\|\mathbf{a}_i\|_2 = 1$ for $i = 1, 2, \dots, n$. Let

$$H := \mu_1 \Phi_{\gamma_1}^{\text{MC}}(\cdot - \mathbf{y} - \boldsymbol{\varepsilon}) \quad (4.32)$$

for given $\gamma_1, \mu_1 > 0$ and for given $\mathbf{y}, \boldsymbol{\varepsilon} \in \mathbb{R}^m$, and let

$$R := \mu_2 \|\cdot\|_1 + \frac{1}{2} \|\cdot\|_2^2 \quad (4.33)$$

for given $\mu_2 \geq 0$. Assume that the following conditions hold:

- (a) $\mu_1 \lambda_{\max}(\mathbf{A}^\top \mathbf{A}) \leq \gamma_1$ (\Leftarrow (K-II)).
- (b) $\hat{x}_i \hat{x}_j > 0$.

Then, it holds that

$$|\hat{x}_i - \hat{x}_j| \leq \mu_1 \sqrt{2m(1 - \mathbf{a}_i^\top \mathbf{a}_j)}. \quad (4.34)$$

According to [83, Theorem 1], for the naïve elastic net, which corresponds to $H(\mathbf{z}) := \frac{\mu_1}{2} \|\mathbf{z} - \mathbf{y}\|_2^2$ and $R(\mathbf{x}) := \mu_2 \|\mathbf{x}\|_1 + \frac{1}{2} \|\mathbf{x}\|_2^2$, it holds that

$$|\hat{x}_i - \hat{x}_j| \leq \mu_1 \|\mathbf{y}\|_2 \sqrt{2(1 - \mathbf{a}_i^\top \mathbf{a}_j)}, \quad (4.35)$$

provided that $\hat{x}_i \hat{x}_j > 0$. Comparing (4.34) and (4.35), the upper bound in (4.34) depends solely on the number m of measurements, while that in (4.35) depends on $\|\mathbf{y}\|_2$. Therefore, for a given m , even when the magnitude of the measurements becomes severely large due to the influence of outliers, the upper bound remains constant in the case of the MC loss function. We note that, while a small α yields a larger bound in (4.31), the grouping effect and recovery performance are not proportional in general, and hence the hyperparameters should be set in consideration of these factors at the same time. Proposition 4.4 can directly be applied to the proposed formulation (4.3) because (4.31) holds for an arbitrary $\boldsymbol{\varepsilon} \in \mathbb{R}^m$.

Finally, the following proposition holds independently of Proposition 4.4.

Proposition 4.5. Suppose that each column of \mathbf{A} is normalized as $\|\mathbf{a}_i\|_2 = 1$ for $i = 1, 2, \dots, n$. Set H as in (4.32) for given $\gamma_1, \mu_1 > 0$ and for given $\mathbf{y}, \boldsymbol{\varepsilon} \in \mathbb{R}^m$, and let

$$R := \mu_2 \Phi_{\gamma_2}^{\text{MC}} + \frac{1}{2} \|\cdot\|_2^2 \quad (4.36)$$

for given $\gamma_2 > 0$, respectively. Assume that the following conditions hold:

- (a) $\mu_1 \lambda_{\max}(\mathbf{A}^\top \mathbf{A}) \leq \gamma_1$ (\Leftarrow (K-II)).
- (b) There exist $i, j \in \{1, 2, \dots, n\}$ such that $|\hat{x}_i|, |\hat{x}_j| \geq \gamma_2$.

Then, for any $\mu_2 \geq 0$, it holds that

$$|\hat{x}_i - \hat{x}_j| \leq \mu_1 \sqrt{2m(1 - \mathbf{a}_i^\top \mathbf{a}_j)}. \quad (4.37)$$

Note that $\mathbf{a}_i^\top \mathbf{a}_j \in [-1, 1]$ is the sample correlation.

Proof. The proof is given in Appendix J.6. \square

Insights from Propositions 4.3, 4.4, and 4.5 can be summarized as below:

- For an arbitrary loss function satisfying the assumption that $J(\mathbf{x})$ has a minimizer, Proposition 4.3 states the following for the extreme situation when two features are identical.
 - (a) $\mu_2 \Phi_\gamma^{\text{MC}}(\mathbf{x})$ (with no quadratic function) may select one of the correlated features, and $J(\mathbf{x})$ has infinitely many solutions.
 - (b) $\mu_2 \Phi_\gamma^{\text{MC}}(\mathbf{x}) + \frac{1}{2} \|\mathbf{x}\|_2^2$ successfully selects both correlated features.
- Proposition 4.4 concerns the particular case of the double use of the MC and quadratic functions for both loss and penalty. Specifically, for the case of the MC loss with the elastic net penalty (Corollary 4.2), the difference between the corresponding coefficients to the correlated features decreases when the correlation becomes higher. Remarkably, the bound in (4.34) is independent of the output vector norm $\|\mathbf{y}\|_2$, which can be severely large due to the outliers. This is in sharp contrast to the naïve elastic net.
- Proposition 4.5 gives the same bound as Corollary 4.2 with different assumptions. While assumption (b) is stricter than that of Corollary 4.2, assumption (a) is weaker and the MC penalty is considered in the function R , which is more general than the ℓ_1 norm. This proposition can be helpful, especially in applications in which the grouping effect is crucial (see Remark 4.1 in Section 4.5.2).

4.4 Convexity Conditions for a General Model

This section shows the convexity conditions for a general model including the formulation of S-SORR as a special case within the framework of LiMES model (see Appendix F.2). Let $(\mathcal{X}, \langle \cdot, \cdot \rangle_{\mathcal{X}})$ and $(\mathcal{Z}_i, \langle \cdot, \cdot \rangle_{\mathcal{Z}_i})$, $i = 1, 2, \dots, Q + 1$, be finite-dimensional real Hilbert spaces. Let $\Psi_i \in \Gamma_0(\mathcal{Z}_i)$, $i = 2, \dots, Q + 1$, be functions defined over \mathcal{Z}_i , and $D_i : \mathcal{Z}_i \rightarrow \mathcal{Z}_i$ be diagonal positive-definite operators. For $i = 1, 2, \dots, Q + 1$, let $\mathcal{A}_i : \mathcal{X} \rightarrow \mathcal{Z}_i : x \mapsto M_i x + c_i$ be an affine operator with a bounded linear operator $(0 \neq) M_i : \mathcal{X} \rightarrow \mathcal{Z}_i$ and a vector $c_i \in \mathcal{Z}_i$. Here, \mathcal{Z}_1 corresponds to \mathcal{Y} appearing in Appendix F.2.

Let us now consider the following problem:

$$\min_{x \in \mathcal{X}} \left[\frac{1}{2} \|\mathcal{A}_1 x\|_{\mathcal{Z}_1}^2 + \sum_{i=2}^{Q+1} \nu_i (\Psi_i)_{D_i}(\mathcal{A}_i x) \right], \quad (4.38)$$

where $\nu_i > 0$. The problem (4.38) reduces to the S-SORR formulation in (4.8) when $\mu_2 > 0$ by letting $Q := 2$, $\mathcal{X} := \mathcal{Z}_2 := \mathcal{Z}_3 := \mathbb{R}^{n+m}$, $\mathcal{Z}_1 := \mathbb{R}^{n+2m}$,

$x := \boldsymbol{\xi}$, $\nu_2 := \mu_1$, $\nu_3 := \mu_2$, $\mathcal{A}_1 : x \mapsto \begin{bmatrix} \boldsymbol{\Theta}_3^{1/2} \\ (1-\alpha)^{1/2}\boldsymbol{\Theta}_1 \end{bmatrix} x - \begin{bmatrix} \mathbf{0}_{n+m} \\ (1-\alpha)^{1/2}\mathbf{y} \end{bmatrix}$,
 $\mathcal{A}_2 : x \mapsto \boldsymbol{\Theta}_1 x - \mathbf{y}$, $\mathcal{A}_3 := \boldsymbol{\Theta}_2$, $D_2 := \gamma_1^{-1/2}\mathbf{I}_m$, $D_3 := \gamma_2^{-1/2}\mathbf{I}_{n+m}$, and
 $\Psi_2 := \Psi_3 := \|\cdot\|_1$. The formulation (4.38) with $\nu_2 := \mu_1$ and $Q := 1$
(equivalent to the LiMES model in (F.2.1)) reduces to the SORR formulation
(which corresponds to the case of $\mu_2 = 0$ in (4.8)).

Another possible application of the problem (4.38) is the robust sparse classification. Let $\mathcal{X} := \mathcal{Z}_1 := \mathcal{Z}_3 := \mathbb{R}^n$, $\mathcal{Z}_2 := \mathbb{R}^m$, $\nu_2 := \mu_1$, $\nu_3 := \mu_2$, $\mathcal{A}_1 := \mathcal{A}_3 := \mathbf{I}_n$, $\mathcal{A}_2 : x \mapsto \mathbf{M}_2 x - \mathbf{1}_m$, where $\mathbf{M}_2 := [y_1 \mathbf{a}_1 \ y_2 \mathbf{a}_2 \ \cdots \ y_m \mathbf{a}_m]^\top \in \mathbb{R}^{m \times n}$, $\Psi_2 : \mathbb{R}^m \rightarrow \mathbb{R} : z := [z_1, z_2, \dots, z_m]^\top \mapsto \sigma_{[-1,0]^m}(z) := \sum_{i=1}^m \sup_{v \in [-1,0]} v z_i = \sum_{i=1}^m \max\{0, -z_i\}$, $\Psi_3 := \|\cdot\|_1$, $D_2 := \gamma_1^{-1/2}\mathbf{I}_m$, $D_3 := \gamma_2^{-1/2}\mathbf{I}_n$, and $Q = 2$. Then, the problem (4.38) reduces to

$$\min_{\mathbf{x} \in \mathbb{R}^n} \mu_1 \underbrace{(\sigma_{[-1,0]^m})}_{=:\Psi_2} \gamma_1^{-1/2} \mathbf{I}_m \underbrace{([y_1 \mathbf{a}_1 \ \cdots \ y_m \mathbf{a}_m]^\top \mathbf{x} - \mathbf{1}_m)}_{=:\mathcal{A}_2 \mathbf{x}} + \mu_2 \Phi_{\gamma_2}^{\text{MC}}(\mathbf{x}) + \frac{1}{2} \|\mathbf{x}\|_2^2, \quad (4.39)$$

which is equivalent to

$$\min_{\mathbf{x} \in \mathbb{R}^n} \mu_1 \sum_{i=1}^m \underbrace{(\sigma_{[-1,0]} \circ (y_i \mathbf{a}_i^\top \cdot -1))}_{= \text{hinge loss}} \gamma_1^{-1/2} \mathbf{I}_n(\mathbf{x}) + \mu_2 \Phi_{\gamma_2}^{\text{MC}}(\mathbf{x}) + \frac{1}{2} \|\mathbf{x}\|_2^2. \quad (4.40)$$

Here, $\sigma_{[-1,0]} \circ (y_i \mathbf{a}_i^\top \cdot -1)$ is the popular hinge loss, and thus each summand is Moreau enhanced hinge loss [35]. The formulation (4.40) can be seen as an extension of the work in [133], where the (Huberized) hinge loss, ℓ_1 norm, and squared ℓ_2 norm are used simultaneously.

The model given in (4.38) can be expressed as a special case of the LiMES model [35] shown below. Define the Hilbert space $\mathcal{Z} := \mathcal{Z}_2 \times \cdots \times \mathcal{Z}_{Q+1} = \{(z_2, \dots, z_{Q+1}) \mid z_i \in \mathcal{Z}_i, i = 2, \dots, Q+1\}$ equipped with the addition $\mathcal{Z} \times \mathcal{Z} \rightarrow \mathcal{Z} : (z, w) \mapsto (z_2 + w_2, \dots, z_{Q+1} + w_{Q+1})$, the scalar multiplication $\mathbb{R} \times \mathcal{Z} \rightarrow \mathcal{Z} : (a, z) \mapsto (az_2, \dots, az_{Q+1})$, and the inner product $\langle \cdot, \cdot \rangle_{\mathcal{Z}} : \mathcal{Z} \times \mathcal{Z} \rightarrow \mathbb{R} : (z, w) \mapsto \sum_{i=2}^{Q+1} \langle z_i, w_i \rangle_{\mathcal{Z}_i}$. Define the function and the operator

$$\Gamma_0(\mathcal{Z}) \ni \Psi := \bigoplus_{i=2}^{Q+1} \nu_i \Psi_i : z \mapsto \sum_{i=2}^{Q+1} \nu_i \Psi_i(z_i), \quad (4.41)$$

$$D : \mathcal{Z} \rightarrow \mathcal{Z} : z \mapsto (\nu_2^{1/2} D_2 z_2, \dots, \nu_{Q+1}^{1/2} D_{Q+1} z_{Q+1}). \quad (4.42)$$

Then, it follows from (4.41) and (4.42) that, for any $z \in \mathcal{Z}$,

$$\begin{aligned} \min_{w \in \mathcal{Z}} \left(\Psi(w) + \frac{1}{2} \|D(z - w)\|_{\mathcal{Z}}^2 \right) &= \min_{w \in \mathcal{Z}} \sum_{i=2}^{Q+1} \left(\nu_i \Psi_i(w_i) + \frac{1}{2} \|\nu_i^{1/2} D_i(z_i - w_i)\|_{\mathcal{Z}_i}^2 \right) \\ &= \sum_{i=2}^{Q+1} \nu_i \min_{w_i \in \mathcal{Z}_i} \left(\Psi_i(w_i) + \frac{1}{2} \|D_i(z_i - w_i)\|_{\mathcal{Z}_i}^2 \right), \end{aligned} \quad (4.43)$$

from which it holds that

$$\begin{aligned} \Psi_D(z) &:= \Psi(z) - \min_{w \in \mathcal{Z}} \left(\Psi(w) + \frac{1}{2} \|D(z - w)\|_{\mathcal{Z}}^2 \right) \\ &= \sum_{i=2}^{Q+1} \nu_i \left[\Psi_i(z_i) - \min_{w_i \in \mathcal{Z}_i} \left(\Psi_i(w_i) + \frac{1}{2} \|D_i(z_i - w_i)\|_{\mathcal{Z}_i}^2 \right) \right] \\ &= \sum_{i=2}^{Q+1} \nu_i (\Psi_i)_{D_i}(z_i). \end{aligned} \quad (4.44)$$

We define the operators

$$\mathcal{A} : \mathcal{X} \rightarrow \mathcal{Z} : x \mapsto (\mathcal{A}_2 x, \dots, \mathcal{A}_{Q+1} x), \quad (4.45)$$

$$M : \mathcal{X} \rightarrow \mathcal{Z} : x \mapsto (M_2 x, \dots, M_{Q+1} x). \quad (4.46)$$

Then, by (4.44), the cost function in Problem (4.38) can be regarded as the LiMES model (F.2.1) in Appendix F.2 by letting $\mathcal{A}_2 := \mathcal{A}$ and $\mathcal{L} := \text{Id}$ as follows:

$$J_{\Psi_D \circ \mathcal{A}}^{\mathcal{A}_1} : \mathcal{X} \rightarrow (-\infty, +\infty] : x \mapsto \frac{1}{2} \|\mathcal{A}_1 x\|_{\mathcal{Z}_1}^2 + \Psi_D(\mathcal{A} x). \quad (4.47)$$

On the other hand, since it holds that

$$\begin{aligned} \min_{w \in \mathcal{Z}} \left(\Psi(w) + \frac{1}{2} \|D(z - w)\|_{\mathcal{Z}}^2 \right) &= \min_{\tilde{w} \in \mathcal{Z}} \left(\Psi(D^{-1} \tilde{w}) + \frac{1}{2} \|Dz - \tilde{w}\|_{\mathcal{Z}}^2 \right) \\ &= {}^1(\Psi \circ D^{-1})(Dz), \end{aligned} \quad (4.48)$$

(4.47) can be represented as a sum of the smooth terms and a nonsmooth convex term as follows:

$$J_{\Psi_D \circ \mathcal{A}}^{\mathcal{A}_1} = \underbrace{\frac{1}{2} \|\mathcal{A}_1 \cdot\|_{\mathcal{Z}_1}^2 - {}^1(\Psi \circ D^{-1}) \circ D \mathcal{A}}_{=: F \text{ (smooth)}} + \underbrace{\Psi \circ \mathcal{A}}_{\text{(nonsmooth)}}. \quad (4.49)$$

Our results on the convexity of the smooth part F are given below.

Proposition 4.6. The following statements hold.

$$(a) F = \frac{1}{2} \|\mathcal{A}_1 \cdot\|_{\mathcal{Z}_1}^2 - \sum_{i=2}^{Q+1} 1(\nu_i \Psi_i \circ (\nu_i^{-1/2} D_i^{-1})) \circ (\nu_i^{1/2} D_i \mathcal{A}_i) \in \Gamma_0(\mathcal{X}) \text{ if}$$

$$(\clubsuit) M_1^* M_1 - \sum_{i=2}^{Q+1} \nu_i M_i^* D_i^2 M_i \succeq 0.$$

(b) Suppose that $\Psi_i := \|\cdot\|_{\mathcal{Z}_i}$ is a norm defined on \mathcal{Z}_i , and $K_i := \{x \in \mathcal{X} \mid \|D_i^2 \mathcal{A}_i x\|_{\mathcal{Z}_i, *}, \leq 1\} \neq \emptyset$ for $i = 2, \dots, Q+1$, where $\|\cdot\|_{\mathcal{Z}_i, *}$ is the dual norm of $\|\cdot\|_{\mathcal{Z}_i}$. Assume that $\text{int}\left(\bigcap_{i=2}^{Q+1} K_i\right) \neq \emptyset$. Then, $F \in \Gamma_0(\mathcal{X})$ if and only if (\clubsuit) is satisfied.

Proof. See Appendix J.7. □

We remark that the convexity of the smooth part F immediately implies the overall convexity of the entire cost function since the nonsmooth term $\Psi_D \circ \mathcal{A}$ is always convex. Proposition 4.6 reproduces the results of Proposition 3.1 by letting $Q := 2$, $\mathbf{X} \in \mathcal{X} := \mathcal{Z}_1 := \mathcal{Z}_3 := \mathbb{R}^{n \times d}$, $\mathcal{Z}_2 := \mathcal{Y} := \mathbb{R}^{n \times m}$, $\mu_2 := \lambda_2^{-1}$, $\mu_3 := \lambda_1 \lambda_2^{-1}$, $\mathcal{A}_1 := \mathcal{A}_3 := \text{Id}$, $\mathcal{A}_2 : \mathbf{X} \mapsto \mathbf{B} - \mathbf{X}\mathbf{A}$ with given matrices $\mathbf{A} \in \mathbb{R}^{d \times m}$ and $\mathbf{B} \in \mathcal{Y}$, $\mathcal{D}_1 := L$, $\mathcal{D}_2 := M$, and Ψ_2, Ψ_3 be the $\ell_{2,1}$ norms. Note that K in Proposition 3.1 is equivalent to $\{\mathbf{X} \in \mathcal{X} \mid \|\mathbf{L}(\mathbf{Y} - \mathbf{X}\mathbf{A})\|_{2,\infty} \leq 1, \|\mathbf{M}\mathbf{X}\|_{2,\infty} \leq 1\}$.

4.5 Numerical Examples

First, the performance of the proposed method in the robust signal recovery task is evaluated using toy data and compared with the methods based on different cost functions in Experiment 4-A. Then, we show the impact of hyperparameters on the performance of S-SORR in Experiment 4-B. Finally, we show the performance of the proposed method in application to speech denoising in Experiment 4-C. The hyperparameter α is set to 1 in Experiments 4-A and 4-C to focus on robustness, and the case of $\alpha \in (0, 1)$ is discussed in Experiment 4-B.

The measurement vector \mathbf{y} is constructed as the model in (4.1). Here, $\mathbf{A} \in \mathbb{R}^{m \times n}$ is the i.i.d. standard Gaussian input matrix, $\mathbf{x}_\diamond \in \mathbb{R}^n$ is a sparse vector with κ_x nonzero components generated from i.i.d. $\mathcal{N}(0, 1)$ for $\kappa_x \in \mathbb{N}^*$, and each component of $\boldsymbol{\varepsilon}_\star \in \mathbb{R}^m$ is generated from i.i.d. $\mathcal{N}(0, \sigma_{\boldsymbol{\varepsilon}_\star}^2)$ for a given SNR. The outlier $\mathbf{o}_\diamond \in \mathbb{R}^m$ is a sparse vector with κ_o nonzero components for $\kappa_o \in \mathbb{N}^*$, and we consider the following four models for nonzero components:

model O₁: $-M_o$ or M_o with the same probability for a given $M_o > 0$,

model O₂: M_o follows a uniform distribution $\mathcal{U}(0.5\bar{M}_o, 1.5\bar{M}_o)$ in model O₁ for a given $\bar{M}_o > 0$,

Table 4.2: Formulations for sparse signal recovery.

method	problem formulation
lasso [21]	$\min_{\mathbf{x} \in \mathbb{R}^n} \frac{1}{2} \ \mathbf{y} - \mathbf{Ax}\ _2^2 + \mu \ \mathbf{x}\ _1$
GMC [25]	$\min_{\mathbf{x} \in \mathbb{R}^n} \frac{1}{2} \ \mathbf{y} - \mathbf{Ax}\ _2^2 + \mu \Phi_{\sqrt{\gamma/\mu\mathbf{A}}}^{\text{GMC}}(\mathbf{x})$

Table 4.3: Formulations for sparse outlier-robust signal regression.

method	problem formulation
YALL1 [80]	$\min_{\mathbf{x} \in \mathbb{R}^n} \ \mathbf{y} - \mathbf{Ax}\ _1 + \mu \ \mathbf{x}\ _1$
extended lasso [82]	$\min_{\mathbf{x} \in \mathbb{R}^n, \mathbf{o} \in \mathbb{R}^m} \frac{1}{2m} \ \mathbf{y} - \mathbf{Ax} - \sqrt{n}\mathbf{o}\ _2^2 + \mu_{m,x} \ \mathbf{x}\ _1 + \mu_{m,o} \ \mathbf{o}\ _1$
Huber FISTA [81]	$\min_{\mathbf{x} \in \mathbb{R}^n} \sum_{i=1}^m \Phi_{\gamma}^{\text{Huber}}(y_i - (\mathbf{Ax})_i) + \mu \ \mathbf{x}\ _1$
RPGG [72]	$\min_{\mathbf{x} \in \mathbb{R}^n, \mathbf{o} \in \mathbb{R}^m} \Phi_{\gamma_1}^{\text{MC}}(\mathbf{o}) + \mu \Phi_{\gamma_2}^{\text{MC}}(\mathbf{x})$ s.t. $\mathbf{o} = \mathbf{y} - \mathbf{Ax}$
S-SORR	See (4.3)

model O₃: \bar{M}_o follows $\mathcal{N}(\mu_{\bar{M}_o}, \sigma_{\bar{M}_o}^2)$ with $\sigma_{\bar{M}_o}^2 := (\bar{M}_o/4)^2$ in model O₂ for a given $\mu_{\bar{M}_o} > 0$,

model O₄: i.i.d. $\mathcal{N}(0, \sigma_{o_\diamond}^2)$ for a given $\sigma_{o_\diamond}^2 > 0$.

Experiment 4-A uses O₁-O₄, Experiment 4-B uses models O₁-O₃, and Experiment 4-C uses model O₄, respectively. The indices of nonzero components in \mathbf{x}_\diamond and \mathbf{o}_\diamond are chosen randomly³ [72, 82].

4.5.1 Experiment 4-A: Toy data

The performance of the proposed method is compared with the existing sparse outlier-robust recovery methods: YALL1 [80], Huber FISTA [81], extended lasso [82], and RPGG [72]. We also test sparse signal recovery methods: lasso [21] and GMC [25] for reference. The formulations for these methods are listed in Tables 4.2 and 4.3. The regularization parameter μ of each method, γ of GMC and Huber FISTA, $\mu_{m,x}$ and $\mu_{m,o}$ of extended lasso, and γ_1, γ_2 of RPGG are tuned to attain the best performance. For

³Other popular models for outliers include *contaminated Gaussian distribution* and *heavy-tailed distributions*. In *contaminated Gaussian distribution*, the residuals obeys $(1 - \epsilon)D_1 + \epsilon D_2$, where ϵ is the occurrence probability of outliers, $D_1 := \mathcal{N}(0, \sigma^2)$, and D_2 may be an arbitrary distribution [65]. In *heavy-tailed distributions*, the density tail of the residuals tends to zero more slowly than the Gaussian distribution, such as the symmetric alpha-stable (S α S) distribution [65, 73]. It is beyond the scope of the present study to consider those models.

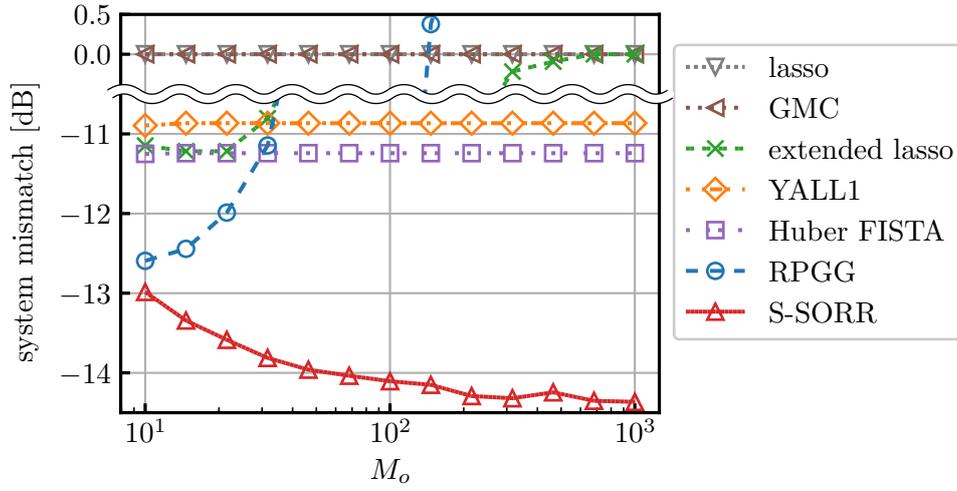
S-SORR, ρ , γ_1 , γ_2 , and μ_1 are tuned to attain the best performance under convexity condition (K-I). The parameter μ_2 of S-SORR is set to the theoretical bound to satisfy convexity condition (K-II), and τ and λ are set to slightly smaller values than the theoretical upper bound for the convergence condition. The step size of RPKG is tuned so that the convergence speed is comparable to the other methods. RPKG is the only nonconvex formulation, and the initial point of its algorithm is set to the zero vector (in analogy to the original chapter). The results are averaged over 300 independent trials, and the evaluation metric is the system mismatch defined as

$$\text{system mismatch} := \frac{\|\hat{\mathbf{x}} - \mathbf{x}_\diamond\|_2^2}{\|\mathbf{x}_\diamond\|_2^2}, \quad (4.50)$$

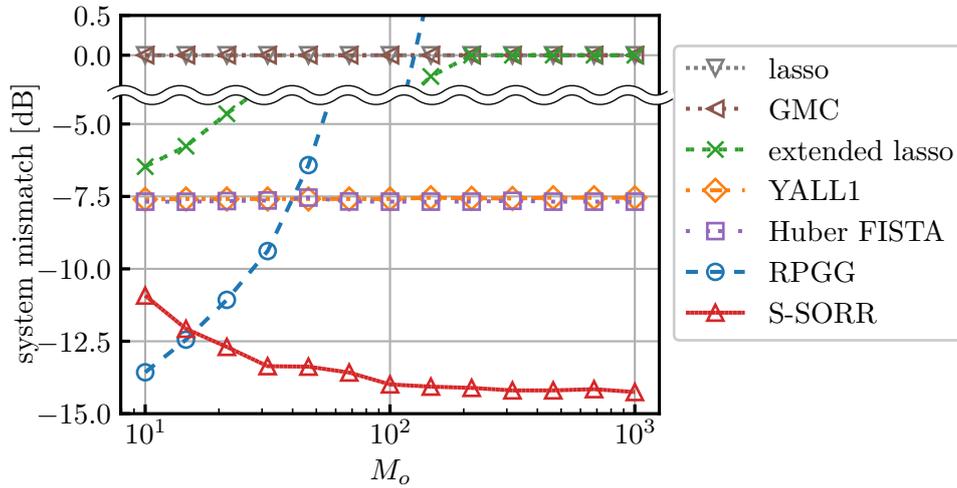
where $\hat{\mathbf{x}}$ is an estimate of \mathbf{x}_\diamond .

4.5.1.1 Performance with Different Outlier Magnitudes

First, we consider the outliers of model O_1 . Figure 4.1 shows the system mismatch across the outlier magnitude M_o for the overdetermined and underdetermined cases. In both cases, S-SORR achieves remarkably robust performance especially when M_o is fairly high. It should be remarked that S-SORR has a unique feature that the system mismatch decreases as M_o increases, unlike other methods. In general, a small γ_1 leads to a small μ_1 due to convexity condition (K-I), and this can cause excessive regularization. However, when M_o is large, γ_1 can be large as well, allowing μ_1 to be large enough. While extended lasso behaves similarly to Huber FISTA when M_o is small, the system mismatch of extended lasso increases sharply as M_o increases. This is because the term $\mu_{m,o}\|\mathbf{o}\|_1$ (see Table 4.3) introduces additional bias due probably to the limited robustness, and it seriously affects the quadratic loss $(1/(2m))\|\mathbf{y} - \mathbf{A}\mathbf{x} - \sqrt{n}\mathbf{o}\|_2^2$. While RPKG outperforms YALL1 and Huber FISTA when $M_o < 20$ in Figure 4.1(a) and $M_o < 50$ in Figure 4.1(b), its performance deteriorates as M_o becomes larger. We remark that the critical differences between RPKG and S-SORR are (i) stability against noise and (ii) convexity. Specifically, due to its nonconvexity, RPKG involves a tradeoff between computational complexity and recovery performance. According to [72, Theorem 3], under an appropriate condition for \mathbf{A} and with the number of iterations of at most $\mathcal{O}(r/\kappa)$, RPKG estimates the solution with reconstruction errors bounded by $\mathcal{O}(\kappa) + \mathcal{O}(\|\boldsymbol{\varepsilon}_\star\|_2)$, where $r := \sqrt{\|\hat{\mathbf{x}}_0 - \mathbf{x}_\diamond\|_2^2 + \|\hat{\mathbf{o}}_0 - \mathbf{o}_\diamond\|_2^2}$ with the initial point $(\hat{\mathbf{x}}_0, \hat{\mathbf{o}}_0)$ and κ is the step size. Since we tune the step size in the range where the convergence speed is comparable to the other methods (see Figure 4.2), the step size tends to be larger as M_o becomes larger, resulting in lower reconstruction errors. In contrast, S-SORR can be optimized by the efficient algorithm in the framework of convex analysis without such a tradeoff. Figure 4.2 shows the learning curve of each method. It can be seen that the convergence speed



(a)



(b)

Figure 4.1: System mismatch across M_o for $\kappa_x = [0.05n]$ and $\kappa_o = [0.3m]$. (a) Overdetermined case ($m = 128$, $n = 64$) under SNR 5 dB. (b) Underdetermined case ($m = 64$, $n = 128$) under SNR 15 dB.

of S-SORR is faster than the other methods. Figure 4.3 shows the system mismatch across \bar{M}_o for outliers of model O_2 and $\mu_{\bar{M}_o}$ for model O_3 . In both cases, S-SORR outperforms the other methods as consistent with the case of model O_1 .

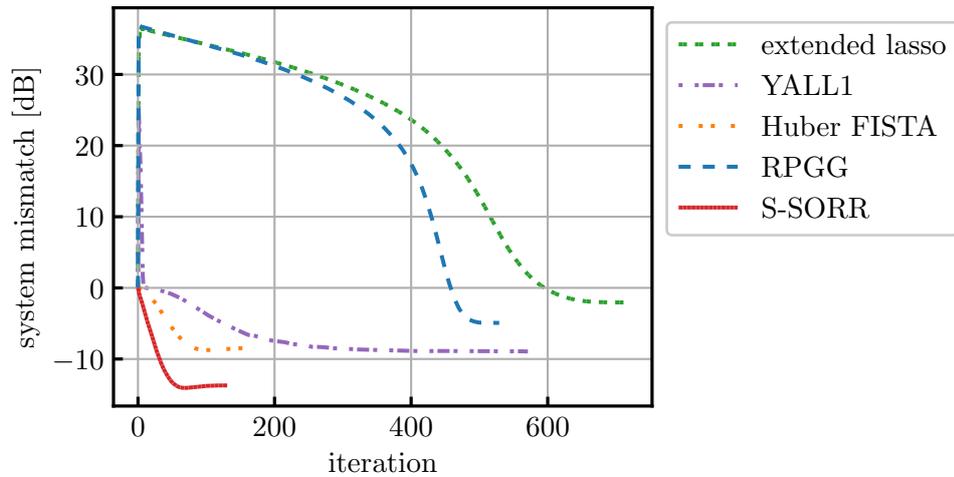
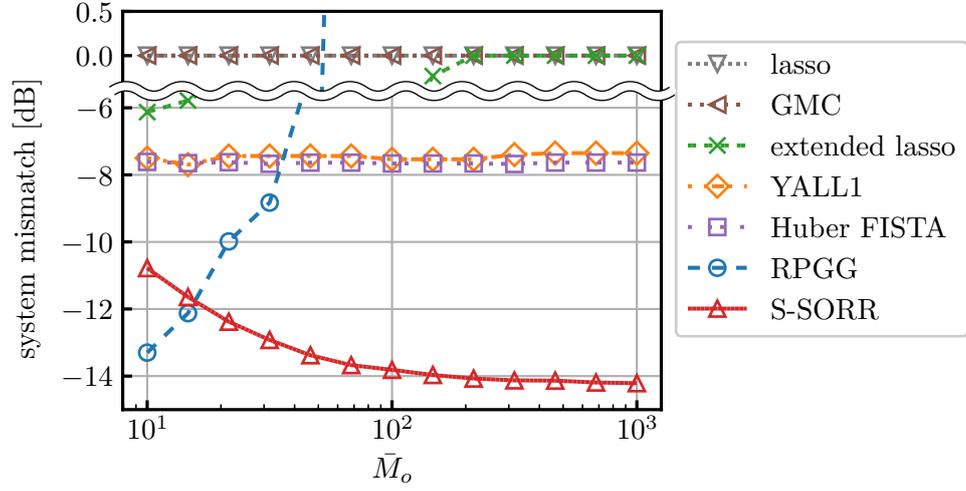


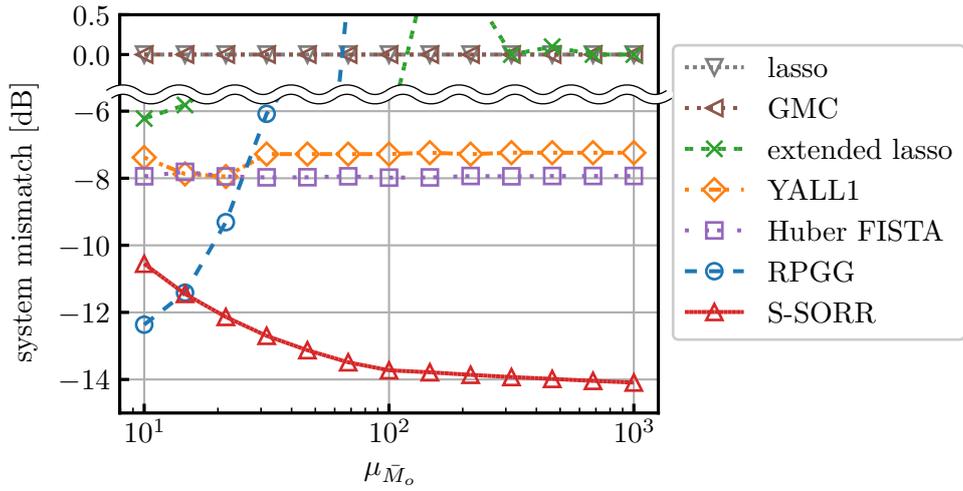
Figure 4.2: Learning curves for $m = 256$, $n = 512$, $\kappa_x = \lfloor 0.05n \rfloor$, $\kappa_o = \lfloor 0.3m \rfloor$, and $M_o = 100$ under SNR 15 dB.

4.5.1.2 Performance with Different Outlier Sparsity, m/n , and SNR

Figure 4.4 shows the system mismatch across outlier sparsity for both overdetermined and underdetermined cases for outliers of model O_1 . In both cases, S-SORR exhibits highly accurate and stable performance, and it outperforms the other methods significantly. Lasso and GMC are highly sensitive to outliers due to the use of the quadratic loss function. The performance of RPGG and extended lasso are worse than YALL1, Huber FISTA, and S-SORR over a wide range of outlier sparsity due to the large values of M_o . S-SORR outperforms YALL1 and Huber FISTA, and their differences increase as the outliers become denser. This is because the ℓ_1 and Huber's losses have limited robustness since these losses increase linearly as the residual increases. In contrast, the MC loss is highly robust since it stays constant for large values. Figures 4.5 and 4.6 show the system mismatch across m/n and SNR, respectively, for outliers of model O_1 . As consistent with the results in Fig. 4.4, S-SORR significantly outperforms the other methods.

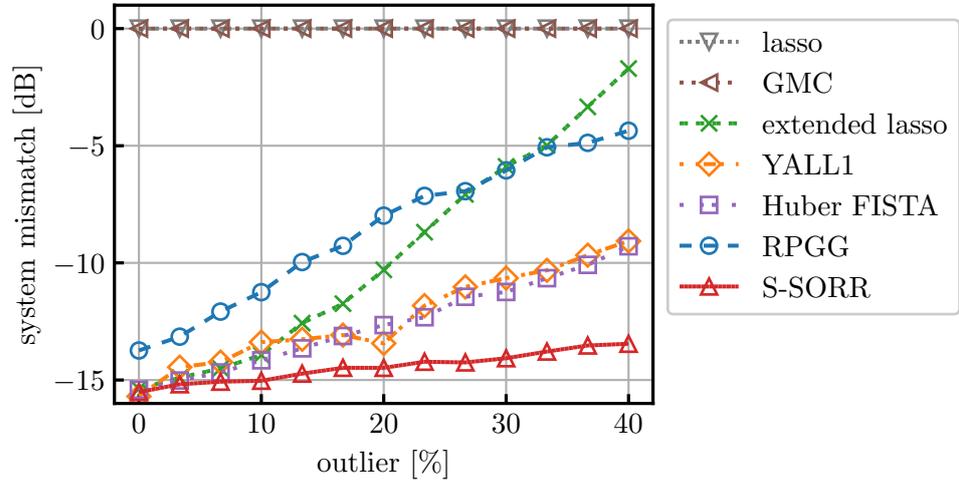


(a)

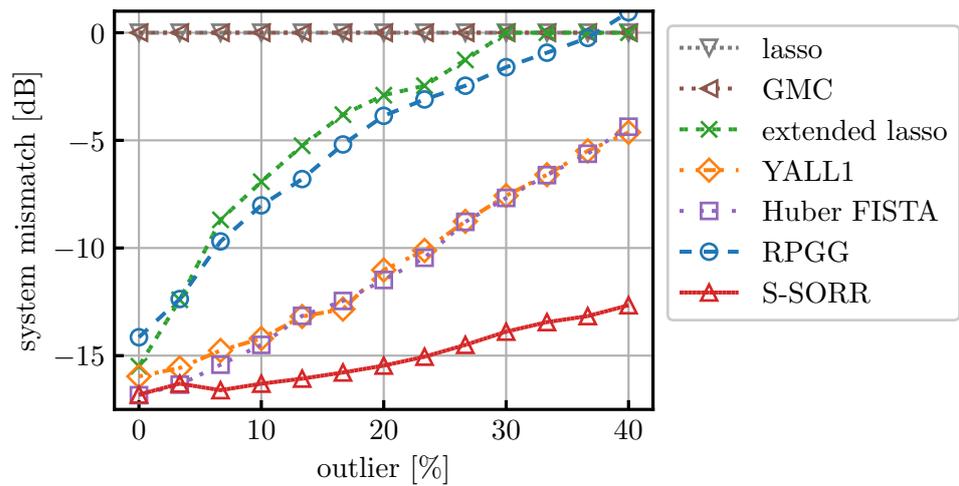


(b)

Figure 4.3: System mismatch across (a) \bar{M}_o for outliers of model O_2 and (b) $\mu\bar{M}_o$ for model O_3 , respectively, for $\kappa_x = \lfloor 0.05n \rfloor$, $\kappa_o = \lfloor 0.3m \rfloor$, $m = 128$, and $n = 64$ under SNR 15 dB.



(a)



(b)

Figure 4.4: System mismatch across outlier sparsity for $\kappa_x = \lfloor 0.05n \rfloor$ and $M_o = 100$. (a) Overdetermined case ($m = 128$, $n = 64$) under SNR 5 dB. (b) Underdetermined case ($m = 64$, $n = 128$) under SNR 15 dB.

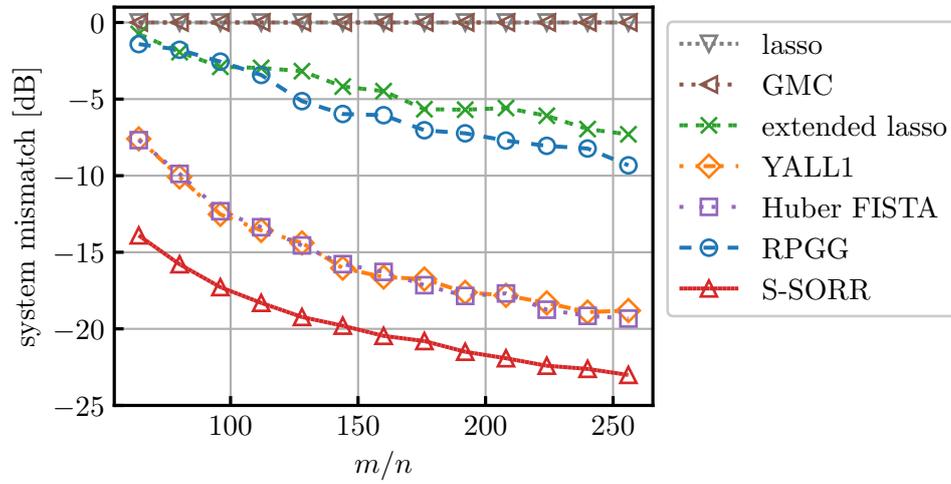


Figure 4.5: System mismatch across m/n for $n = 128$, $\kappa_x = [0.05n]$, $\kappa_o = [0.3m]$, and $M_o = 100$ under SNR 15 dB.

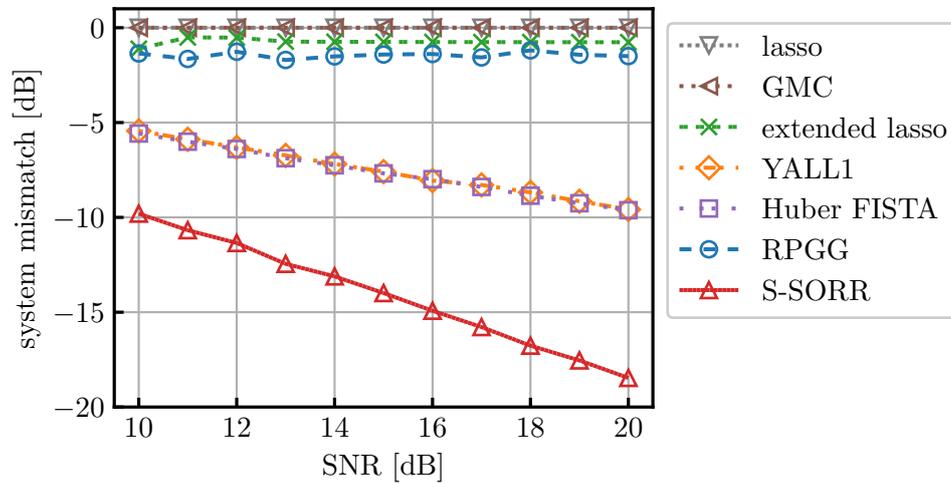


Figure 4.6: System mismatch across SNR for $m = 64$, $n = 128$, $\kappa_x = [0.05n]$, $\kappa_o = [0.3m]$, and $M_o = 100$.

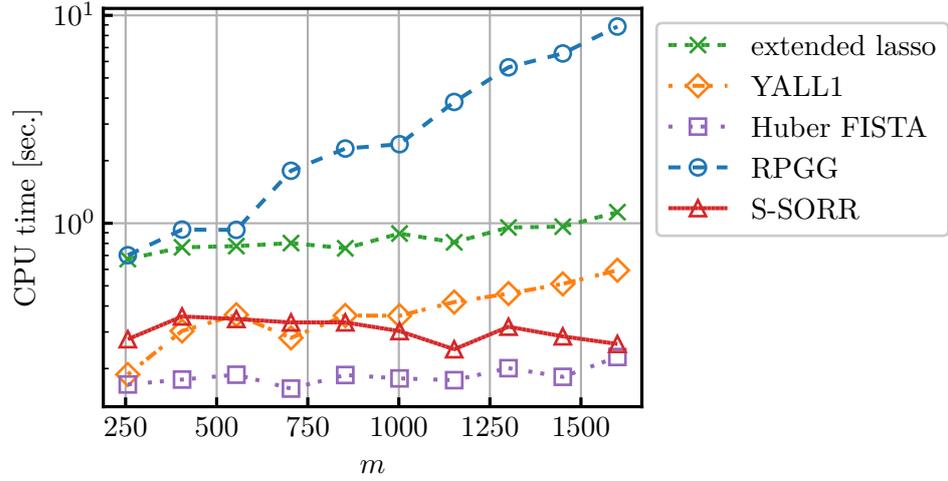


Figure 4.7: CPU time across m for $n = 512$, $\kappa_x = \lfloor 0.05n \rfloor$, $\kappa_o = \lfloor 0.3m \rfloor$, and $M_o = 100$ under SNR 15 dB for outliers of model O_1 .

4.5.1.3 Comparison of Execution Time

Figure 4.7 shows the execution time required for sparse outlier-robust recovery methods to converge. It can be seen that the CPU time of S-SORR is comparable to the other methods based on convex optimization (YALL1, Huber FISTA, and extended lasso), and outperforms RPGG significantly. This is because RPGG requires the computation of the pseudoinverse matrix at the initialization step, while S-SORR needs to compute only simple operators.

4.5.1.4 Grouping Effect in the Absence of Outliers

We generate the vector $\mathbf{z}_1 \in \mathbb{R}^m$ randomly from the i.i.d. uniform distribution $\mathcal{U}(0, 20)$ and its correlated vector $\mathbf{z}_2 \in \mathbb{R}^m$ randomly so as to satisfy

$$\frac{|\bar{\mathbf{z}}_1^\top \bar{\mathbf{z}}_2|}{\|\bar{\mathbf{z}}_1\|_2 \|\bar{\mathbf{z}}_2\|_2} = 0.7, \quad (4.51)$$

where $\bar{\mathbf{z}}_1$ and $\bar{\mathbf{z}}_2$ are centered vectors by subtracting the mean from \mathbf{z}_1 and \mathbf{z}_2 , respectively. Here, the procedure for generating \mathbf{z}_2 is as follows:

- (i) generate $\boldsymbol{\zeta} \in \mathbb{R}^n$ from the i.i.d. uniform distribution $\mathcal{U}(0, 20)$,
- (ii) yield $\bar{\boldsymbol{\zeta}}$ by centering $\boldsymbol{\zeta}$,
- (iii) set

$$\bar{\boldsymbol{\zeta}}^\perp := \bar{\boldsymbol{\zeta}} - \frac{\bar{\boldsymbol{\zeta}}^\top \bar{\mathbf{z}}_1}{\|\bar{\mathbf{z}}_1\|_2^2} \bar{\mathbf{z}}_1, \quad (4.52)$$

and

(iv) set

$$\mathbf{z}_2 := \bar{\boldsymbol{\zeta}}^\perp + \frac{\|\bar{\boldsymbol{\zeta}}^\perp\|_2}{\tan(\arccos(0.7)) \|\bar{\mathbf{z}}_1\|_2} \bar{\mathbf{z}}_1. \quad (4.53)$$

See Appendix J.8 for proof of (4.51).

The observation vector is generated by $\mathbf{y} = \mathbf{z}_1 + 0.5\mathbf{z}_2 + \boldsymbol{\epsilon}_*$, where $\boldsymbol{\epsilon}_*$ is a zero-mean Gaussian noise vector with SNR set to 15 dB. The input matrix $\mathbf{A} := [\mathbf{a}_1 \mathbf{a}_2 \dots \mathbf{a}_9] \in \mathbb{R}^{m \times n}$ ($m = 100$, $n = 9$) is generated by normalizing the following vectors:

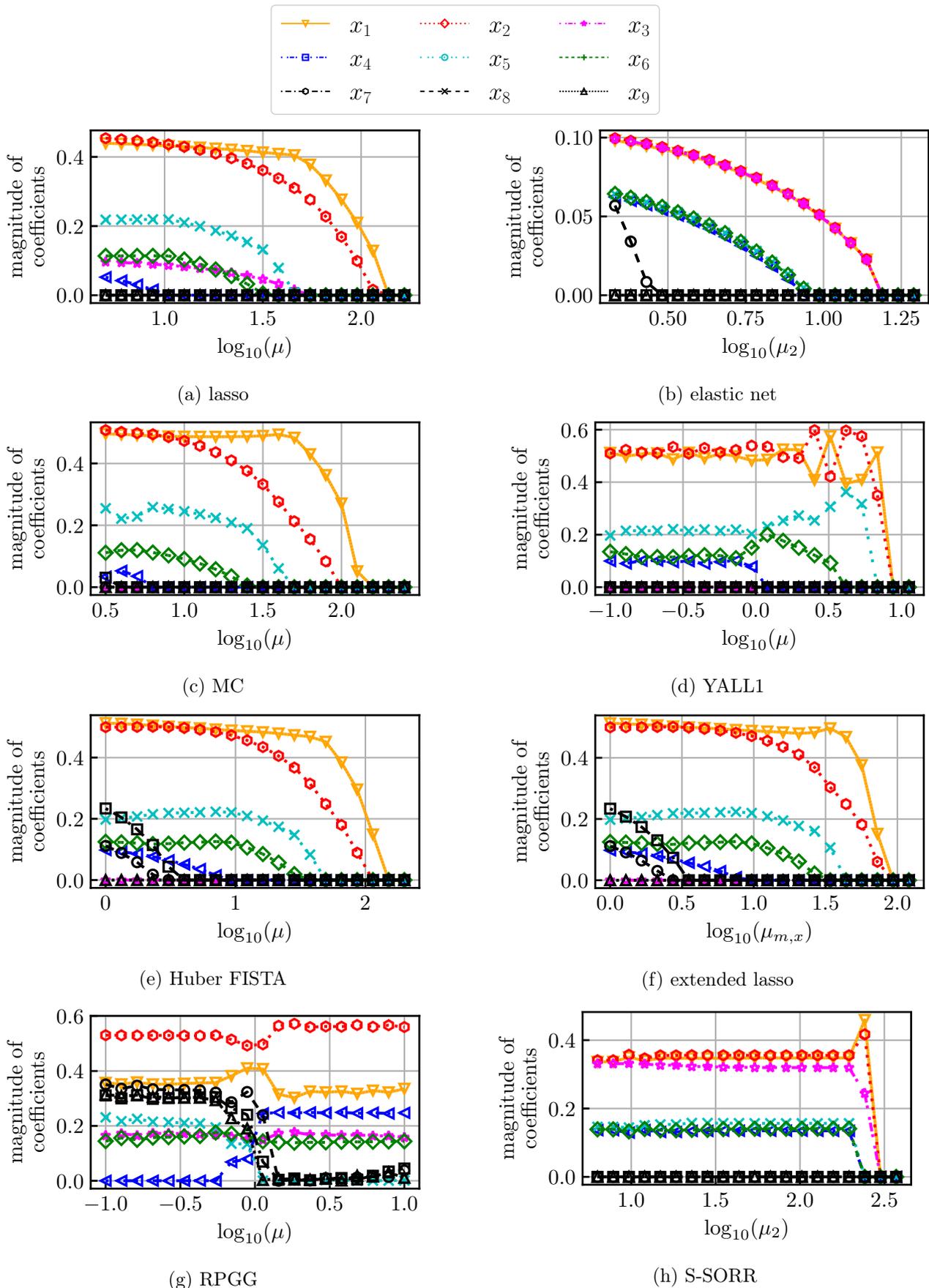
$$\begin{aligned} \tilde{\mathbf{a}}_1 &= \mathbf{z}_1 + \boldsymbol{\epsilon}_1, & \tilde{\mathbf{a}}_2 &= -\mathbf{z}_1 + \boldsymbol{\epsilon}_2, & \tilde{\mathbf{a}}_3 &= \mathbf{z}_1 + \boldsymbol{\epsilon}_3, \\ \tilde{\mathbf{a}}_4 &= \mathbf{z}_2 + \boldsymbol{\epsilon}_4, & \tilde{\mathbf{a}}_5 &= -\mathbf{z}_2 + \boldsymbol{\epsilon}_5, & \tilde{\mathbf{a}}_6 &= \mathbf{z}_2 + \boldsymbol{\epsilon}_6, \\ \tilde{\mathbf{a}}_7 &= \mathbf{q}_3 + \boldsymbol{\epsilon}_7, & \tilde{\mathbf{a}}_8 &= \mathbf{q}_4 + \boldsymbol{\epsilon}_8, & \tilde{\mathbf{a}}_9 &= \mathbf{q}_5 + \boldsymbol{\epsilon}_9. \end{aligned} \quad (4.54)$$

Here, $\boldsymbol{\epsilon}_i$ ($i = 1, 2, \dots, 9$) are generated from i.i.d. standard normal distribution, and $\mathbf{q}_3, \mathbf{q}_4, \mathbf{q}_5 \in \text{span}^\perp\{\mathbf{z}_1, \mathbf{z}_2\}$ are taken from the orthogonal matrix $\mathbf{Q} := [\mathbf{q}_1 \mathbf{q}_2 \dots \mathbf{q}_5]$ which is generated by orthonormalizing the columns of the matrix $[\mathbf{z}_1 \ \mathbf{z}_2 \ \mathbf{w}_1 \ \mathbf{w}_2 \ \mathbf{w}_3]$ by the QR decomposition, where $\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3 \in \mathbb{R}^m$ are generated from the i.i.d. standard normal distribution. The set of vectors $\{\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3\}$ are grouped, which is referred to as group A, because all of them are correlated with \mathbf{z}_1 . Similarly, $\{\mathbf{a}_4, \mathbf{a}_5, \mathbf{a}_6\}$ are referred to as group B. The correlations within each group and between the groups are approximately 1 and 0.7, respectively. Since the observation vector \mathbf{y} is generated with \mathbf{z}_1 and \mathbf{z}_2 , the vectors belonging to group A or group B are relevant to \mathbf{y} , while the other vectors $\mathbf{a}_7, \mathbf{a}_8$, and \mathbf{a}_9 are irrelevant. Hence, the group-A variables x_1, x_2 , and x_3 (and the group-B variables x_4, x_5 , and x_6) are desired to have nearly the same amplitudes, while x_7, x_8 , and x_9 are desired to be zero.

The parameters in each method and ρ , γ_1 , γ_2 , and μ_1 of S-SORR are tuned to minimize the residual errors $\|(\mathbf{z}_1 + 0.5\mathbf{z}_2) - \mathbf{A}\hat{\mathbf{x}}_*\|_2^2$, where $\hat{\mathbf{x}}_*$ is the estimate of the coefficient vector \mathbf{x}_\diamond . The parameter μ_2 of S-SORR is set to the theoretical bound to satisfy the convexity condition (see the appendix), τ and λ are set to a slightly smaller value than the theoretical upper bound for the convergence condition. The step size of the RPGG is tuned so that the convergence speed is comparable to the other methods. The initial point of RPGG is set to the zero vector.

Figure 4.8 shows the solution paths for each algorithm. To remove the influence of normalization, the i th component of $\hat{\mathbf{x}}_*$ is divided by $\|\tilde{\mathbf{a}}_i\|_2$ for $i = 1, 2, \dots, n$. The S-SORR estimator, as well as the elastic net estimator, succeeds in capturing the group structures; *i.e.*, it yields the group variables, the group-A variables $\{x_1, x_2, x_3\}$ and the group-B variables $\{x_4, x_5, x_6\}$, sharing nearly the same magnitude. More precisely, for S-SORR, the three variables in each group switch from inactive to active, as μ_2 decreases, simultaneously and quickly. None of the other methods succeeds. Specifically,

\mathbf{a}_3 in group A is missing in the results of MC, YALL1, Huber FISTA, and extended lasso, while the magnitude of x_3 is smaller than the other group-A variables x_1 and x_2 . RPGG fails to capture the three variables in each group simultaneously. We mention that the solution path of RPGG does not converge to the zero vector as μ increases due to the linear constraint $\mathbf{o} = \mathbf{y} - \mathbf{A}\mathbf{x}$.



4.5.1.5 Grouping Effect in the Presence of Outliers

We consider the situation when the data are contaminated by outliers as well as Gaussian noise. Specifically, the observation vector is generated by $\mathbf{y} = \mathbf{z}_1 + 0.5\mathbf{z}_2 + \boldsymbol{\varepsilon}_* + \mathbf{o}_\diamond$, for outliers of model O_4 . The vectors \mathbf{z}_1 and \mathbf{z}_2 and the matrix \mathbf{A} are generated in the same way as Section 4.5.1.4. The outlier sparsity $\text{supp}(\mathbf{o}_\diamond)/m$ is set to 0.05, where

$$\text{supp} : \mathbb{R}^m \rightarrow [0, +\infty) : \mathbf{x} \mapsto \text{card}(\{i \in \{1, 2, \dots, m\} \mid x_i \neq 0\}). \quad (4.55)$$

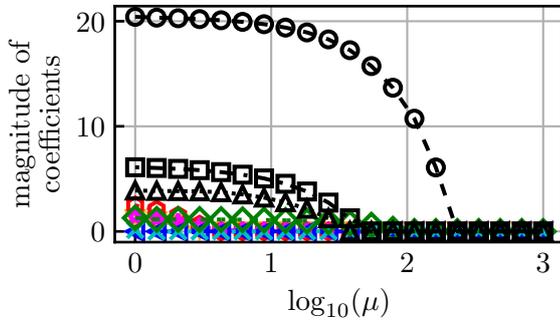
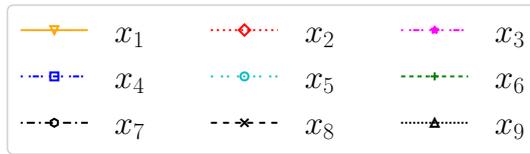
The SOR is defined as

$$\text{SOR} := \frac{\mathbb{E}[\|\mathbf{z}_1 + 0.5\mathbf{z}_2\|_2^2]/m}{\mathbb{E}[\|\mathbf{o}_\diamond\|_2^2]/\text{supp}(\mathbf{o}_\diamond)}, \quad (4.56)$$

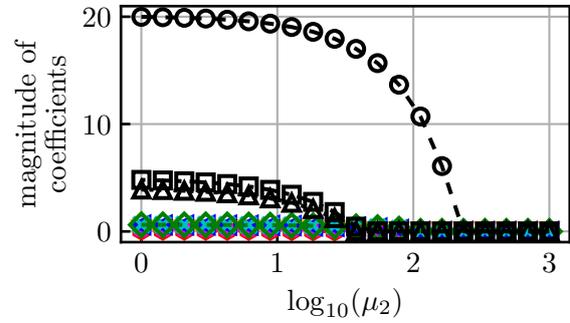
and is set to -30 dB. The parameters for each method are chosen in the same way as in Experiment 4-A.

Figure 4.9 shows the magnitude of coefficients for different values of the regularization parameter. S-SORR estimates the coefficients correctly, and it also maintains the grouping effect properly. This is consistent with the theoretical results of Proposition 4.5, which indicates that the upper bound of the discrepancy of the coefficients is independent of $\|\mathbf{y}\|_2$. All the methods excluding S-SORR fail in this outlier case. In particular, the elastic net unfortunately fails due to the presence of outliers. The outlier-robust methods (While YALL1, Huber FISTA, extended lasso, and RPGG) fail to capture the group structure as in Experiment 4-A.

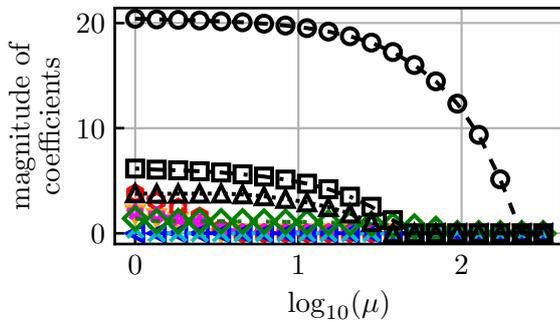
We finally mention that, while the naïve elastic net needs *rescaling* to remove the extra bias caused by the double shrinkage associated respectively with ℓ_1 and ℓ_2 norms, S-SORR requires no such a rescaling procedure. This is because the MC function equally penalizes those coefficients larger than the threshold γ and therefore it produces no extra bias. This is an additional practical advantage for S-SORR because it is free from the hyperparameter tuning for rescaling.



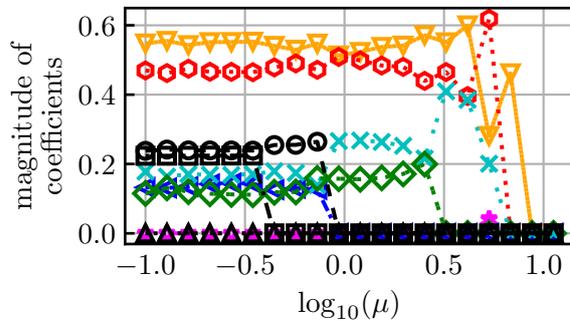
(a) lasso



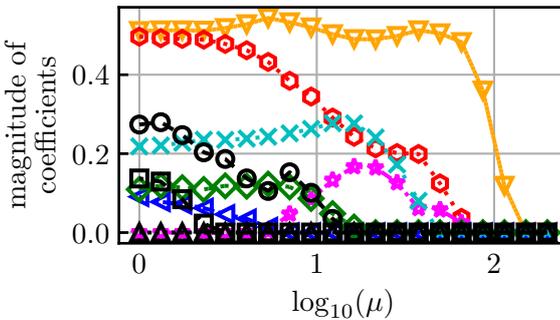
(b) elastic net



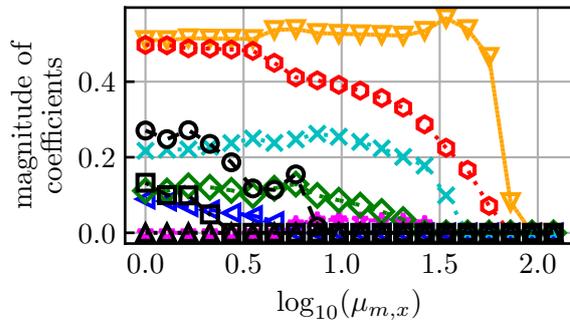
(c) MC



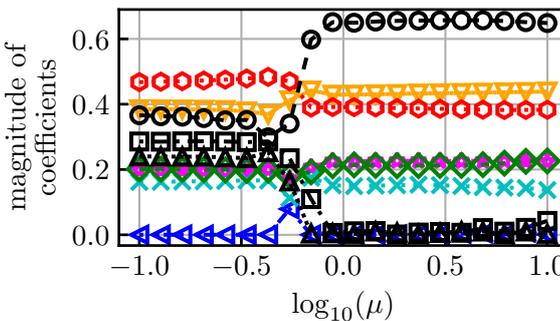
(d) YALL1



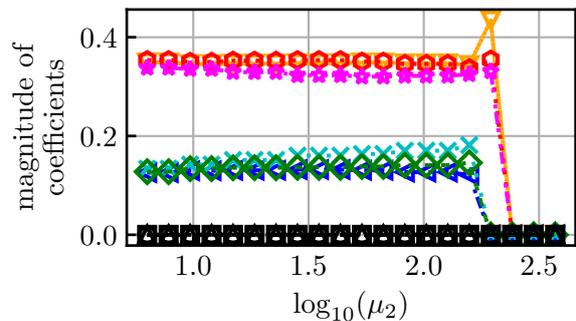
(e) Huber FISTA



(f) extended lasso



(g) RPGG



(h) S-SORR

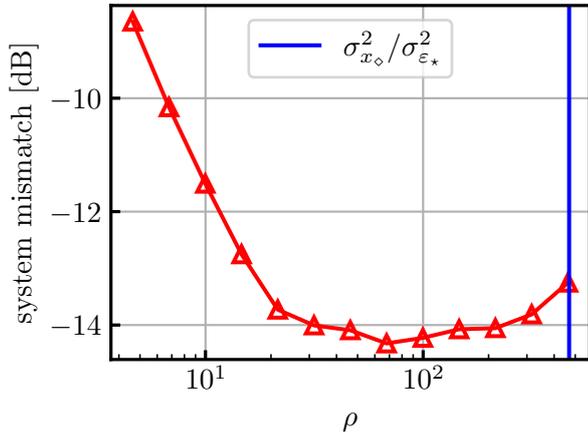
Figure 4.9: Solution paths for (a) lasso, (b) elastic net, (c) MC, (d) YALL1, (e) Huber FISTA, (f) extended lasso, (g) RPGG, and (h) S-SORR.

4.5.2 Experiment 4-B: Fluctuations of Hyperparameters

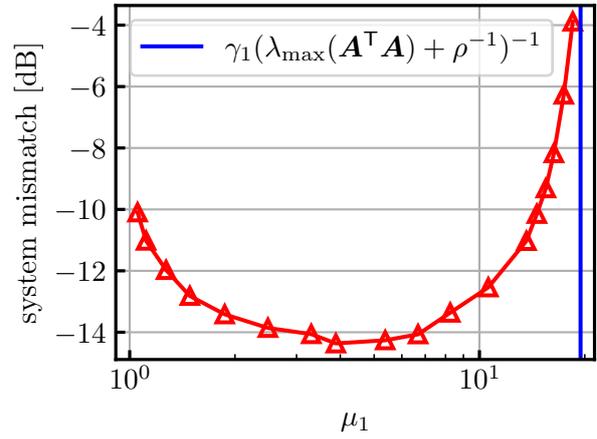
4.5.2.1 Hyperparameters ρ , μ_1 , γ_1 , and γ_2

We study how the performance of S-SORR changes due to fluctuations of the hyperparameters ρ , μ_1 , γ_1 , and γ_2 . Figure 4.10 shows the system mismatch across each hyperparameter. All the other hyperparameters than the targeted one are tuned to attain the best performance. Figure 4.10(a) shows the system mismatch across ρ . It can be seen that the fluctuation of the performance is smaller when ρ increases from the optimal value than when it decreases, and the performance is reasonably insensitive to the fluctuation of the ρ value. Figure 4.10(b) shows the system mismatch across μ_1 . The blue line shows the value of $\gamma_1(\lambda_{\max}(\mathbf{A}^\top \mathbf{A}) + \rho^{-1})^{-1}$, which is the upper bound of μ_1 to guarantee the convexity of the cost function. While the performance degrades as μ_1 approaches the bound, it is insensitive for a wide range ($1.2 < \mu_1 < 9$ in this case). Figure 4.10(c) shows the system mismatch across γ_1 . It can be seen that the best performance is achieved when $\gamma_1 = M_o$. When γ_1 is smaller, it leads to a smaller μ_1 due to convexity condition (K-I), and this may cause excessive regularization. When $\gamma_1 > M_o$, the derivative of the MC loss at the outlier value does not vanish, and this may cause extra bias (see Figure 4.11). Therefore, the optimal value of γ_1 depends on the balance between regularization and robustness. Figure 4.10(d) shows the system mismatch across γ_2 . The best performance is achieved when $\gamma_2 \geq \|\mathbf{x}_\diamond\|_\infty$ since the quadratic penalty can cause a shrinking bias when $\gamma_2 < \|\mathbf{x}_\diamond\|_\infty$. Therefore, the number of hyperparameters can be reduced in the case of $\alpha = 1$ by setting $\gamma_2 = +\infty$. We note that the hyperparameters ρ , μ_1 , and γ_2 can easily be set to obtain reasonable performance, while γ_1 should be tuned carefully. Figure 4.12 shows the values of the tuned hyperparameters across \bar{M}_o for outliers of model O_2 and across $\mu_{\bar{M}_o}$ for model O_3 , respectively. It can be seen that the performance is almost insensitive to μ_1 , μ_2 , γ_2 , and ρ under different outlier magnitudes. Moreover, $\gamma_1/\bar{M}_o \leq 1$ is satisfied for a large \bar{M}_o ($\gamma_1/\mu_{\bar{M}_o} \leq 1$ for a large $\mu_{\bar{M}_o}$) to achieve remarkable robustness, and $\gamma_1/\bar{M}_o \geq 1$ is satisfied for a small \bar{M}_o ($\gamma_1/\mu_{\bar{M}_o} \geq 1$ for a small $\mu_{\bar{M}_o}$) to avoid excessive regularization.

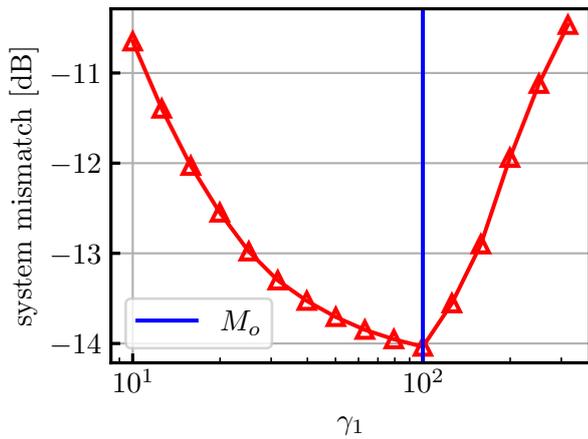
Remark 4.1. The best performance of S-SORR can be achieved when $\gamma_2 \geq \|\mathbf{x}_\diamond\|_\infty$ according to Figure 4.10(d), which violates assumption (b) of Proposition 4.5. However, the performance deteriorates only gradually when γ_2 becomes slightly smaller than $\|\mathbf{x}_\diamond\|_\infty$. Moreover, assumption (a) of Proposition 4.5 is weaker than that of Corollary 4.2 (which considers the case $\gamma_2 \rightarrow +\infty$). Therefore, Proposition 4.5 can be useful in applications in which the grouping effect is crucial.



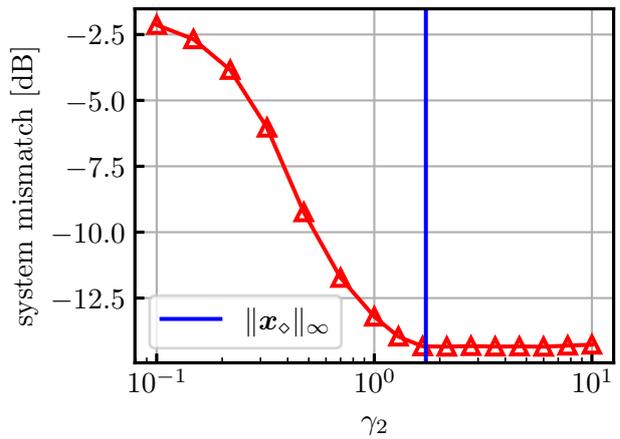
(a)



(b)

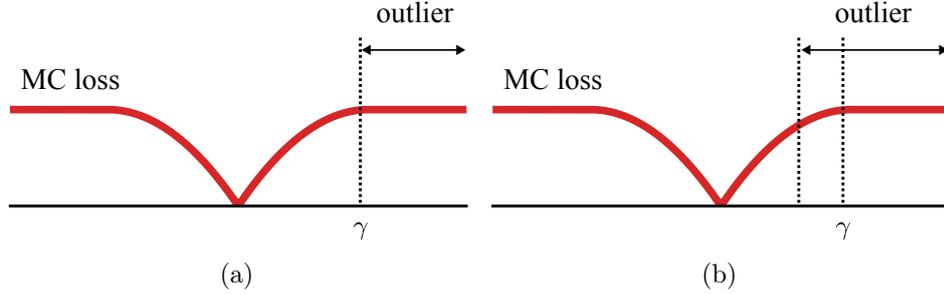


(c)



(d)

Figure 4.10: System mismatch across hyperparameters of S-SORR for $m = 64$, $n = 128$, $\kappa_x = \lceil 0.05n \rceil$, $\kappa_o = \lceil 0.3m \rceil$, and $\bar{M}_o = 100$ under SNR 15 dB.

Figure 4.11: Case with (a) $\gamma_1 = M_o$ and (b) $\gamma_1 > M_o$.

4.5.2.2 Hyperparameter α

In this subsection, we compare the performance of S-SORR ($\alpha < 1$) with the methods used in Experiment 4-A and the following formulation:

$$\min_{\mathbf{x} \in \mathbb{R}^n, \boldsymbol{\varepsilon} \in \mathbb{R}^m} \mu_1 \|\mathbf{y} - (\mathbf{A}\mathbf{x} + \boldsymbol{\varepsilon})\|_1 + \mu_2 \|\mathbf{x}\|_1 + \frac{1}{2} \|\boldsymbol{\varepsilon}\|_2^2, \quad (4.57)$$

which is denoted as stable YALL1. Figure 4.13 plots system mismatch and sparseness measure of each method across M_o for outliers of model O_1 . Here, the sparseness measure is defined as $\text{card}(\{i \mid |x_i| \geq 10^{-3}\})$, which returns the number of dominant components; the sparseness of \mathbf{x}_\diamond is 19. The hyperparameters of each method are tuned to attain the best system mismatch as well as Experiment A. It can be seen that, while S-SORR ($\alpha = 1$) outperforms the other methods due to high robustness, its sparseness is limited. While Stable YALL1 and Huber FISTA estimate sparser solutions than S-SORR, their robustness is limited. In contrast, S-SORR ($\alpha = 0.75$) yields sparser estimates than S-SORR ($\alpha = 1$) while maintaining the same level of system mismatch when the outlier magnitude M_o increases. S-SORR ($\alpha = 0.75$) outperforms the other methods except for S-SORR ($\alpha = 1$) when $M_o \geq 100$. This indicates that, even when the quadratic functions are introduced for convexity, the remarkable robustness and sparseness of the estimates can be balanced appropriately without introducing extra bias. We note that the sparseness of RPGG estimate is relatively high because the hyperparameters are tuned to attain the best system mismatch at a comparable convergence speed to the other methods.

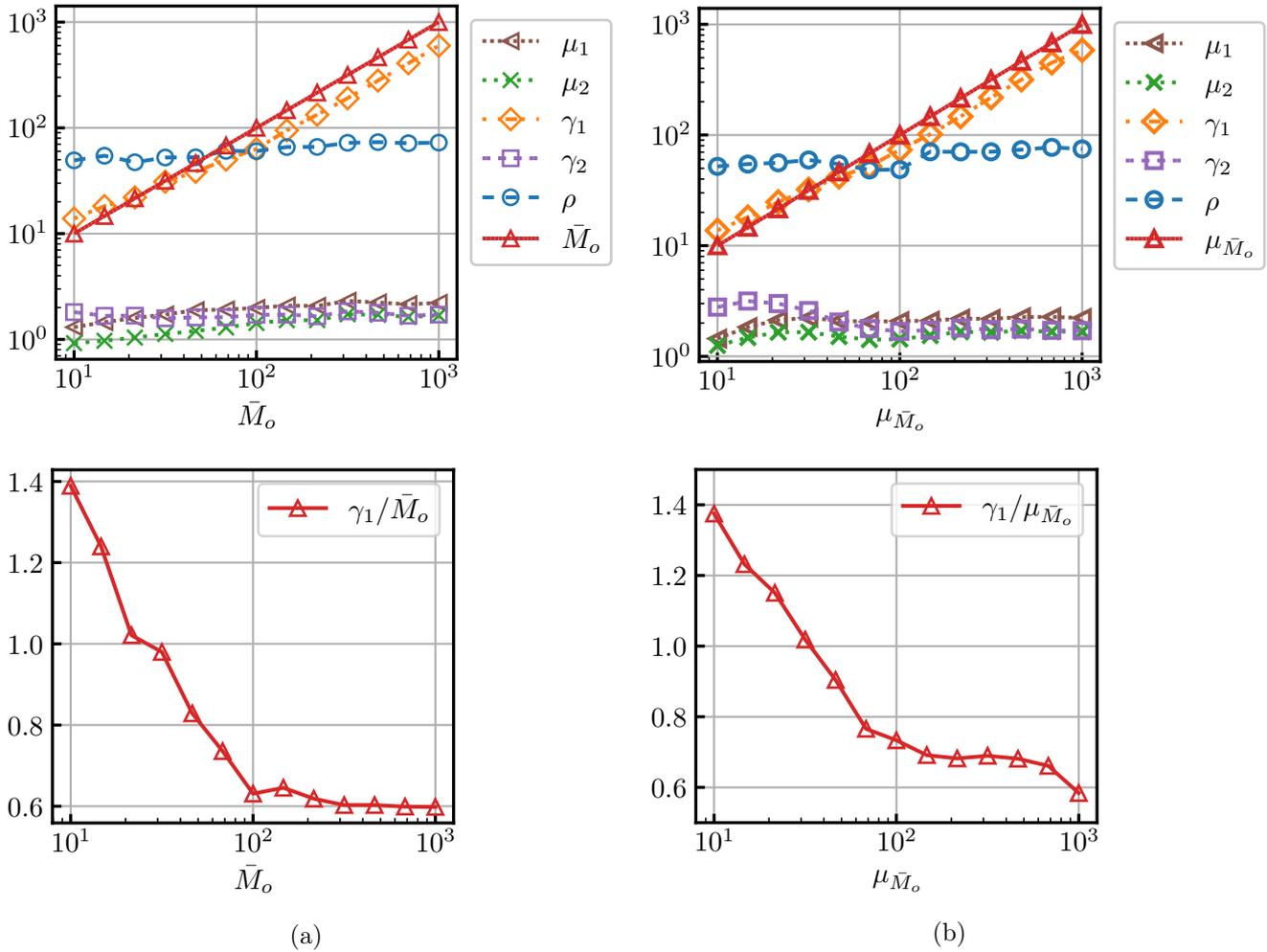
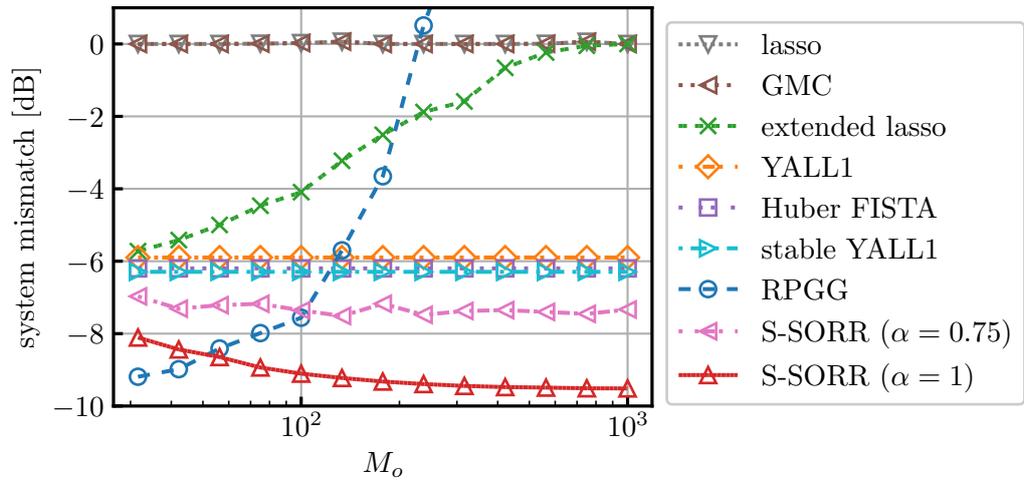
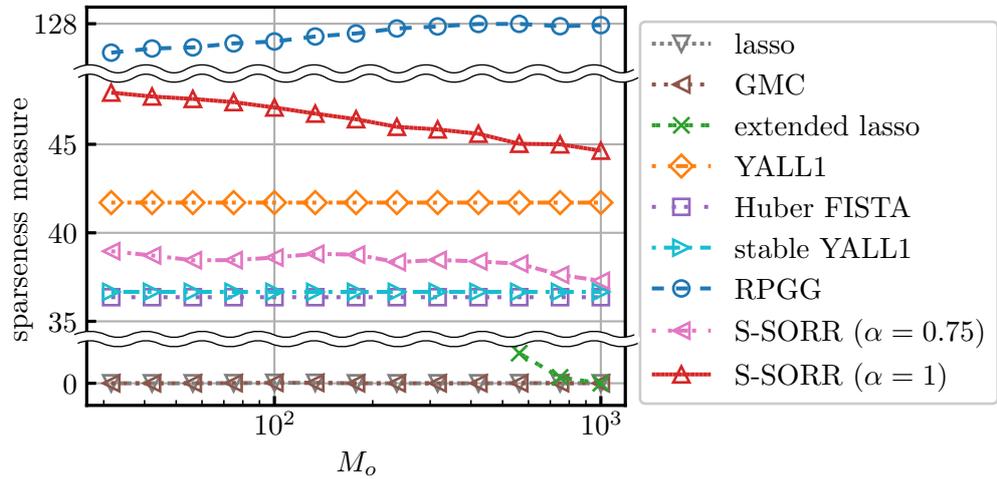


Figure 4.12: The values of tuned hyperparameters across (a) \bar{M}_o for outliers of model O_2 and (b) $\mu\bar{M}_o$ for model O_3 , respectively, for $m = 64$, $n = 128$, $\kappa_x = \lfloor 0.05n \rfloor$, $\kappa_o = \lfloor 0.3m \rfloor$, and $\bar{M}_o = 100$ under SNR 15 dB. The values of μ_2 , \bar{M}_o , and $\mu\bar{M}_o$ are plotted for reference.



(a)



(b)

Figure 4.13: (a) System mismatch and (b) sparseness measure across M_o for $m = 64$, $n = 128$, $\kappa_x = \lceil 0.15m \rceil$, and $\kappa_o = \lceil 0.1m \rceil$ under SNR 15 dB.

4.5.3 Experiment 4-C: Real Data (Application to Speech Denoising)

We consider a speech denoising task using the publicly available dataset from Interspeech 2020 Deep Noise Suppression Challenge [134]. The measured

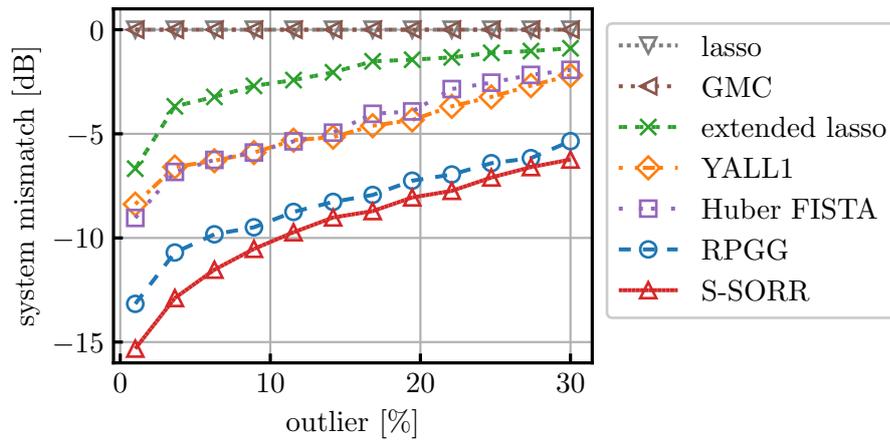
signal $\mathbf{y} \in \mathbb{R}^m$ is corrupted by noise and outliers as follows:

$$\mathbf{y} = \mathbf{z}_s + \boldsymbol{\varepsilon}_* + \mathbf{o}_\diamond, \quad (4.58)$$

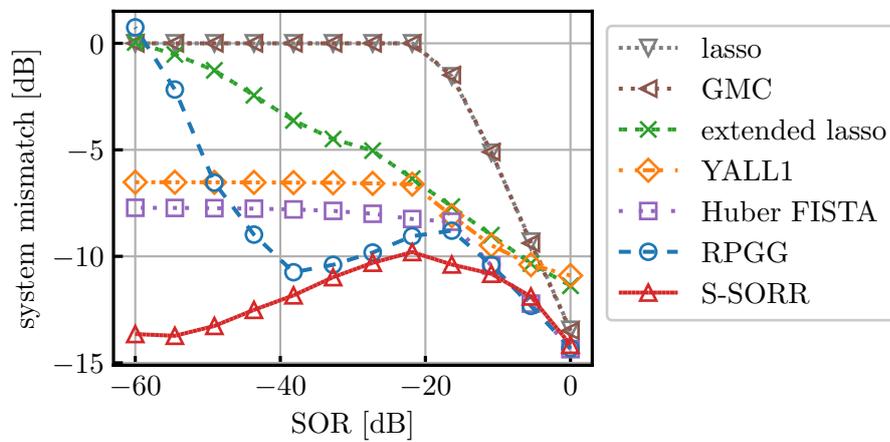
where $\mathbf{z}_s \in \mathbb{R}^m$, $\boldsymbol{\varepsilon}_* \in \mathbb{R}^m$, and $\mathbf{o}_\diamond \in \mathbb{R}^m$ are the clean speech signal, noise, and outliers, respectively, of size $m := 32768$. Each component of the noise vector $\boldsymbol{\varepsilon}_*$ follows i.i.d. $\mathcal{N}(0, \sigma_{\boldsymbol{\varepsilon}_*}^2)$, and \mathbf{o}_\diamond is sparse with nonzero components following i.i.d. $\mathcal{N}(0, \sigma_{\mathbf{o}_\diamond}^2)$ with a given SOR defined as $(\|\mathbf{z}_s\|_2^2/m)/(\|\mathbf{o}_\diamond\|_2^2/\kappa_o)$. The Haar wavelet decomposition is indicated by matrix $\mathbf{W} \in \mathbb{R}^{m \times m}$ with the decomposition level 5. The model in (4.58) can then be rewritten with the sparse wavelet coefficients $\mathbf{x}_\diamond := \mathbf{W}\mathbf{z}_s \in \mathbb{R}^m$ as

$$\mathbf{y} = \mathbf{W}^\top \mathbf{x}_\diamond + \boldsymbol{\varepsilon}_* + \mathbf{o}_\diamond, \quad (4.59)$$

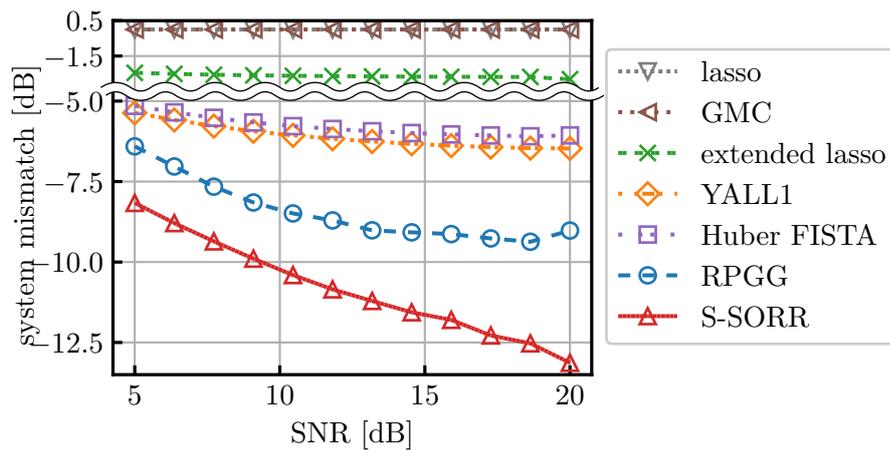
which corresponds to (4.1) for $\mathbf{A} := \mathbf{W}^\top$ ($n = m = 32768$). We consider the problem of recovering the clean signal \mathbf{z}_s from given \mathbf{y} and \mathbf{W} . The parameter settings are the same as in Section 4.5.1. Figure 4.14 depicts the system mismatch across (a) outlier sparsity, (b) SOR, and (c) SNR, respectively. In all cases, the proposed method outperforms the other methods significantly over a wide range of scenarios.



(a)



(b)



(c)

Figure 4.14: Application to speech denoising. (a) SNR 15 dB and SOR -30 dB, (b) SNR 15 dB and $\kappa_o = [0.05m]$, and (c) SOR -40 dB and $\kappa_o = [0.05m]$.

4.6 Conclusion

In this chapter, the S-SORR formulation was proposed as an integration of the MC-based sparse signal recovery and SORR for the sparse outlier-robust recovery problem. This addresses (Q2) which was raised in Chapter 1.2. While the proposed approach enjoyed remarkable robustness, the sparseness of the estimates was controlled by taking into consideration a tradeoff between sparseness and robustness. The ℓ_2 penalty of the auxiliary noise vector yielded stability under heavy Gaussian noise. The necessary and sufficient conditions for convexity of the smooth part of the cost were derived under the nonempty-interior assumption via the product space formulation within the LiMES framework. It was shown that, under an appropriate parameter choice, the simultaneous use of the MC and ℓ_2 penalties of the coefficient vector led to the grouping effect, and remarkably, the upper bound is independent of the observation vector, unlike the original result for the elastic net. The numerical examples showed the efficacy of S-SORR in its application to speech denoising as well as toy problems.

Chapter 5

External Division of Two Proximity Operators: An Application to Feature Grouping

5.1 Introduction

The goal of this chapter is to present a mathematical framework to derive methods for sparse regression with low estimation bias by exploiting a new class of operators called *the external division operator*. We start by presenting the background of operator splitting algorithms briefly.

5.1.1 Operator Splitting Algorithms

A popular approach to solve (1.2) is to utilize the proximity operators in the operator splitting algorithms. For instance, lasso can be solved by iterative shrinkage-thresholding algorithm (ISTA) [135] based on the standard soft-shrinkage operator, which is the proximity operator of the ℓ_1 norm. The convergence analysis of ISTA has been generalized in [112] to the case with the proximity operator of a weakly convex penalty. The proximity operator of the MC penalty is known to be the firm-shrinkage operator, which is intensively used in various algorithms to obtain less biased estimates of sparse signals than the soft-shrinkage operator [112, 25, 136, 137, 37, 138]. In general, the proximity operators of a weakly convex function (such as the firm-shrinkage operator) can be expressed as the gradient of a certain smooth convex function. This class of operators is called *monotone Lipschitz gradient (MoL-Grad) denoiser* (see Appendix G), and this property plays a crucial role in the convergence analysis of the operator splitting algorithms [139]. Actually, MoL-Grad denoiser can be expressed as the proximity operator of

a weakly convex function under certain conditions [139] (a related result can also be found in [140])¹.

5.1.2 Contributions

In this chapter, by answering (Q3), which is raised in Section 1.2, we aim to develop a method achieving the selection of groups of features and the bias reduction effect simultaneously with a convergence guarantee. Although one may consider the use of the Moreau enhancement of the OSCAR regularizer (see Section 1.1.5 and (2.67)) would be suitable, no direct discrete measure corresponding to OSCAR is known, unlike the ℓ_1 norm. This implies that its bias reduction effect is unclear. We first show that the Moreau enhancement of the OSCAR regularizer bridges the OSCAR regularizer and a certain ideal discrete measure (see Proposition 5.1).

Following this, we consider a novel debiasing approach which goes beyond the Moreau enhancement. We start by reconsidering the firm-shrinkage operator (the proximity operator of the MC penalty) from a different perspective. It is shown that the firm-shrinkage operator can be expressed as an external division of two soft-shrinkage operators with different thresholds (see Proposition 5.2 below). By analogy with this, we propose a new operator, $T_{\lambda_1, \lambda_2, \omega, \eta}^{\text{DOSCAR}}$, by considering an external division of two proximity operators of the scaled OSCAR regularizer, and refer it to as debiased OSCAR (DOSCAR) (see Section 5.2.3). A geometric interpretation of DOSCAR is provided to explain its debiasing effect. The motivation of this study is summarized in Figure 5.1. Primitive questions here are (i) *Can DOSCAR be expressed as a proximity operator of a certain function?*, and (ii) *how does the OSCAR regularizer relate to that function?* We repeat that the property of being expressed as the proximity operator of a certain weakly convex function plays a crucial role in the convergence analysis of the operator splitting algorithms.

To answer the above questions, we study a general class of nonlinear operators given in the following form:

$$\Delta_\omega := \omega \text{Prox}_{g_1} - (\omega - 1) \text{Prox}_{g_2}, \quad (5.1)$$

which we refer to as *the external division operator*. Here, $\omega > 1$, and $\text{Prox}_{g_1}, \text{Prox}_{g_2} : \mathbb{R}^n \rightarrow \mathbb{R}^n$ are the proximity operators of proper, lower-semicontinuous, and convex functions $g_1, g_2 : \mathbb{R}^n \rightarrow (-\infty, +\infty]$, respectively (see Section 2.1.2). DOSCAR is a specific example of (5.1) by setting g_1 and g_2 to the scaled OSCAR regularizers. See Section 5.1.3 for the relation of Δ_ω to previous works related to the operators composed of multiple proximity operators.

¹Note that the proximity operator is defined as the set-valued operator in [140], while it is defined as the single-valued operator (see (2.20)) in [139].

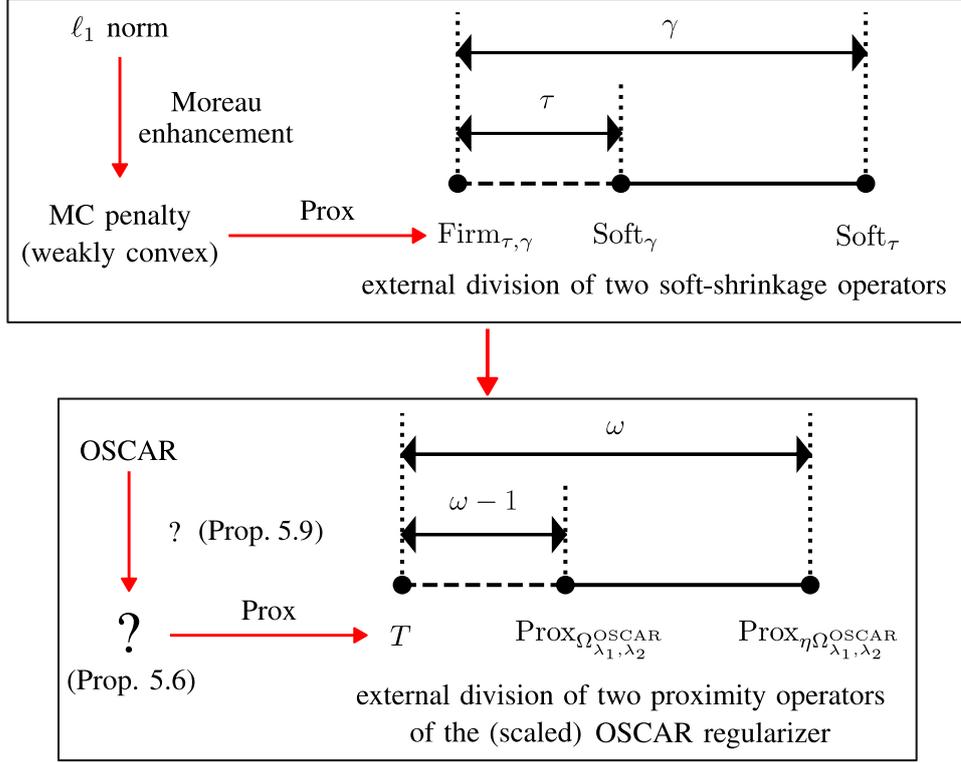


Figure 5.1: Motivation of this study.

We show that the external division operator is Lipschitz continuous and can be expressed as the gradient of the external division (with the same ratio as in (5.1)) of the Moreau envelopes of the conjugate functions g_1^* and g_2^* , say ψ_ω . If ψ_ω is convex, the external division operator is a MoL-Grad denoiser, and thus it is the proximity operator of a certain weakly convex function, say φ_ω (see Proposition 5.6). In this case, thanks to the results of [139], it is ensured that the gradient algorithm involving Δ_ω converges to a global minimizer of the cost function involving φ_ω and a given strongly convex fidelity function $f : \mathbb{R}^n \rightarrow \mathbb{R}$. Moreover, we generalize this algorithm to guarantee convergence even when the strong convexity of f is restricted to some subspace (see Proposition 5.7). Specifically, this generalized algorithm is applicable to solve (1.2) for the underdetermined case. The above algorithm can be applied for any g_1, g_2 if the corresponding ψ_ω is convex. We investigate the sufficient conditions for ψ_ω to be convex for given g_1, g_2 , and it turns out that the decomposition of proximity operator, *i.e.*, $\text{Prox}_{g_2} = \text{Prox}_{g_2 - g_1} \circ \text{Prox}_{g_1}$ is a nontrivially crucial assumption (see Proposition 5.8). This assumption holds for some widely-used sparsity-promoting convex functions such as the ℓ_1 norm, (a generalization of) OSCAR, and the total variation norm (see Example 5.1).

In fact, owing to the above decomposition, a closed-form expression of φ_ω is provided when g_1 and g_2 are the support functions (see Proposition 5.9). This closed-form expression allows us to show that φ_ω can be viewed as a generalization of the Moreau enhancement when g_1 and g_2 are scaled OSCAR (see Proposition 5.10). In numerical examples, DOSCAR is applied to the sparse signal regression with highly correlated features in both overdetermined and underdetermined cases. We show that DOSCAR achieves less biased estimation of sparse signals than the other methods, and yields significant improvements from OSCAR.

5.1.3 Relation to Previous Works

There are some previous works related to the operators composed of multiple proximity operators. Let $g_k : \mathbb{R}^n \rightarrow (-\infty, +\infty]$ be proper, lower-semicontinuous, and convex functions, and $\omega_k \geq 0$ such as $\sum_{k=1}^p \omega_k = 1$ for $k = 1, 2, \dots, p$ for $p \in \mathbb{N}^*$. The function $g : \mathbb{R}^n \rightarrow (-\infty, +\infty]$ which satisfies

$$\text{Prox}_g = \sum_{k=1}^p \omega_k \text{Prox}_{g_k} \quad (5.2)$$

is known as the proximal average [141, 142]. A generalization of (5.2) involving linear operators has been studied as the proximal mixture [142] (see also [143]). The case when $p = 2$ and $\omega_1 = \omega_2 = 1$ is studied as the proximal sum [143, 144]. However, these operators cover only a convex or conical combination of multiple proximity operators. To the best of our knowledge, the properties of the external division, defined as the affine combination with a negative weight of two proximity operators, remain unexplored.

5.2 Approaches to Bias Reduction of OSCAR

First, we show that the Moreau enhancement of the OSCAR regularizer bridges the OSCAR regularizer and a certain ideal nonconvex function. Then, we explore the way of developing a new approach which goes beyond the Moreau enhancement. We observe that the firm-shrinkage operator can be viewed as an external division of two soft-shrinkage operators. By analogy with that, we propose a new operator for bias reduction of OSCAR.

5.2.1 Moreau Enhancement of the OSCAR Regularizer

To reduce the estimation bias of OSCAR, a possible approach would be the use of the Moreau enhancement of the OSCAR regularizer. However, unlike the ℓ_1 norm, no direct discrete measure corresponding to OSCAR is known to the best of our knowledge. This implies that the bias reduction effect of

the Moreau enhancement of OSCAR is unclear. We show below that the Moreau enhancement of the OSCAR regularizer

$$(\Omega_{\lambda_1, \lambda_2}^{\text{OSCAR}})_{\gamma^{-1/2} \mathbf{I}_n} := \Omega_{\lambda_1, \lambda_2}^{\text{OSCAR}} - \gamma \Omega_{\lambda_1, \lambda_2} \quad (5.3)$$

bridges the $\Omega_{\lambda_1, \lambda_2}^{\text{OSCAR}}$ and a certain ideal discrete measure. Note that it holds by Fact 2.4(e) that

$$\lim_{\gamma \rightarrow +\infty} (\Omega_{\lambda_1, \lambda_2}^{\text{OSCAR}})_{\gamma^{-1/2} \mathbf{I}_n}(\mathbf{x}) = \Omega_{\lambda_1, \lambda_2}^{\text{OSCAR}}(\mathbf{x}), \quad \forall \mathbf{x} \in \mathbb{R}^n. \quad (5.4)$$

The following proposition shows the convergence of the rescaled Moreau enhancement of the OSCAR regularizer. Due to symmetry, it is enough to show the case when $\mathbf{x} \in \mathcal{K}_{\geq 0}^n$ since

$$(\Omega_{\lambda_1, \lambda_2})_{\gamma^{-1/2} \mathbf{I}_n}(\mathbf{x}) = (\Omega_{\lambda_1, \lambda_2})_{\gamma^{-1/2} \mathbf{I}_n}(|\mathbf{x}|_{\downarrow}). \quad (5.5)$$

Proposition 5.1. Let n be an arbitrary integer greater than 1. Then, it holds for any $\mathbf{x} \in \mathcal{K}_{\geq 0}^n$ that

$$\begin{aligned} & \lim_{\gamma \rightarrow +0} 2\gamma^{-1} (\Omega_{\lambda_1, \lambda_2}^{\text{OSCAR}})_{\gamma^{-1/2} \mathbf{I}_n}(\mathbf{x}) \\ &= \begin{cases} \|\mathbf{w}\|_2^2, & \text{if } \mathbf{x} \in \mathbb{R}_{++}^n \cap \mathcal{K}_{>}^n, \\ & \text{(case 1)} \\ \|\mathbf{w}\|_2^2 - \sum_{l=1}^q \sum_{j \in S_l} \left(w_j - \frac{\sum_{k \in S_l} w_k}{\text{card}(S_l)} \right)^2, & \text{if } \mathbf{x} \in \mathbb{R}_{++}^n \cap (\mathcal{K}_{>}^n)^c, \\ & \text{(case 2)} \\ \|\mathbf{w}\|_2^2 - w_n^2, & \text{if } \mathbf{x} \in (\mathbb{R}_{++}^n)^c \cap \mathcal{K}_{>}^n, \\ & \text{(case 3a)} \\ \|\mathbf{w}\|_2^2 - \sum_{l=1}^q \sum_{j \in S_l} \left(w_j - \frac{\sum_{k \in S_l} w_k}{\text{card}(S_l)} \chi_{\mathbb{R}_{++}}(x_j) \right)^2 \\ \quad - w_n^2 \chi_{\mathbb{R}_{++}}(x_{n-1}), & \text{if } \mathbf{x} \in \mathcal{K}_{\geq 0}^n \setminus (\mathbb{R}_{++}^n \\ & \quad \cup \mathcal{K}_{>}^n \cup \{\mathbf{0}\}), \\ & \text{(case 3b)} \\ 0, & \text{if } \mathbf{x} = \mathbf{0}, \\ & \text{(case 4)} \end{cases} \end{aligned} \quad (5.6)$$

where

$$\mathcal{K}_{>}^n := \{\mathbf{x} \in \mathbb{R}^n \mid x_1 > x_2 > \dots > x_n\}, \quad (5.7)$$

$$\chi_{\mathbb{R}_{++}} : \mathbb{R}_{++} \rightarrow \{0, 1\} : x \mapsto \begin{cases} 1, & \text{if } x \in \mathbb{R}_{++}, \\ 0, & \text{if } x \notin \mathbb{R}_{++}, \end{cases} \quad (5.8)$$

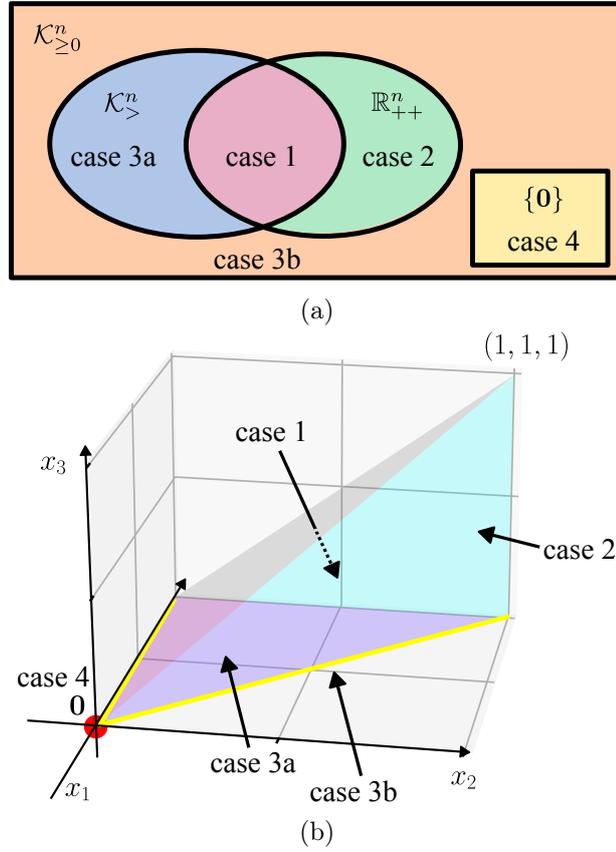


Figure 5.2: (a) Inclusion relation among the five cases of (5.6). (b) Visualization of the five cases in the three-dimensional case.

$\mathbf{w} \in \mathbb{R}^n$ is defined by (2.69), S^c denotes the complement set of any set $S \subset \mathbb{R}^n$, and S_l with $\text{card}(S_l) \geq 2$ is the l th group of consecutive indices for $l = 1, 2, \dots, q$, $q \in \mathbb{N}$, such that

$$x_j = x_k, \forall j, k \in S_l, \text{ and}, \quad (5.9)$$

$$x_j \neq x_k, \forall j \in S_l, \forall k \in \{1, 2, \dots, n\} \setminus S_l. \quad (5.10)$$

For example, it holds that $S_1 = \{3, 4, 5\}$, $S_2 = \{7, 8\}$, $S_3 = \{8, 9\}$ for

$$x_1 > x_2 > \underbrace{x_3 = x_4 = x_5}_{\text{first group}} > x_6 > \underbrace{x_7 = x_8}_{\text{second group}} > \underbrace{x_9 = x_{10}}_{\text{third group}} > x_{11}. \quad (5.11)$$

Proof. The proof is given in Appendix K.1. \square

The inclusion relation of the cases of (5.6) and visualization of the cases in the three-dimensional case are depicted in Figures 5.2(a) and 5.2(b), respectively.

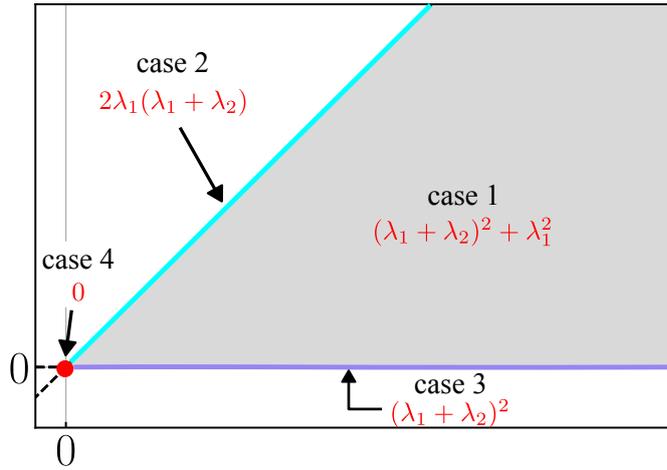


Figure 5.3: The function value of (5.12).

Corollary 5.1. Let $n := 2$. Then, for any $\mathbf{x} \in \mathcal{K}_{\geq 0}^2$, it holds that

$$\begin{aligned} & \lim_{\gamma \rightarrow +0} 2\gamma^{-1} (\Omega_{\lambda_1, \lambda_2}^{\text{OSCAR}})_{\gamma^{-1/2} \mathbf{I}_n}(\mathbf{x}) \\ &= \begin{cases} (\lambda_1 + \lambda_2)^2 + \lambda_1^2, & \text{if } x_1 > x_2 > 0, \\ 2\lambda_1(\lambda_1 + \lambda_2), & \text{if } x_1 = x_2 > 0, \\ (\lambda_1 + \lambda_2)^2, & \text{if } x_1 > x_2 = 0, \\ 0, & \text{if } x_1 = x_2 = 0. \end{cases} \end{aligned} \quad (5.12)$$

We first explain the case $n := 2$. Equation (5.12) indicates that the rescaled Moreau enhancement converges to a certain discrete function shown in Figure 5.3. Unlike (5.6), (5.12) is divided into four cases (case 3b does not exist for the two-dimensional case). The function returns lower values on the sets $\{\mathbf{x} \in \mathcal{K}_{\geq 0}^2 \mid x_1 = x_2 \geq 0\}$ and $\{\mathbf{x} \in \mathcal{K}_{\geq 0}^2 \mid x_1 \geq x_2 = 0\}$ than the other regions. Hence, when (5.12) is used as a penalty, the solutions $\hat{\mathbf{x}}$ tend to satisfy

$$\hat{x}_1 = \hat{x}_2 \geq 0 \text{ or } \hat{x}_1 \geq \hat{x}_2 = 0. \quad (5.13)$$

Figure 5.4 shows the contours of the Moreau enhancement of the OSCAR regularizer (5.12) for the two-dimensional case. It can be seen that the contours of the Moreau enhancement sharpen those of the OSCAR regularizer. Hence, the solutions are more likely to satisfy (5.13).

Moreover, Proposition 5.1 indicates that the limitation of the rescaled Moreau enhancement of the OSCAR regularizer is bounded above by a constant $\|\mathbf{w}\|_2^2$, and is lower than this bound on the set $(\mathcal{K}_{>}^n \cap \mathbb{R}_{++}^n)^c$. Hence, when (5.6) is used as a penalty, the solutions tend to lie on this set.

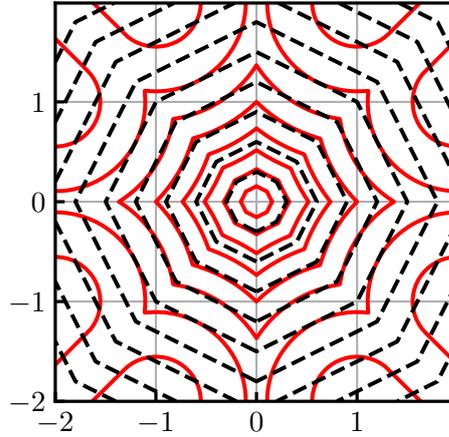


Figure 5.4: The contours of the Moreau enhancement of OSCAR (red) for $\lambda_1 = \lambda_2 := 0.5$ and $\gamma := 2$ and OSCAR (black) for the two-dimensional case.

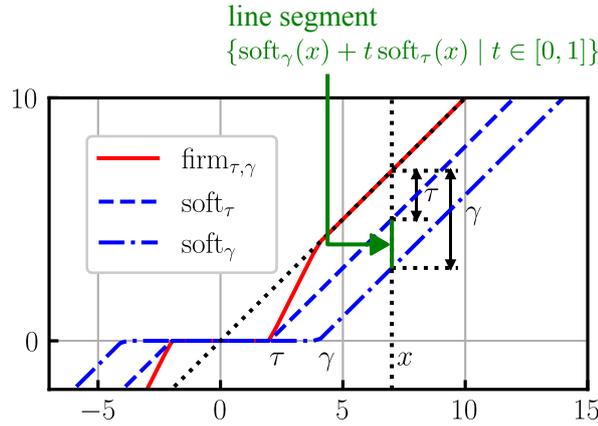


Figure 5.5: A relation between the firm-shrinkage and two soft-shrinkage operators.

5.2.2 Firm-Shrinkage Operator as External Division

Figure 5.5 shows the relation between the firm-shrinkage operator and two associated soft-shrinkage operators. It can be seen that, for any $x \in \mathbb{R}$, $\text{firm}_{\tau, \gamma}(x)$ externally divides the line segment $\{\text{soft}_{\gamma}(x) + t \text{soft}_{\tau}(x) \mid t \in [0, 1]\}$ between the points $\text{soft}_{\gamma}(x)$ and $\text{soft}_{\tau}(x)$ in the ratio of γ/τ . This is verified in the following proposition by expressing the firm-shrinkage operator as an external division of two soft-shrinkage operators.

Proposition 5.2. For any γ, τ satisfying $\gamma > \tau > 0$, it holds that

$$\text{Firm}_{\tau, \gamma} = \frac{\gamma}{\gamma - \tau} \text{Soft}_{\tau} - \frac{\tau}{\gamma - \tau} \text{Soft}_{\gamma}. \quad (5.14)$$

Proof. By (2.76), it holds for any $x \in \mathbb{R}$ that

$$\text{firm}_{\tau,\gamma}(x) = \begin{cases} x, & \text{if } |x| \geq \gamma, \\ \frac{\gamma}{\gamma - \tau} \text{soft}_{\tau}(x), & \text{if } |x| < \gamma. \end{cases} \quad (5.15)$$

For the case when $|x| \geq \gamma$, it holds that

$$\begin{aligned} x &= \frac{\gamma}{\gamma - \tau}(x - \tau \text{sign}(x)) - \frac{\tau}{\gamma - \tau}(x - \gamma \text{sign}(x)) \\ &= \frac{\gamma}{\gamma - \tau} \text{soft}_{\tau}(x) - \frac{\tau}{\gamma - \tau} \text{soft}_{\gamma}(x). \end{aligned} \quad (5.16)$$

On the other hand, for the case when $|x| < \gamma$, $\text{soft}_{\gamma}(x) = 0$ yields that

$$\frac{\gamma}{\gamma - \tau} \text{soft}_{\tau}(x) = \frac{\gamma}{\gamma - \tau} \text{soft}_{\tau}(x) - \frac{\tau}{\gamma - \tau} \text{soft}_{\gamma}(x). \quad (5.17)$$

Hence, (5.14) follows from (5.15), (5.16), and (5.17). \square

Note that Proposition 5.2 also gives the relation regarding an inner division as follows:

$$\text{soft}_{\tau} = \frac{\gamma - \tau}{\gamma} \text{firm}_{\tau,\gamma} + \frac{\tau}{\gamma} \text{soft}_{\gamma}. \quad (5.18)$$

It is known that the firm-shrinkage operator is the proximity operator of the MC penalty which is the Moreau enhancement of the ℓ_1 norm [25, 34]. Proposition 5.2 provides an alternative view of the firm-shrinkage operator as an external division of Soft_{τ} and Soft_{γ} . The operator $\text{Firm}_{\tau,\gamma}$ is reproduced by Δ_{ω} defined in (5.1) for $g_1 := \tau \|\cdot\|_1$, $g_2 := \gamma \|\cdot\|_1$, and $\omega := \gamma/(\gamma - \tau)$.

5.2.3 Proposed Bias Reduction Approach for OSCAR: De-biased OSCAR

By analogy with the firm-shrinkage operator, we propose the following external division operator and refer to it as DOSCAR:

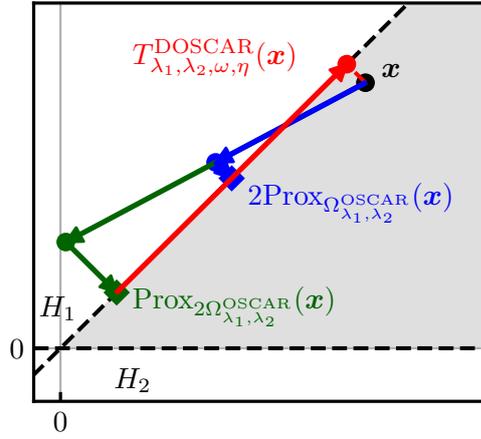
$$T_{\lambda_1, \lambda_2, \omega, \eta}^{\text{DOSCAR}} := \omega \text{Prox}_{\Omega_{\lambda_1, \lambda_2}^{\text{OSCAR}}} - (\omega - 1) \text{Prox}_{\eta \Omega_{\lambda_1, \lambda_2}^{\text{OSCAR}}} \quad (5.19)$$

$$= \text{Prox}_{\Omega_{\lambda_1, \lambda_2}^{\text{OSCAR}}} + (\omega - 1) \left(\text{Prox}_{\Omega_{\lambda_1, \lambda_2}^{\text{OSCAR}}} - \text{Prox}_{\eta \Omega_{\lambda_1, \lambda_2}^{\text{OSCAR}}} \right). \quad (5.20)$$

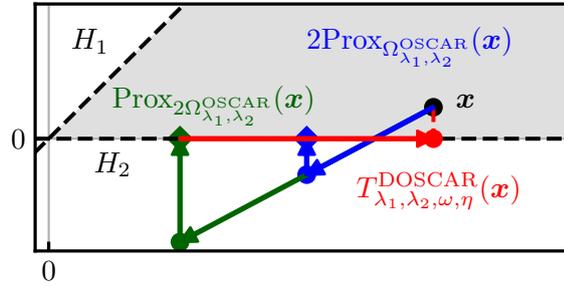
Figure 5.6 depicts a geometric interpretation of $T_{\lambda_1, \lambda_2, \omega, \eta}^{\text{DOSCAR}}(\mathbf{x})$ for $\mathbf{x} \in \mathbb{R}^2$. We only consider the case when $x_1 \geq x_2 \geq 0$ due to symmetry, and set λ_1 and λ_2 to not too large values and $\omega = \eta = 2$ for simplicity. Figure 5.6(a) illustrates the case when \mathbf{x} is close to the hyperplane

$$H_1 := \{\mathbf{x} \in \mathbb{R}^2 \mid x_1 = x_2\}, \quad (5.21)$$

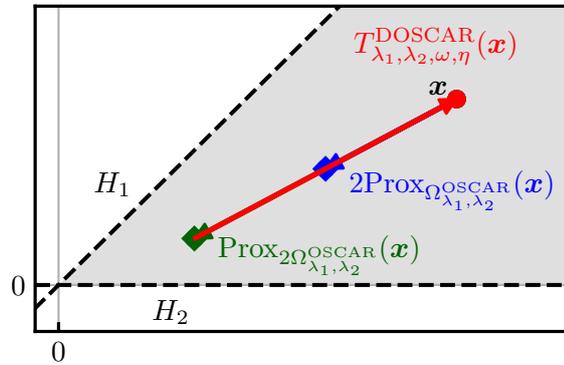
i.e., x_1 and x_2 are in the same group. In this case, $\text{Prox}_{\Omega_{\lambda_1, \lambda_2}^{\text{OSCAR}}}$ yields a point



(a)



(b)



(c)

Figure 5.6: A geometric interpretation of $T_{\lambda_1, \lambda_2, \omega, \eta}^{DOSCAR}(\mathbf{x})$ for the two-dimensional case. (a) x_1 and x_2 are in the same group, (b) x_2 is close to 0, and (c) x_1 and x_2 are in the different groups. The gray regions show $\mathbb{R}_{++}^2 \cap \mathcal{K}_{\geq 0}^2$.

where two components become identical, *i.e.*, $\text{Prox}_{\Omega_{\lambda_1, \lambda_2}^{\text{OSCAR}}}(\mathbf{x}) \in H_1$. However, this shrinkage causes estimation bias by shrinking the large component with a large weight and shrinking the small component with a small weight. In contrast, for DOSCAR, the second term of (5.20) pushes $\text{Prox}_{\Omega_{\lambda_1, \lambda_2}^{\text{OSCAR}}}(\mathbf{x})$ back to the current estimate \mathbf{x} along the hyperplane H_1 . As a result, $T_{\lambda_1, \lambda_2, \omega, \eta}^{\text{DOSCAR}}(\mathbf{x})$ coincides with the projection of \mathbf{x} onto H_1 , so that the estimation bias is reduced. Figure 5.6(b) illustrates the case when x_2 is close to 0. In a similar way, the second term of (5.20) pushes the first component of $\text{Prox}_{\Omega_{\lambda_1, \lambda_2}^{\text{OSCAR}}}(\mathbf{x})$ back to x_1 while letting the second component stay 0. As a result, the estimation bias is reduced since $T_{\lambda_1, \lambda_2, \omega, \eta}^{\text{DOSCAR}}(\mathbf{x})$ coincides with the projection of \mathbf{x} onto

$$H_2 := \{\mathbf{x} \in \mathbb{R}^2 \mid x_2 = 0\}. \quad (5.22)$$

Figure 5.6(c) illustrates the case when \mathbf{x} is distant from H_1 and H_2 , *i.e.*, x_1 and x_2 correspond to different groups of important features. In this case, $\text{Prox}_{\Omega_{\lambda_1, \lambda_2}^{\text{OSCAR}}}$ shrinks \mathbf{x} diagonally. This is undesirable since the two coefficients become more similar although they correspond to the different groups. In contrast, $T_{\lambda_1, \lambda_2, \omega, \eta}^{\text{DOSCAR}}$ keeps \mathbf{x} at the same position. On the basis of these arguments, $T_{\lambda_1, \lambda_2, \omega, \eta}^{\text{DOSCAR}}$ is expected to reduce the estimation bias while preserving the advantages of OSCAR.

5.3 Proposed Method: External Division Operator

First, we establish some basic properties for the external division of general nonexpansive operators. Following this, we show that the external division operator is expressed as the proximity operator of the weakly convex function φ_ω (defined in (5.31) below) if the function ψ_ω (defined in (5.29) below) is convex. Then, the convergence of the gradient algorithm based on the external division operator to a minimizer of the cost function involving φ_ω as a penalty is guaranteed. Subsequently, we present the convexity conditions of ψ_ω and a closed-form expression of φ_ω . Finally, the relation between φ_ω and the Moreau enhancement is investigated. Although we focus on the finite-dimensional case for simplicity in this chapter, most results can be extended for an infinite-dimensional Hilbert space case. For convenience, the relations of the theoretical results to be presented in this chapter are summarized in Figure 5.7.

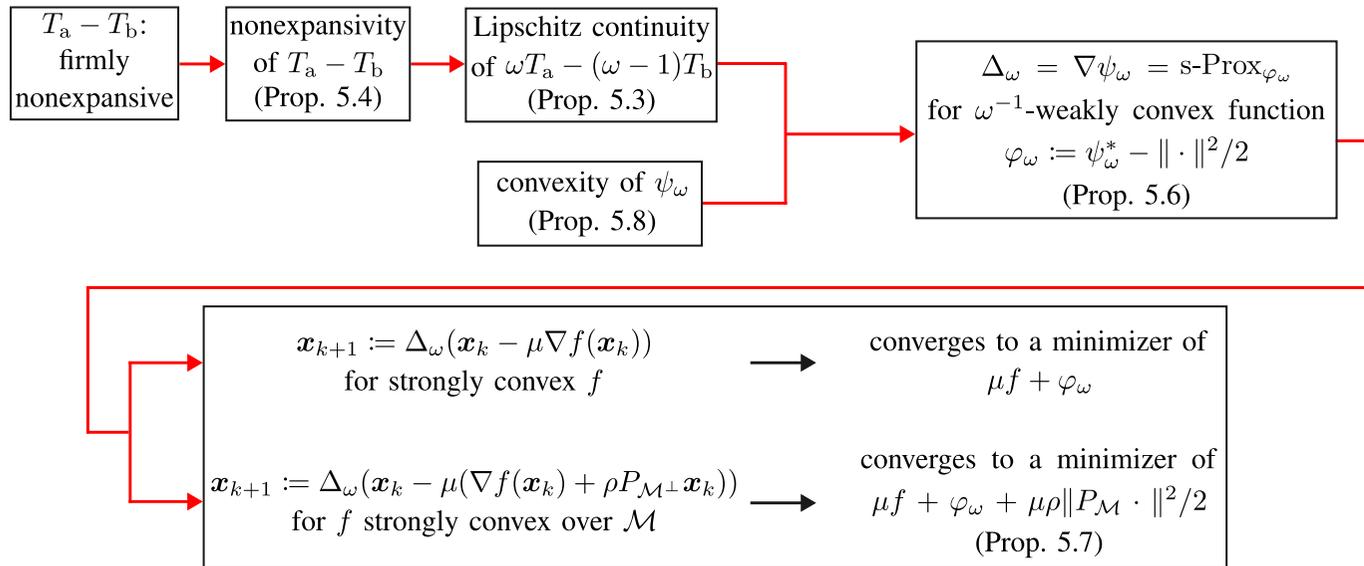


Figure 5.7: An overview of relations of the theoretical results in this chapter.

5.3.1 External Division of Two Nonexpansive Operators

First, we show some basic properties of the external division of two general nonexpansive operators. The following two propositions indicate that the external division of firmly nonexpansive operators is Lipschitz continuous.

Proposition 5.3. Let $T_a, T_b : \mathbb{R}^n \rightarrow \mathbb{R}^n$ be nonexpansive operators such that $T_a - T_b$ is nonexpansive. Then, $\omega T_a - (\omega - 1)T_b$ is ω -Lipschitz continuous for given $\omega > 1$.

Proof. Since it holds that

$$\omega T_a - (\omega - 1)T_b = T_a + (\omega - 1)(T_a - T_b), \quad (5.23)$$

it holds for any $\mathbf{x}, \boldsymbol{\xi} \in \mathbb{R}^n$ that

$$\begin{aligned} & \|(\omega T_a - (\omega - 1)T_b)(\mathbf{x}) - (\omega T_a - (\omega - 1)T_b)(\boldsymbol{\xi})\|_2 \\ & \leq \|T_a \mathbf{x} - T_a \boldsymbol{\xi}\|_2 + (\omega - 1)\|(T_a - T_b)(\mathbf{x}) - (T_a - T_b)(\boldsymbol{\xi})\|_2 \\ & \leq \|\mathbf{x} - \boldsymbol{\xi}\|_2 + (\omega - 1)\|\mathbf{x} - \boldsymbol{\xi}\|_2 \\ & = \omega \|\mathbf{x} - \boldsymbol{\xi}\|_2, \end{aligned} \quad (5.24)$$

where the second inequality is due to the nonexpansiveness of T_a and $T_a - T_b$. This implies that $\omega T_a - (\omega - 1)T_b$ is ω -Lipschitz continuous. \square

Proposition 5.4. Let $T_a, T_b : \mathbb{R}^n \rightarrow \mathbb{R}^n$ be firmly nonexpansive operators. Then, $T_a - T_b$ is nonexpansive.

Proof. Since T_a, T_b are firmly nonexpansive operators, there exist nonexpansive operators N_1 and N_2 satisfying $T_a = (1/2)\text{Id} + (1/2)N_1$ and $T_b = (1/2)\text{Id} + (1/2)N_2$. Hence, it holds, for any $\mathbf{x}, \boldsymbol{\xi} \in \mathbb{R}^n$, that

$$\begin{aligned} \|(T_a - T_b)(\mathbf{x}) - (T_a - T_b)(\boldsymbol{\xi})\|_2 & = \|(1/2)(N_1 \mathbf{x} - N_1 \boldsymbol{\xi}) - (1/2)(N_2 \mathbf{x} - N_2 \boldsymbol{\xi})\|_2 \\ & \leq (1/2)\|N_1 \mathbf{x} - N_1 \boldsymbol{\xi}\|_2 + (1/2)\|N_2 \mathbf{x} - N_2 \boldsymbol{\xi}\|_2 \\ & \leq \|\mathbf{x} - \boldsymbol{\xi}\|_2, \end{aligned} \quad (5.25)$$

which implies that $T_a - T_b$ is nonexpansive. \square

The following proposition shows the relation between the set of common fixed points of given two nonexpansive operators and that of the external division of the two operators.

Proposition 5.5. Let $T_a, T_b : \mathbb{R}^n \rightarrow \mathbb{R}^n$ be nonexpansive operators. Assume that $\text{Fix}(T_a), \text{Fix}(T_b) \neq \emptyset$. Then,

$$\text{Fix}(\omega T_a - (\omega - 1)T_b) \supset \text{Fix}(T_a) \cap \text{Fix}(T_b), \quad (5.26)$$

for a given $\omega > 1$.

Proof. Let $\mathbf{x} \in \text{Fix}(T_a) \cap \text{Fix}(T_b)$. Then,

$$\begin{aligned} (\omega T_a - (\omega - 1)T_b)(\mathbf{x}) &= \omega T_a \mathbf{x} - (\omega - 1)T_b \mathbf{x} \\ &= \omega \mathbf{x} - (\omega - 1)\mathbf{x} \\ &= \mathbf{x}, \end{aligned} \tag{5.27}$$

which yields $\mathbf{x} \in \text{Fix}(\omega T_a - (\omega - 1)T_b)$. Hence, (5.26) holds. \square

This proposition states that the common fixed points of given two non-expansive operators are also fixed points of the external division. However, the opposite does not hold in general as shown in the following remark.

Remark 5.1. The opposite inclusion of (5.26) does not hold in general. For instance, let $T_a = \text{Id} - \text{soft}_2$ and $T_b = \text{Id} - \text{soft}_1$. Then, T_a and T_b are nonexpansive, and $\text{Fix}(T_a) = [-2, 2]$ and $\text{Fix}(T_b) = [-1, 1]$. On the other hand, $(2T_a - T_b)(3) = (\text{Id} - 2\text{soft}_b + \text{soft}_a)(3) = 3 - 2 + 2 = 3$, and hence $3 \in \text{Fix}(2T_a - T_b)$.

5.3.2 External Division Operator of Two Proximity Operators

We present a sufficient condition for the external division operator to be expressed as the proximity operator of a certain weakly convex function. This property is a key ingredient in convergence analysis to be presented in Section 5.3.3. A key idea is to use the framework of MoL-Grad denoisers [139] (see Appendix G).

Proposition 5.6. Let $g_1, g_2 \in \Gamma_0(\mathbb{R}^n)$ and $\omega > 1$. Set $\Delta_\omega := \omega \text{Prox}_{g_1} - (\omega - 1) \text{Prox}_{g_2}$ as in (5.1). Then, Δ_ω is ω -Lipschitz continuous, and it holds that

$$\Delta_\omega = \nabla \psi_\omega, \tag{5.28}$$

where

$$\psi_\omega := \omega \text{Prox}_{g_1} - (\omega - 1) \text{Prox}_{g_2} \tag{5.29}$$

(see Figure 5.8). If ψ_ω is convex, it holds that

$$\Delta_\omega = \text{s-Prox}_{\varphi_\omega} \tag{5.30}$$

in the sense of (2.20) with the $(1 - \omega^{-1})$ -weakly convex function

$$\varphi_\omega := \psi_\omega^* - \|\cdot\|_2^2/2. \tag{5.31}$$

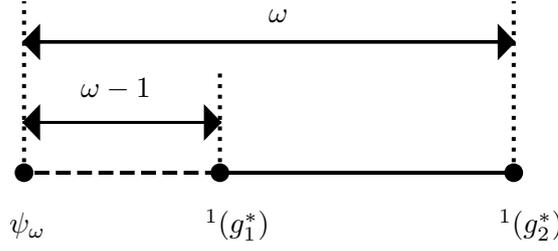


Figure 5.8: The relation among ψ_ω given in (5.29) and the Moreau envelopes $\nabla^{-1}(g_1^*)$ and $\nabla^{-1}(g_2^*)$ of the conjugate functions g_1^* and g_2^* .

Proof. Since Prox_{g_1} and Prox_{g_2} are firmly nonexpansive, $\text{Prox}_{g_1} - \text{Prox}_{g_2}$ is nonexpansive by Proposition 5.4. Hence, Δ_ω is ω -Lipschitz continuous by Propositions 5.3. It holds by Fact 2.4(b) that

$$\begin{aligned}
 \Delta_\omega &= \omega \text{Prox}_{g_1} - (\omega - 1) \text{Prox}_{g_2} \\
 &= \omega(\text{Id} - \text{Prox}_{g_1^*}) - (\omega - 1)(\text{Id} - \text{Prox}_{g_2^*}) \\
 &= \omega \nabla^{-1}(g_1^*) - (\omega - 1) \nabla^{-1}(g_2^*) \\
 &= \nabla \left(\underbrace{\omega(\nabla^{-1}(g_1^*)) - (\omega - 1)(\nabla^{-1}(g_2^*))}_{=\psi_\omega} \right). \tag{5.32}
 \end{aligned}$$

If ψ_ω is convex, (5.30) holds by Fact G.1. \square

The expression of φ_ω given in (5.31) involves the conjugate of ψ_ω , which is not computable in general. An alternative expression of φ_ω will be given in Proposition 5.9 below. In the following subsection, we will provide the convergence analysis of the gradient algorithm based on the external division operator under the assumption of convexity of ψ_ω . The convexity conditions of ψ_ω will be discussed in Section 5.3.5.

5.3.3 Gradient Algorithm With External Division Operator: Case of Strongly Convex Fidelity Term

Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be the fidelity function which is ρ -strongly convex with κ -Lipschitz continuous gradient for $\rho, \kappa > 0$. A popular approach to suppress the fidelity f while accommodating the prior information into a regularizer $R \in \Gamma_0(\mathbb{R}^n)$ is to use the proximal gradient algorithm, *i.e.*, the recursion

$$\mathbf{x}_{k+1} := \text{Prox}_{\mu R}(\mathbf{x}_k - \mu \nabla f(\mathbf{x}_k)). \tag{5.33}$$

By analogy, we consider replacing the proximity operator of the proximal gradient algorithm by Δ_ω , obtaining the recursion

$$\mathbf{x}_{k+1} := \Delta_\omega(\mathbf{x}_k - \mu \nabla f(\mathbf{x}_k)), \tag{5.34}$$

Algorithm 5.1 : Gradient algorithm with the external division operator.

Set $\mathbf{x}_0 \in \mathbb{R}^n$, $\omega, \rho, \mu > 0$.

For $k = 0, 1, 2, \dots$

$\mathbf{x}_{k+1} := \Delta_\omega(\mathbf{x}_k - \mu(\nabla f(\mathbf{x}_k) + \rho P_{\mathcal{M}^\perp} \mathbf{x}_k))$.

where $\mu > 0$. Thanks to Proposition 5.6, under the convexity of ψ_ω , the convergence of the algorithm given in (5.34) to a minimizer (if exists) of $\mu f + \varphi_\omega$ is guaranteed by Fact G.2 under the conditions $\omega \in (1, (\kappa + \rho)/(\kappa - \rho))$ and $\mu \in [(1 - \omega^{-1})/\rho, (1 + \omega^{-1})/\kappa)$.

5.3.4 Gradient Algorithm With External Division Operator: Case When Strong Convexity is Restricted to Subspace

We consider a more general case when the strong convexity of f is restricted to some subspace $\mathcal{M} \subset \mathbb{R}^n$ with the case of underdetermined linear system envisioned (a more specific discussion is given below). In this case, the cost function $\mu f + \varphi_\omega$ is nonconvex since μf is not strongly convex on \mathcal{M}^\perp while φ_ω is weakly convex on \mathcal{M}^\perp . To guarantee the global optimality, we modify the penalty term as $\tilde{\varphi}_\omega$ which will be defined in (5.35). Along with this modification, we modify the recursion in (5.34) and consider Algorithm 5.1. The convergence of Algorithm 5.1 is guaranteed as shown in the following proposition.

Proposition 5.7. Let $\mathcal{M} \subset \mathbb{R}^n$ be a subspace. Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be a differentiable function such that

- (i) $f - \rho \|P_{\mathcal{M}} \cdot\|_2^2/2$ is convex,
- (ii) $\nabla f(\mathbf{x}) \in \mathcal{M}$ for all $\mathbf{x} \in \mathbb{R}^n$,
- (iii) $\nabla f + \rho \|P_{\mathcal{M}^\perp} \cdot\|_2^2/2$ is κ -Lipschitz continuous for $\kappa \geq \rho > 0$.

Let $g_1, g_2 \in \Gamma_0(\mathbb{R}^n)$ be convex such that $\psi_\omega (:= \omega(\cdot - g_1^*) - (\omega - 1)(\cdot - g_2^*))$ defined in (5.29) is convex. Let $\mu \in (0, 2/(\kappa + \rho))$ be the step size parameter, and set $\omega := (1 - \mu\rho)^{-1} > 1$. Suppose that Δ_ω given in (5.1) is employed in Algorithm 5.1. Let

$$\tilde{\varphi}_\omega := \varphi_\omega + \frac{\mu\rho}{2} \|P_{\mathcal{M}^\perp} \cdot\|_2^2 \quad (5.35)$$

for φ_ω defined in (5.31). Then, given an arbitrary $\mathbf{x}_0 \in \mathbb{R}^n$, the sequence $(\mathbf{x}_k)_{k \in \mathbb{N}}$ produced by the algorithm converges to a minimizer, if exists, of

$$\mu f + \tilde{\varphi}_\omega. \quad (5.36)$$

Here, μf and $\tilde{\varphi}_\omega$ are ρ -strongly convex and ρ -weakly convex on \mathcal{M} , respectively. Moreover, μf and $\tilde{\varphi}_\omega$ are both convex on \mathcal{M}^\perp , respectively.

Proof. Proof is given in Appendix K.2. \square

Let us consider the specific case of squared errors $f : \mathbb{R}^n \rightarrow \mathbb{R} : \mathbf{x} \mapsto \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2/2$, where $\mathbf{y} \in \mathbb{R}^m$ and $\mathbf{A} \in \mathbb{R}^{m \times n}$ are the observation vector and the input matrix, respectively. In this case, f is not strongly convex if the linear system is underdetermined ($m < n$). Assumptions (i), (ii), and (iii) in Proposition 5.7 are automatically satisfied for $\mathcal{M} := \text{range } \mathbf{A}^\top$ even in the underdetermined case since (i) $\|\mathbf{y} - \mathbf{A} \cdot\|_2^2/2 - \rho\|P_{\mathcal{M}} \cdot\|_2^2/2$ is convex if ρ is not greater than $\lambda_{\min}^{++}(\mathbf{A}^\top \mathbf{A})$, (ii) $\nabla(\|\mathbf{y} - \mathbf{A} \cdot\|_2^2/2)(\mathbf{x}) = \mathbf{A}^\top(\mathbf{A}\mathbf{x} - \mathbf{y}) \in \mathcal{M}$, and (iii) $\nabla(\|\mathbf{y} - \mathbf{A} \cdot\|_2^2/2 + \rho\|P_{\mathcal{M}^\perp} \cdot\|_2^2/2)(\mathbf{x}) = \mathbf{A}^\top(\mathbf{A}\mathbf{x} - \mathbf{y}) + \rho P_{\mathcal{M}^\perp}$ is κ -Lipschitz continuous, where $\kappa := \lambda_{\max}(\mathbf{A}^\top \mathbf{A})$. Hence, convergence is guaranteed by Proposition 5.7 under $\mu \in (0, 2/(\kappa + \rho))$. This is an essential difference from [112, Proposition 6] which requires the ρ -strong convexity of f .

The idea of Proposition 5.7 is to abandon the weak convexity of the regularizer over \mathcal{M}^\perp to guarantee the convexity of the whole cost function. In the case of $\mathcal{M} := \mathbb{R}^n$, we have $\mathcal{M}^\perp := \{\mathbf{0}\}$, so that $\mu\rho\|P_{\mathcal{M}^\perp} \cdot\|_2^2/2$, as well as $\rho P_{\mathcal{M}^\perp} \mathbf{x}_k$ in the algorithm, vanishes. Hence, in this specific case, Algorithm 5.1 reduces to the algorithm given in (5.34), and the sequence $(\mathbf{x}_k)_{k \in \mathbb{N}}$ converges to a minimizer of $\mu f + \varphi_\omega$, as already mentioned in Section 5.3.3.

5.3.5 Convexity Condition of ψ_ω Given in (5.29)

For given $g_1, g_2 \in \Gamma_0(\mathbb{R}^n)$, the convexity of ψ_ω is a key property to ensure the cocoercivity of Δ_ω and hence the convergence of Algorithm 5.1. However, it is not straightforward to establish the convexity conditions of ψ_ω in general. By comparing (5.1) and (5.14), it is verified that the firm-shrinkage operator is the external division operator for $g_1 = \tau\|\cdot\|_1$ and $g_2 = \gamma\|\cdot\|_1$ for $\gamma > \tau > 0$. One would by analogy expect that ψ_ω is convex if $g_2 := \eta g_1 \in \Gamma_0(\mathbb{R}^n)$ for $\eta > 1$. However, this condition is not sufficient in general (see Remark 5.2). The following proposition gives sufficient conditions for ψ_ω to be convex.

Proposition 5.8. Let $g_1, g_2 \in \Gamma_0(\mathbb{R}^n)$, and $\omega > 1$. Assume that the following hold jointly:

$$(C-1) \quad g_2 - g_1 \in \Gamma_0(\mathbb{R}^n).$$

$$(C-2) \quad \text{Prox}_{g_2} = \text{Prox}_{g_2 - g_1} \circ \text{Prox}_{g_1}.$$

$$(C-3) \quad \text{Prox}_{(g_2 - g_1)^*} \circ \text{Prox}_{g_1} \text{ is monotone.}$$

Then, the following statements hold.

(a) The function ψ_ω given in (5.29) is convex if (C-1), (C-2), and (C-3) are jointly satisfied.

(b) (C-3) is satisfied if one of the following conditions is satisfied:

$$(i) \quad \text{Prox}_{(g_2 - g_1)^*} \circ \text{Prox}_{g_1} = \text{Prox}_{(g_2 - g_1)^* + g_1}.$$

(ii) g_1 and g_2 are separable, *i.e.*, there exist $\tilde{g}_1, \tilde{g}_2 \in \Gamma_0(\mathbb{R})$ such that

$$g_1(\mathbf{x}) = \sum_{i=1}^n \tilde{g}_1(x_i), \quad g_2(\mathbf{x}) = \sum_{i=1}^n \tilde{g}_2(x_i), \quad (5.37)$$

for any $\mathbf{x} \in \mathbb{R}^n$.

Proof. The proof is given in Appendix K.3. \square

Some sufficient conditions for the decomposition of the proximity operator (assumptions (C-2) and (i)) are investigated in [145, 122, 96, 144] (see Appendix H). The following example gives several examples of g_1 and g_2 which satisfy assumptions (C-1)–(C-3).

Example 5.1. Assumptions (C-1)–(C-3) of Propositions 5.8 are satisfied if $g_1, g_2 : \mathbb{R}^n \rightarrow \mathbb{R}$ are set as follows:

- (a) (Generalization of OSCAR) $g_1 : \mathbf{x} \mapsto \sum_{i,j \in \mathcal{I}} w_{i,j} \max\{|x_i|, |x_j|\}$, where² $w_{i,j} \geq 0$ for $i, j \in \mathcal{I} := \{1, 2, \dots, n\}$, and g_2 is any permutation invariant function which satisfies (i) $g_2(\mathbf{x}) = g_2(|\mathbf{x}|)$ for any $\mathbf{x} \in \mathbb{R}^n$ and (ii) $g_2 - g_1 \in \Gamma_0(\mathbb{R}^n)$. Here, for any $g : \mathbb{R}^n \rightarrow \mathbb{R}$, g is called permutation invariant if $g(\mathbf{x}) = g(\mathbf{Q}\mathbf{x})$ for any permutation \mathbf{Q} (under some fixed basis) for any $\mathbf{x} \in \mathbb{R}^n$.

The specific choice of

$$w_{i,j} := \begin{cases} \lambda_1, & \text{if } i = j, \\ \lambda_2, & \text{if } i < j, \\ 0, & \text{otherwise,} \end{cases} \quad (5.38)$$

for $\lambda_1, \lambda_2 > 0$, for g_1 reproduces the OSCAR regularizer $\Omega_{\lambda_1, \lambda_2}^{\text{OSCAR}}$ defined in (2.67). Since the OSCAR regularizer is permutation invariant and satisfies $\Omega_{\lambda_1, \lambda_2}^{\text{OSCAR}}(\mathbf{x}) = \Omega_{\lambda_1, \lambda_2}^{\text{OSCAR}}(|\mathbf{x}|)$ for any $\mathbf{x} \in \mathbb{R}^n$, a specific choice of g_2 is $g_2 := \eta g_1$, where $\eta > 1$.

- (b) (Total variation norm or the clustered lasso) $g_1 : \mathbf{x} \mapsto \sum_{i,j \in \mathcal{I}} w_{i,j} |x_i - x_j|$ for any $w_{i,j} \geq 0$ for $i, j \in \mathcal{I} := \{1, 2, \dots, n\}$, and g_2 is a function such that $g_2 - g_1 \in \Gamma_0(\mathbb{R}^n)$ is permutation invariant.
- (c) (Hinge loss function) $g_1 := \Phi^{\text{hinge}} := \sum_{i=1}^n \phi^{\text{hinge}}$, and $g_2 := \eta g_1$, where $\Gamma_0(\mathbb{R}) \ni \phi^{\text{hinge}} : x \mapsto \max\{0, 1 - x\}$ and $\eta > 1$.

Proof. (a) Since $g_2 - g_1 \in \Gamma_0(\mathbb{R}^n)$, assumption (C-1) of Proposition 5.8 is satisfied. Assumption (C-2) of Proposition 5.8 is satisfied by [145, Corollary 5]. Now we show that assumption (C-3) of Proposition 5.8

²When $w_{i,j} = 0$ for all $i, j \in \mathcal{I}$, the function g_1 is not a norm. Otherwise, g_1 is a norm.

is satisfied. By [145, Proposition 4], g is permutation invariant if and only if $\text{Prox}_g(\mathbf{Q}\mathbf{x}) = \mathbf{Q}\text{Prox}_g(\mathbf{x})$ for any $\mathbf{x} \in \mathbb{R}^n$, which implies that

$$\begin{aligned} \text{Prox}_{g^*}(\mathbf{Q}\mathbf{x}) &= \mathbf{Q}\mathbf{x} - \text{Prox}_g(\mathbf{Q}\mathbf{x}) \\ &= \mathbf{Q}\mathbf{x} - \mathbf{Q}\text{Prox}_g(\mathbf{x}) \\ &= \mathbf{Q}\text{Prox}_{g^*}(\mathbf{x}), \end{aligned} \quad (5.39)$$

where the first and third equalities are due to Fact 2.4(c). Then, by using [145, Proposition 4] again, we obtain that g^* is permutation invariant. Hence, assumption (C-3) is satisfied by permutation invariance of g_1 and [145, Corollary 5].

- (b) The proof is given in the same way as in (a).
- (c) For any $\mathbf{x} \in \mathbb{R}^n$, the hinge loss function is expressed, as a translation of the support function, as

$$\begin{aligned} \Phi^{\text{hinge}}(\mathbf{x}) &= \sum_{i=1}^n \max\{0, 1 - x_i\} \\ &= \max_{\mathbf{v} \in [-1, 0]^n} \sum_{i=1}^n v_i(x_i - 1) \\ &= \sigma_{[-1, 0]^n}(\mathbf{x} - \mathbf{1}_n). \end{aligned} \quad (5.40)$$

Now let us show that the assumption (C-2) of Proposition 5.8 is satisfied. Since Φ^{hinge} is separable, it holds from [96, Proposition 24.11] for any $\gamma > 0$ that

$$\text{Prox}_{\gamma\Phi^{\text{hinge}}}(\mathbf{x}) = (\text{Prox}_{\gamma\phi^{\text{hinge}}}(x_i))_{i=1}^n. \quad (5.41)$$

The proximity operator of the hinge loss function is known [96, Example 24.36] as, for any $\xi \in \mathbb{R}$,

$$\text{Prox}_{\gamma\phi^{\text{hinge}}}(\xi) = \begin{cases} \xi + \gamma, & \text{if } \xi < 1 - \gamma, \\ 1, & \text{if } \xi \in [1 - \gamma, 1], \\ \xi, & \text{if } \xi > 1. \end{cases} \quad (5.42)$$

Hence, it holds that

$$\begin{aligned}
& \text{Prox}_{(\eta-1)\phi^{\text{hinge}}} \circ \text{Prox}_{\phi^{\text{hinge}}}(\xi) \\
&= \begin{cases} \text{Prox}_{\phi^{\text{hinge}}}(\xi) + \eta - 1, & \text{if } \text{Prox}_{\phi^{\text{hinge}}}(\xi) < 2 - \eta, \\ 1, & \text{if } \text{Prox}_{\phi^{\text{hinge}}}(\xi) \in [2 - \eta, 1], \\ \text{Prox}_{\phi^{\text{hinge}}}(\xi), & \text{if } \text{Prox}_{\phi^{\text{hinge}}}(\xi) > 1, \end{cases} \\
&= \begin{cases} \xi + \eta, & \text{if } \xi < 1 - \eta, \\ 1, & \text{if } \xi \in [1 - \eta, 1], \\ \xi, & \text{if } \xi > 1, \end{cases} \\
&= \text{Prox}_{\eta\phi^{\text{hinge}}}(\xi). \tag{5.43}
\end{aligned}$$

Hence, assumption (C-2) of Proposition 5.8 is satisfied by (5.41) and (5.43). Finally, assumption (C-3) of Proposition 5.8 is satisfied since assumption (ii) is satisfied. \square

The following remark gives an example of the case when ψ_ω is nonconvex for g_1, g_2 which satisfy assumption (C-1) of Proposition 5.8 but do not satisfy (C-2).

Remark 5.2. Let $g_1 : \mathbb{R} \rightarrow \mathbb{R} : x \mapsto \begin{cases} -\ln(x), & \text{if } x > 0, \\ +\infty, & \text{otherwise,} \end{cases}$ and $g_2 = 2g_1$.

In this case, assumption (C-1) is satisfied because $g_2 - g_1 = g_1 \in \Gamma_0(\mathbb{R})$. The proximity operator of ηg_1 for $\eta > 0$ is known to be expressed as follows [146]:

$$\text{Prox}_{\eta g_1}(x) = \frac{x + \sqrt{x^2 + 4\eta}}{2}, \tag{5.44}$$

which implies that

$$\begin{aligned}
\text{Prox}_{g_2}(x) &= \frac{x + \sqrt{x^2 + 8}}{2} \\
&\neq \frac{\frac{x + \sqrt{x^2 + 4}}{2} + \sqrt{\left(\frac{x + \sqrt{x^2 + 4}}{2}\right)^2 + 4}}{2} \\
&= \text{Prox}_{g_1}(\text{Prox}_{g_2}(x)). \tag{5.45}
\end{aligned}$$

In this case, for every $\omega > 2\sqrt{5}/(3 - \sqrt{5})$, ψ_ω given in (5.29) is nonconvex. To see this, we observe by (5.32) that, for any $x \in \mathbb{R}$,

$$\begin{aligned}
\frac{d^2}{dx^2}\psi(x) &= \frac{d}{dx}(\omega \text{Prox}_{g_1} - (\omega - 1) \text{Prox}_{2g_1})(x) \\
&= \frac{d}{dx} \frac{1}{2} \left(x + \omega\sqrt{x^2 + 4} - (\omega - 1)\sqrt{x^2 + 8} \right) \\
&= \frac{1}{2} \left(1 + \omega \frac{x}{\sqrt{x^2 + 4}} - (\omega - 1) \frac{x}{\sqrt{x^2 + 8}} \right), \tag{5.46}
\end{aligned}$$

from which it follows that

$$\begin{aligned} \frac{d^2}{dx^2}\psi(-1) &= \frac{1}{2} \left(1 - \omega \frac{1}{\sqrt{5}} + (\omega - 1) \frac{1}{3} \right) \\ &= -\frac{1}{3\sqrt{5}} ((3 - \sqrt{5})\omega - 2\sqrt{5}) \\ &< 0. \end{aligned} \tag{5.47}$$

The following remark shows that the set of assumptions (C-1)–(C-3) of Proposition 5.8 is only sufficient, and it is not necessary.

Remark 5.3. Let $g_1 : \mathbb{R}^n \rightarrow \mathbb{R} : \mathbf{x} \mapsto \frac{1}{2}\|\mathbf{x}\|_2^2$, and $g_2 := 3g_1$. Then, it holds for any $\mathbf{x} \in \mathbb{R}^n$ and $\eta \in \mathbb{R}$ that

$$\begin{aligned} {}^1(\eta g_1)(\mathbf{x}) &= \eta g_1(\text{Prox}_{\eta g_1}(\mathbf{x})) + \frac{1}{2}\|\mathbf{x} - \text{Prox}_{\eta g_1}(\mathbf{x})\|_2^2 \\ &= \eta g_1((1 + \eta)^{-1}\mathbf{x}) + \frac{1}{2}\|\mathbf{x} - (1 + \eta)^{-1}\mathbf{x}\|_2^2 \\ &= \eta(1 + \eta)^{-2}\|\mathbf{x}\|_2^2 + \eta^2(1 + \eta)^{-2}\|\mathbf{x}\|_2^2 \\ &= \eta(1 + \eta)^{-1}\|\mathbf{x}\|_2^2, \end{aligned} \tag{5.48}$$

where it follows from (2.35) that

$$\begin{aligned} \text{Prox}_{\eta\|\cdot\|_2^2/2}(\mathbf{x}) &= (\text{Id} + \eta\partial(\|\cdot\|_2^2/2))^{-1}(\mathbf{x}) \\ &= (1 + \eta)^{-1}\mathbf{x}. \end{aligned} \tag{5.49}$$

Then, it holds for any $\mathbf{x} \in \mathbb{R}^n$ that

$$\begin{aligned} \psi_\omega(\mathbf{x}) &= \omega {}^1(g_1^*)(\mathbf{x}) - (\omega - 1) {}^1((3g_1)^*)(\mathbf{x}) \\ &= \omega {}^1(g_1)(\mathbf{x}) - (\omega - 1) {}^1\left(3g_1\left(\frac{1}{3}\cdot\right)\right)(\mathbf{x}) \\ &= \omega {}^1(g_1)(\mathbf{x}) - (\omega - 1) {}^1\left(\frac{1}{3}g_1\right)(\mathbf{x}) \\ &= \frac{\omega}{2}\|\mathbf{x}\|_2^2 - \frac{\omega - 1}{4}\|\mathbf{x}\|_2^2 \\ &= \frac{1 + \omega}{4}\|\mathbf{x}\|_2^2, \end{aligned} \tag{5.50}$$

from which it follows that ψ_ω is convex. However, we have

$$\begin{aligned} \text{Prox}_{g_2}(\mathbf{x}) &= \frac{1}{4}\mathbf{x} \\ &\neq \frac{1}{6}\mathbf{x} \\ &= \text{Prox}_{\|\cdot\|_2^2}\left(\frac{1}{2}\mathbf{x}\right) \\ &= \text{Prox}_{g_2 - g_1}(\text{Prox}_{g_1}(\mathbf{x})), \end{aligned} \tag{5.51}$$

and hence assumption (C-2) of Proposition 5.8 is violated. Therefore, assumptions (C-1)–(C-3) are not a necessary condition for convexity of ψ_ω .

5.3.6 A Closed-Form Expression of φ_ω Defined in (5.31)

Proposition 5.7 states that Algorithm 5.1 converges to a minimizer (if exists) of the cost function which includes φ_ω as a regularization term. However, φ_ω given in (5.31) is not computable in general due to the conjugate of ψ_ω . The following proposition gives an alternative expression of φ_ω which is computable if $g_2 - g_1$ is proximal.

Proposition 5.9. For given $\omega > 1$, nonempty convex subsets $C_1, C_2 \subset \mathbb{R}^n$ such that $g_1 := \sigma_{C_1}$ and $g_2 := \sigma_{C_2}$ satisfying assumptions (C-1)–(C-3) in Proposition 5.8. Assume that $\text{range}(\Delta_\omega) = \mathbb{R}^n$ for the external division operator Δ_ω defined in (5.1). Set φ_ω as in (5.31). Then, it holds for any $\mathbf{x} \in \mathbb{R}^n$ that

$$\begin{aligned} \varphi_\omega(\mathbf{x}) &= \sigma_{C_1}(\mathbf{x} + (\omega - 1)\text{Prox}_{\omega(\sigma_{C_2} - \sigma_{C_1})}(\mathbf{x})) \\ &\quad - (\omega - 1)\sigma_{C_1}(\text{Prox}_{\omega(\sigma_{C_2} - \sigma_{C_1})}(\mathbf{x})) - (1 - \omega^{-1})^1(\omega(\sigma_{C_2} - \sigma_{C_1}))(\mathbf{x}). \end{aligned} \quad (5.52)$$

Proof. Proof is given in Appendix K.4. □

Figure 5.9 shows φ_ω given in (5.52) for $g_1 := |\cdot|$, $g_2 := \eta|\cdot|$, and $\omega := 2$ with $\eta := 2.0, 1.5, 3.0$. The case (a) satisfies $\eta := \omega(\omega - 1)^{-1}$, and hence φ_ω coincides with the MC function defined in (2.45) (which will be verified by Proposition 5.10 by setting $\lambda_2 := 0$). It can be seen that η controls the derivative of φ_ω at the large inputs. For the Moreau enhancement, the derivative of φ_ω at the large inputs is 0, and hence the estimation bias for the large coefficients can be reduced. We note that Proposition 5.9 can be applied to the cases in Examples 5.1(a) and (b) since the norms are the support functions for some nonempty convex set.

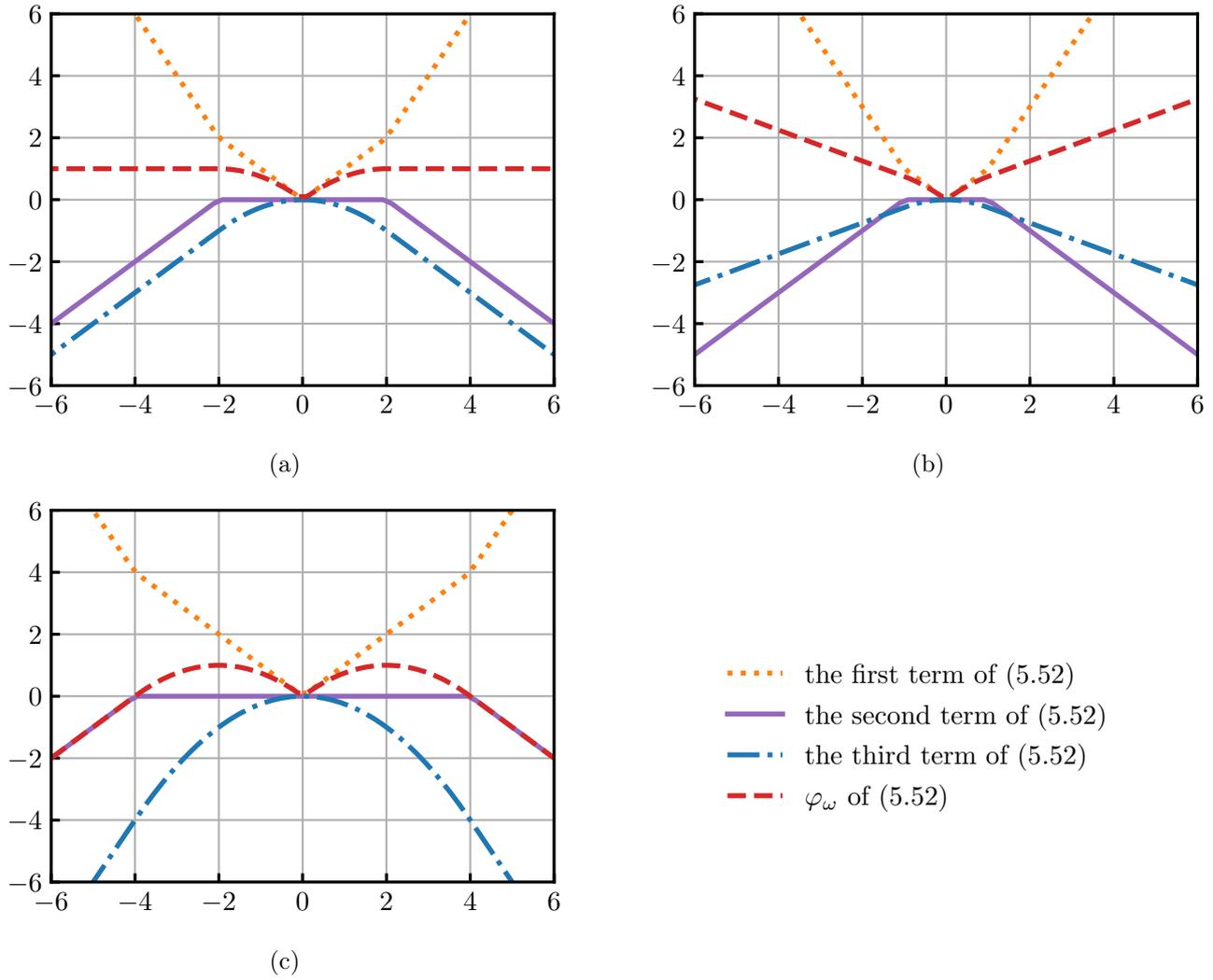


Figure 5.9: The function φ_ω in Proposition 5.9 for $g_1 := |\cdot|$, $g_2 := \eta|\cdot|$, and $\omega := 2$ with (a) $\eta := 2.0$, (b) $\eta := 1.5$, and (c) $\eta := 3.0$.

5.3.7 Relation of φ_ω and the Moreau Enhancement (the Case of $g_2 = \eta g_1$ for $\eta > 1$)

For general $g_1, g_2 \in \Gamma_0(\mathbb{R}^n)$, the following remark provides a connection between φ_ω given in (5.31) and the Moreau enhancement [35, 34]. We use the following lemma.

Lemma 5.1. For any function $g \in \Gamma_0(\mathbb{R}^n)$ and for any $\gamma > 0$, it holds that

$${}^1(\gamma g) = \gamma {}^\gamma g. \quad (5.53)$$

Proof. For any $\mathbf{x} \in \mathbb{R}^n$, it holds that

$$\begin{aligned} {}^1(\gamma g) &= \min_{\mathbf{z} \in \mathbb{R}^n} \left(\gamma g(\mathbf{z}) + \frac{1}{2} \|\mathbf{x} - \mathbf{z}\|_2^2 \right) \\ &= \gamma \min_{\mathbf{z} \in \mathbb{R}^n} \left(g(\mathbf{z}) + \frac{1}{2\gamma} \|\mathbf{x} - \mathbf{z}\|_2^2 \right) \\ &= \gamma {}^\gamma g. \end{aligned} \quad (5.54)$$

□

Remark 5.4. If the functions g_1, g_2 in Proposition 5.9 are support functions, it holds that

$$\varphi_\omega \leq g_1 - (1 - \omega^{-1}) {}^1(\omega(g_2 - g_1)) \quad (5.55)$$

due to (5.52) and the triangle inequality. When $g_2 := \eta g_1$ with

$$\eta = \frac{\omega}{\omega - 1}, \quad (5.56)$$

the majorant (5.55) of φ_ω reduces to the Moreau enhancement of g_1 , *i.e.*,

$$\varphi_\omega \leq g_1 - \omega^{(\omega-1)^{-1}} g_1, \quad (5.57)$$

where

$$\begin{aligned} (1 - \omega^{-1}) {}^1(\omega(\eta - 1)g_1) &= (1 - \omega^{-1}) {}^1(\eta g_1) \\ &= (1 - \omega^{-1}) \eta {}^\eta(g_1) \\ &= {}^\eta g_1 \\ &= \omega^{(\omega-1)^{-1}} g_1. \end{aligned} \quad (5.58)$$

Here, the last equality is due to Lemma 5.1.

Moreover, we show in the following proposition that (5.57) holds with equality when g_1 and g_2 are scaled OSCAR.

Proposition 5.10. Let $g_1 := \Omega_{\lambda_1, \lambda_2}^{\text{OSCAR}}$ and $g_2 := \eta \Omega_{\lambda_1, \lambda_2}^{\text{OSCAR}}$ for $\lambda_1, \lambda_2 \geq 0$ and for $\eta > 1$. Set $\varphi_\omega : \mathbb{R}^n \rightarrow \mathbb{R} : \mathbf{x} \mapsto \psi_\omega^*(\mathbf{x}) - \|\mathbf{x}\|_2^2/2$ as in (5.31) for $\omega > 1$. Then, it holds that

$$\varphi_\omega = \Omega_{\lambda_1, \lambda_2}^{\text{OSCAR}} - ab^b(\Omega_{\lambda_1, \lambda_2}^{\text{OSCAR}}), \quad (5.59)$$

where $a := (1 - \omega^{-1}) \in (0, 1)$, and $b := \omega(\eta - 1) > 0$. Moreover, if $ab = 1$, i.e., $\eta = \omega(\omega - 1)^{-1}$, it holds that

$$\varphi_\omega = \Omega_{\lambda_1, \lambda_2}^{\text{OSCAR}} - b\Omega_{\lambda_1, \lambda_2}^{\text{OSCAR}}. \quad (5.60)$$

Proof. Proof is given in Appendix K.5. \square

This proposition indicates that the corresponding weakly convex function to DOSCAR (see (5.59)) is more general³ than the Moreau enhancement. This general class of weakly convex functions includes those achieving better performance than the Moreau enhancement in general, as demonstrated in Section 5.4.2. Figure 5.10 shows the function φ_ω given in (5.59) with $\lambda_1 = \lambda_2 := 0.5$ and $\omega := 2.0$ ($a := 0.5$) for $\eta := 2.0, 1.5, 3.0$ ($b = 2.0, 1.0, 4.0$, respectively). Since it holds from Fact 2.4(c) that

$$\begin{aligned} b^b(\Omega_{\lambda_1, \lambda_2}^{\text{OSCAR}}) &= b(b^{-1}\|\cdot\|_2^2/2 - b^{-1}((\Omega_{\lambda_1, \lambda_2}^{\text{OSCAR}})^*) \circ (b^{-1}\text{Id})) \\ &= \|\cdot\|_2^2/2 - b^{-1}((\Omega_{\lambda_1, \lambda_2}^{\text{OSCAR}})^*) \circ (b^{-1}\text{Id}), \end{aligned} \quad (5.61)$$

the function φ_ω is a -weakly convex. Hence, the parameter η (or equivalently, b) changes the shapes of φ_ω without changing its concavity for a fixed ω .

³The external division yields a different way of generalization of Moreau enhancement than the generalized Moreau enhancement defined in Definition 2.18.

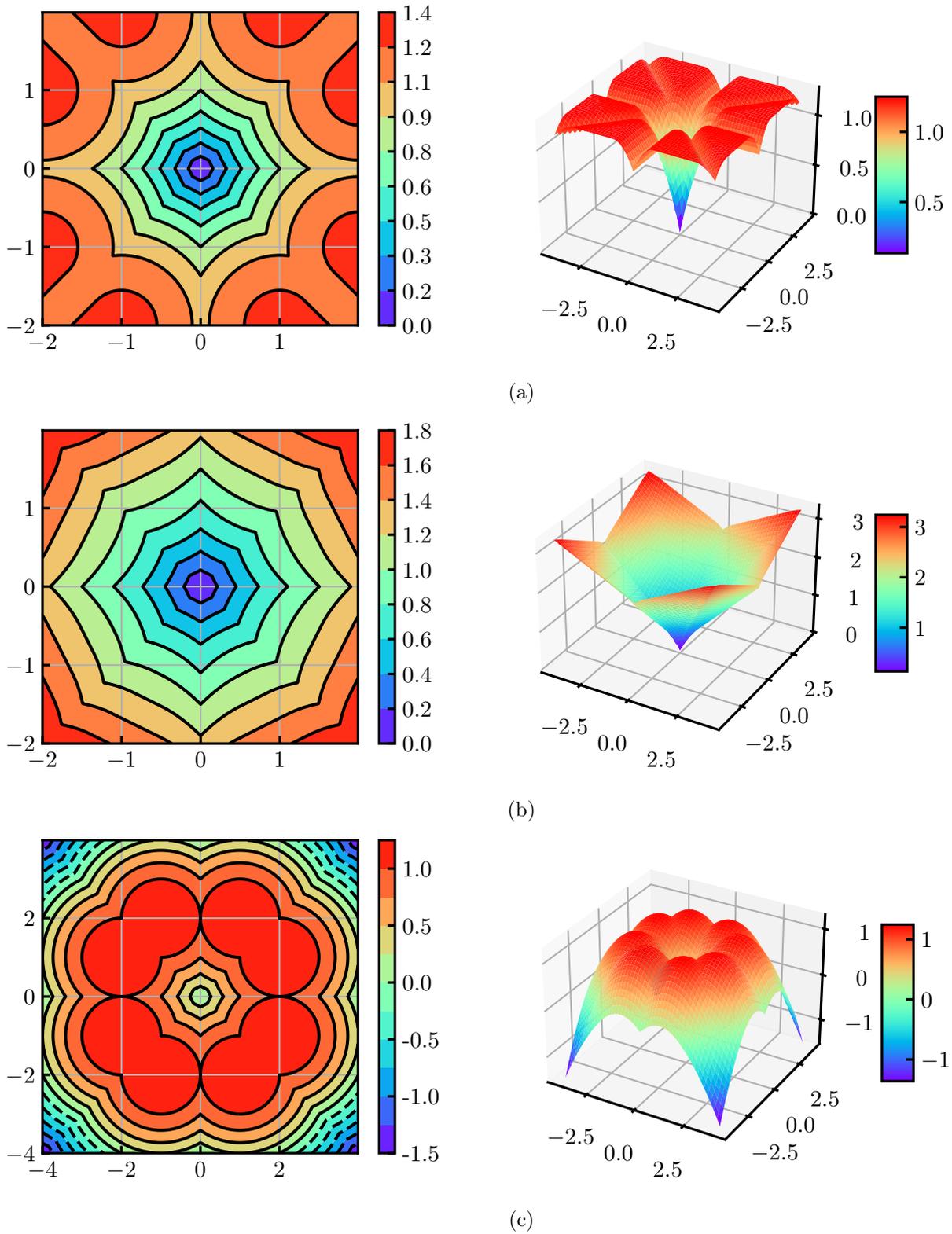


Figure 5.10: The function φ_ω given in (5.59) with $\lambda_1 = \lambda_2 := 0.5$ and $\omega := 2.0$ for (a), $\eta := 2.0$ (the case when φ_ω coincides with the Moreau enhancement of the OSCAR regularizer), (b) $\eta := 1.5$, and (c) $\eta := 3.0$.

5.4 Numerical Examples

We demonstrate the performance of a specific example of the external division operator, DOSCAR, in the application of sparse modeling, particularly when there are groups of highly correlated features. The convergence of Algorithm 5.1 with $T_{\lambda_1, \lambda_2, \omega, \eta}^{\text{DOSCAR}}$ is guaranteed for any $\omega, \eta > 1$ by Propositions 5.6 and 5.8 on the basis of Example 5.1(a). First, the efficacy of the proposed method is evaluated using toy data in Experiment 5-A. Then, we show the influence of hyperparameter η on the performance of the proposed method in Experiment 5-B.

5.4.1 Experiment 5-A: Toy Data

The standard linear model $\mathbf{y} = \mathbf{A}\mathbf{x}_\diamond + \boldsymbol{\epsilon}_\star \in \mathbb{R}^m$ is used. The noise vector $\boldsymbol{\epsilon}_\star \in \mathbb{R}^m$ is generated i.i.d. from the zero-mean Gaussian distribution with a given SNR. We consider the following two toy datasets:

- A. The column vectors of $\mathbf{A} \in \mathbb{R}^{m \times n}$ for $m := 200$ and $n := 80$ are generated as:

$$\mathbf{a}_i := \begin{cases} \tilde{\mathbf{a}}_1 + \boldsymbol{\epsilon}_{\star, i}, & i = 1, 2, \dots, 5, \\ \tilde{\mathbf{a}}_2 + \boldsymbol{\epsilon}_{\star, i}, & i = 6, 7, \dots, 10, \\ \tilde{\mathbf{a}}_3 + \boldsymbol{\epsilon}_{\star, i}, & i = 11, 12, \dots, 15, \\ \tilde{\mathbf{a}}_4, & i = 16, 17, \dots, 80, \end{cases} \quad (5.62)$$

where the components of $\tilde{\mathbf{a}}_j$ ($j = 1, 2, 3, 4$) are generated i.i.d. from standard Gaussian distribution, and those of $\boldsymbol{\epsilon}_{\star, i} \in \mathbb{R}^m$ ($i = 1, 2, \dots, 15$) are generated i.i.d. from $N(0, \sigma_\epsilon^2)$ with $\sigma_\epsilon > 0$. The coefficient vector is set to

$$\mathbf{x}_\diamond := [\underbrace{3 \dots 3}_5, \underbrace{2 \dots 2}_5, \underbrace{1.5 \dots 1.5}_5, \underbrace{0 \dots 0}_{65}]^\top \in \mathbb{R}^n. \quad (5.63)$$

- B. (Toeplitz covariance) Each column vector of $\mathbf{A} \in \mathbb{R}^{m \times n}$ for $n := 80$ is generated from the zero-mean multivariate Gaussian distribution with covariance given by $\text{cov}(\mathbf{a}_i, \mathbf{a}_j) = r^{|i-j|}$ for $r \in [0, 1]$. The coefficient vector is set as

$$\mathbf{x}_\diamond := [\underbrace{1 \dots 1}_{10}, \underbrace{0 \dots 0}_{10}, \underbrace{1 \dots 1}_{10}, \underbrace{0 \dots 0}_{50}]^\top \in \mathbb{R}^n. \quad (5.64)$$

The SNR is set to 20 dB.

We compare DOSCAR with the methods to solve the least squares problem with the ℓ_1 norm (lasso), the MC penalty, and the OSCAR regularizer. Since the overall convexity of the cost function with the MC penalty is not

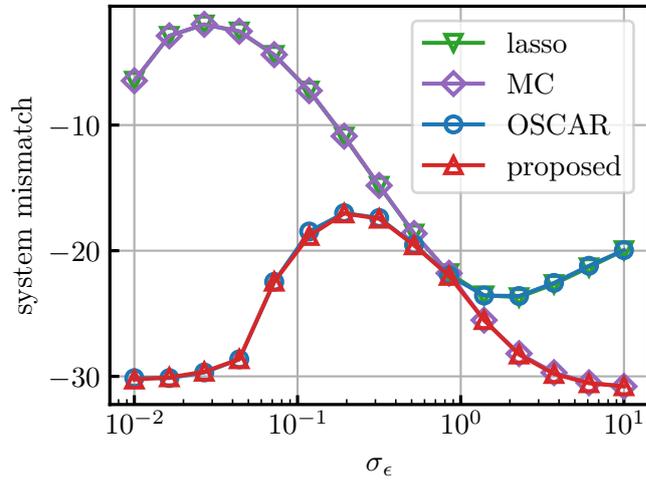
guaranteed for the underdetermined case, the GMC penalty [25, 36] $\mu\Phi_{\sqrt{\gamma/\mu}\mathbf{A}}^{\text{GMC}}$ for $\mu > 0$ and $\gamma \in [0, 1)$ (see (2.47) for definition) is compared in this case. The hyperparameters of the MC and GMC penalties are set so that the whole cost functions are convex. For DOSCAR, the hyperparameters κ and ρ are set to $\lambda_{\max}(\mathbf{A}^\top \mathbf{A})$ and $\lambda_{\min}^{++}(\mathbf{A}^\top \mathbf{A})$, respectively. The hyperparameter ω is set as $\omega := (1 - \mu\rho)^{-1}$, where μ is set to a slightly small value of $2/(\kappa + \rho)$ to guarantee the convergence. All the other hyperparameters are tuned to attain the best performance. The results are averaged over 300 trials, and the evaluation metric is the system mismatch defined in (4.50).

Figure 5.11 shows the system mismatch across σ_ϵ for dataset A for different SNRs. The parameter σ_ϵ controls the correlation of the column vectors of \mathbf{A} : large σ_ϵ corresponds to small correlations, and small σ_ϵ corresponds to high correlations. It can be seen that the proposed method outperforms the other methods in a wide range. For both settings of SNR 20 and 15 dB, the proposed method behaves similarly to the MC when the correlation is low ($\sigma_\epsilon \geq 1.38$), and similarly to OSCAR when the correlation is high ($\sigma_\epsilon \leq 0.51$). When the correlation is low, the proposed method and the MC outperform the others due to the debiasing effect. As the correlation becomes higher, the performance of the proposed method and OSCAR improve due to the grouping effect.

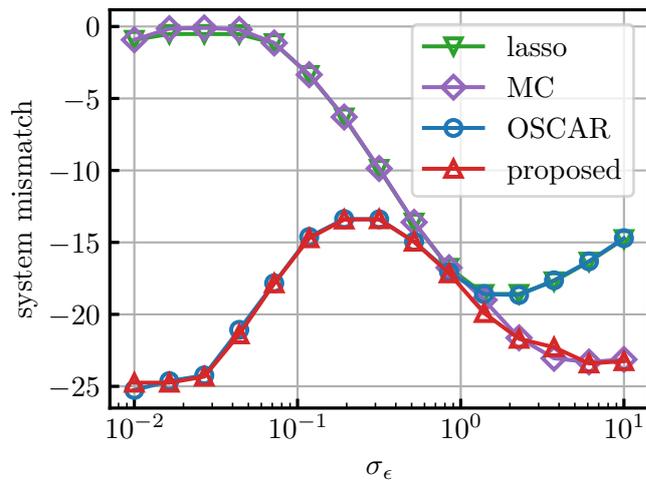
Figure 5.12 shows the system mismatch across r for dataset B. The parameter r controls the correlation of the column vectors of \mathbf{A} : large r corresponds to high correlations, and small r corresponds to small correlations. For both overdetermined and underdetermined cases, the proposed method outperforms the other methods especially when the correlation level is middle. While the proposed method performs well when the correlation is very high for dataset A, such a situation rather causes performance degradation for dataset B. This is because large r implies that even unimportant columns of \mathbf{A} have high correlations with important columns.

5.4.2 Experiment 5-B: Influence of Hyperparameter η

We study the influence of the fluctuations of hyperparameter η on the performance of the proposed method. Figure 5.13 shows the system mismatch of the proposed method across η with tuned λ_1 and λ_2 for datasets A and B. The blue lines indicate $\eta := \omega(\omega - 1)^{-1}$, with which the proposed method reduces to the Moreau enhancement of OSCAR (see Proposition 5.10). Better performance is attained for smaller values of η except for the case of Figure 5.13(c). It can be seen that the behavior of the performance against the fluctuation of η is problem-dependent. In general, the more accurate estimation can be achieved by the proposed method with appropriate tuning of η than by the Moreau enhancement.

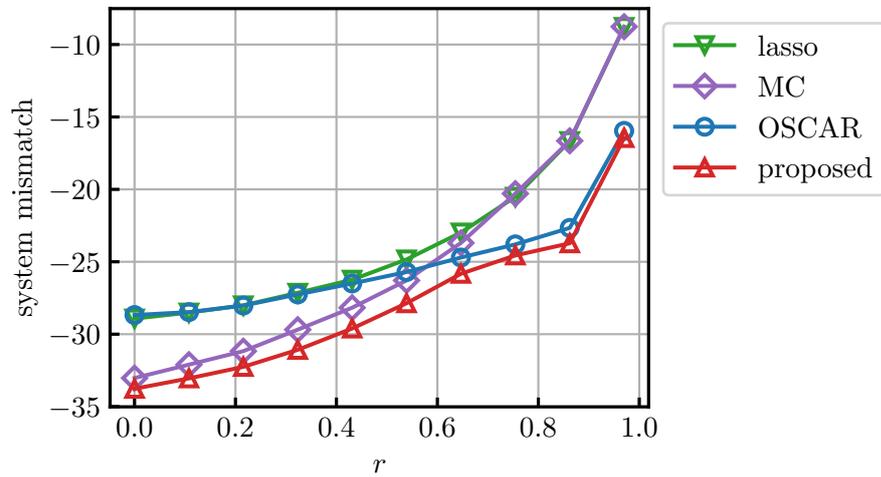


(a)

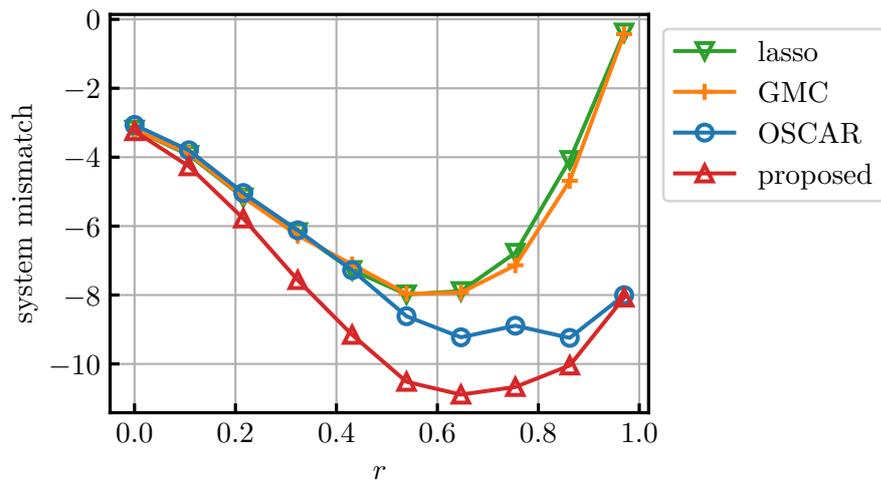


(b)

Figure 5.11: System mismatch across σ_ϵ for dataset A with (a) SNR 20 dB and (b) SNR 15 dB.



(a)



(b)

Figure 5.12: System mismatch across r for dataset B with (a) $m := 400$ and (b) $m := 40$.

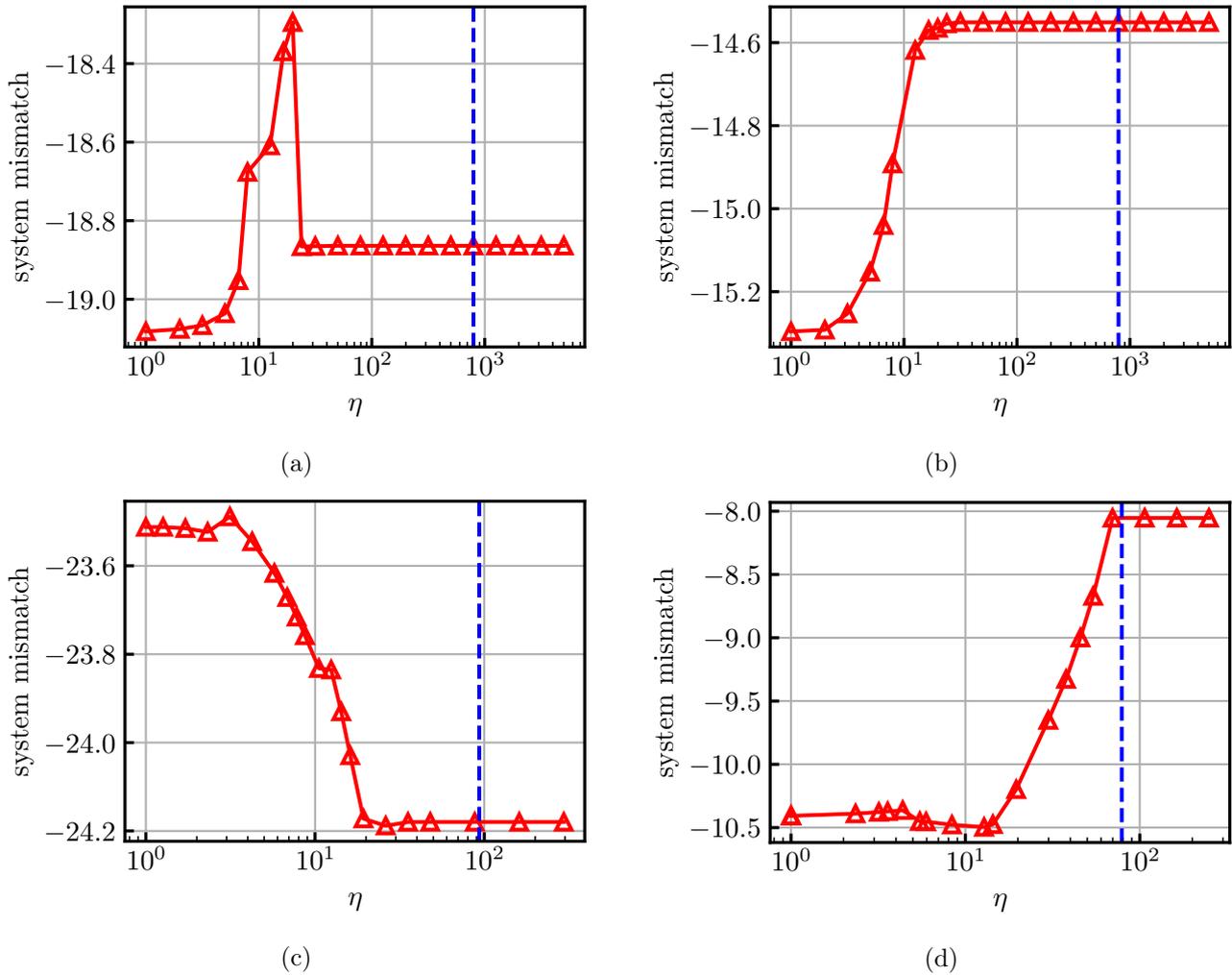


Figure 5.13: System mismatch across hyperparameter η . (a) dataset A with SNR 20 dB and $\kappa := 0.1$, (b) the same as (a) except for SNR 15 dB, (c) dataset B with $m := 400$ and $r := 0.8$, and (d) the same as (c) except for $m := 40$. The blue lines indicate the value $\eta = \omega(\omega - 1)^{-1}$.

5.5 Conclusion

The properties of the external division operator were studied on the basis of the observation that the firm-shrinkage operator can be expressed as an external division of two soft-shrinkage operators. DOSCAR was proposed in the class of the external division operators for an effective feature grouping. This addresses (Q3), which was raised in Chapter 1.2. It was shown that, when ψ_ω defined in (5.29) is convex for given $g_1, g_2 \in \Gamma_0(\mathbb{R}^n)$, the external division operator is expressed as the gradient of some smooth convex function, and it can be expressed as $\Delta_\omega = \nabla\psi_\omega = \text{s-Prox}_{\varphi_\omega}$ for φ_ω defined in (5.31). The gradient algorithm associated with a given fidelity function and the external division operator was proposed, and the convergence to a minimizer of the cost function regularized with φ_ω was proven. This convergence analysis covered the case when the strong convexity of the fidelity is restricted on some subspace. Some sufficient conditions for ψ_ω to be convex were presented, which allowed us to provide a closed-form expression of φ_ω for some cases. Numerical examples demonstrated the efficacy of the proposed operator in recovering signals having structured sparsity, and it was shown that the performance was significantly improved when g_1 and g_2 were defined with scaled OSCAR.

Chapter 6

General Conclusion

This thesis was devoted to robust debiasing methods for sparse modeling. The proposed methods addressed the following two issues which generally occur in outlier-robust estimation and sparse signal estimation, respectively: (i) a severe tradeoff between robustness and global optimality, and (ii) the risk of missing some important groups of highly correlated features. The key idea was to explore an effective way of utilizing the Moreau enhancement technique and to define a new class of operators which extends this technique.

First, to fundamentally solve the above tradeoff problem, we proposed a robust method for jointly sparse signal estimation based on the MC functions. The key idea was using the combination of three terms: the MC loss function, the MC penalty, and the Tikhonov penalty terms. The influence of outliers was significantly reduced by the MC loss function. Moreover, the convexity of the cost function was guaranteed by exploiting the combination of the weak convexity of the MC functions and the strong convexity of the Tikhonov regularizer. Numerical examples including the application of multi-lead electrocardiogram with real data demonstrated the remarkable robustness of the proposed method.

Second, aiming to achieve high performance even in highly noisy environments, we integrated the sparse signal estimation method based on the MC penalty given in (1.3) and SORR given in (F.1.3). This appropriately distinguished the statistical difference between Gaussian noise and outliers by introducing an auxiliary vector to model the noise. In addition, the joint use of the MC and Tikhonov regularizers provided the grouping effect. In contrast to the popular elastic net, it was shown that the upper bound of the discrepancy between the corresponding coefficients is independent of the norm of the observation vector, which could be large owing to huge outliers. The necessary and sufficient conditions for convexity of the smooth part of the cost were derived under a certain nonempty-interior assumption via the product space formulation using the LiMES framework. Numerical examples including the application to speech denoising with real data showed

that the proposed method achieves robust sparse signal estimation even in highly noisy environments.

Third, we introduced a new notion of “the external division operator”, and we presented a method to extract all correlated features accurately. The external division operator was motivated by the observation that the firm-shrinkage operator can be expressed as an external division of two soft-shrinkage operators. The external division operator for OSCAR turned out to be a generalization of its Moreau enhancement. We consider the gradient algorithm associated with a given fidelity function based on the external division operator, and the convergence to a minimizer of the corresponding cost function was proven. This convergence analysis covered the case that strong convexity of the fidelity is restricted to some subspace. It was shown that the external division operator of the OSCAR regularizer is a generalization of the Moreau enhancement of that regularizer. Numerical examples demonstrated that it improves performance dramatically by reducing the estimation bias.

Future Work: There are many ways to continue the study of this thesis. The following are our future work:

- In this thesis, the noise was assumed to obey i.i.d. zero-mean Gaussian distribution. However, it is possible that the proposed methods suffer from errors with other statistical properties, such as the ones obeying heavy-tailed distributions including the Cauchy distribution and symmetric alpha-stable distribution. Further experiments using such errors would be of interest.
- Model selection is a key ingredient in data analysis. In this thesis, most of the hyperparameters of the proposed methods were tuned by a simple grid search based on the information of the ground truth. To apply the proposed methods to various kinds of real data, we would like to explore effective approaches to model selection such as information criterion for the proposed methods. Another possible approach would be an integration of the proposed methods and deep unfolding methods [147].
- Many problems in signal processing can be cast as the online estimation of sparse signals. For instance, a sparsity-aware adaptive filtering method based on the MC penalty has been proposed in [136]. Although this thesis focused on batch methods, extensions to such online methods will expand the range of applications.
- In this thesis, the efficacy of the proposed methods was demonstrated only through numerical examples. We would like to derive an upper bound for the system mismatch, as in [72]. Furthermore, we would like to figure out sufficient conditions for the proposed approaches to cor-

rectly solve the corresponding problems (similar conditions have been derived in other robust signal recovery approaches [148]).

- This thesis focused on the feature grouping for an application of the external division operators. We will further explore the possibility of improvements from existing methods by using the external division operators in other applications. A bottleneck of this is the limitation of the convexity condition of ψ_ω given in Proposition 5.8. Hence, an extension of that proposition would be required.
- This thesis only considered the proximal gradient algorithm (and its generalization) to exploit the external division operator. However, the external division operator may be useful in other popular algorithms such as Douglas-Rachford, Chambolle-Pock, and the primal-dual splitting algorithms. An application of the external division operator to deep neural networks and deep unrolling [149] would also be of interest.

Appendix A

Contours of the Quadratic Loss Function

In this section, we show that the contours of the quadratic loss are ellipses centered at the least square solution, and the axes of the ellipses are tilted 45 degrees from the coordinate axis in the two-dimensional case when each column of \mathbf{A} is standardized. Fix $\mathbf{x} \in \{\boldsymbol{\xi} \in \mathbb{R}^2 \mid \|\mathbf{y} - \mathbf{A}\boldsymbol{\xi}\|_2^2 = c\}$ for some constant $c > 0$. Assume that $\mathbf{A}^\top \mathbf{A}$ is invertible. Then, it holds that

$$\begin{aligned} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 &= \mathbf{y}^\top \mathbf{y} - 2\mathbf{y}^\top \mathbf{A}\mathbf{x} + \mathbf{x}^\top \mathbf{A}^\top \mathbf{A}\mathbf{x} \\ &= \left(\mathbf{x} - (\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{A}^\top \mathbf{y} \right)^\top \mathbf{A}^\top \mathbf{A} \left(\mathbf{x} - (\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{A}^\top \mathbf{y} \right) \\ &\quad + \underbrace{\mathbf{y}^\top \left(\mathbf{I}_n - \mathbf{A} (\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{A}^\top \right) \mathbf{y}}_{\text{independent of } \mathbf{x}}. \end{aligned} \quad (\text{A.1})$$

Let $\tilde{c} := c - \mathbf{y}^\top \left(\mathbf{I}_n - \mathbf{A} (\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{A}^\top \right) \mathbf{y}$. Since \tilde{c} is independent of \mathbf{x} , (A.1) is equivalent to

$$\left(\mathbf{x} - (\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{A}^\top \mathbf{y} \right)^\top \mathbf{A}^\top \mathbf{A} \left(\mathbf{x} - (\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{A}^\top \mathbf{y} \right) = \tilde{c}, \quad (\text{A.2})$$

which is the equation of an ellipse, of which the center is the least square solution $(\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{A}^\top \mathbf{y}$. Since each column of \mathbf{A} is standardized, it holds that

$$\begin{aligned} \mathbf{A}^\top \mathbf{A} &= \begin{bmatrix} 1 & \mathbf{a}_1^\top \mathbf{a}_2 \\ \mathbf{a}_1^\top \mathbf{a}_2 & 1 \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} 1 + |\mathbf{a}_1^\top \mathbf{a}_2| & 0 \\ 0 & 1 - |\mathbf{a}_1^\top \mathbf{a}_2| \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix} \end{aligned} \quad (\text{A.3})$$

Hence, the major and minor axes of the ellipse are along $\left[\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}\right]^T$ and $\left[\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right]^T$, respectively.

Appendix B

Weak Convexities of Some Nonconvex Penalties

We show below that (i) the ℓ_q quasi-norm ($0 < q < 1$) defined in (2.53) and (ii) the capped ℓ_1 norm defined in (2.55) are not weakly convex, and (iii) the CEL0 penalty defined in (2.54) is γ^2 -weakly convex.

- (i) The ℓ_q quasi-norm $\phi^{\ell_q} : \mathbb{R} \rightarrow \mathbb{R} : x \mapsto (1/q)|x|^q$ is differentiable at $x \in \mathbb{R} \setminus \{0\}$. The second derivative at $x > 0$ is given by

$$\begin{aligned} (\phi^{\ell_q})''(x) &= (q-1)|x|^{q-2} \\ &< 0. \end{aligned} \tag{B.1}$$

Since

$$\lim_{x \rightarrow +0} (q-1)|x|^{q-2} = -\infty, \tag{B.2}$$

there does not exist any constant $\rho > 0$ such that

$$(q-1)|x|^{q-2} + \rho > 0, \quad \forall x > 0. \tag{B.3}$$

Since there exists $x > 0$ at which the Hessian of $\phi^{\ell_q} + (\rho/2)x^2$ is negative, $\phi^{\ell_q} + (\rho/2)x^2$ is nonconvex for any $\rho > 0$. Thus, the ℓ_q quasi-norm is not weakly convex.

- (ii) Let $h : \mathbb{R} \rightarrow \mathbb{R} : x \mapsto \phi_\gamma^{\text{cap}}(x) + (\rho/2)x^2$ for some $\rho > 0$. For any $x, \xi \in \mathbb{R}$, it holds that

$$\begin{aligned} \frac{\rho}{2} \left(\frac{1}{2}x^2 + \frac{1}{2}\xi^2 - \left(\frac{x+\xi}{2} \right)^2 \right) &= \frac{\rho}{2} \left(\frac{1}{2}x^2 + \frac{1}{2}\xi^2 - x\xi \right) \\ &= \frac{\rho}{4}(x-\xi)^2. \end{aligned} \tag{B.4}$$

Hence, it holds for any $x \in (0, \gamma)$ and $\xi > \gamma$ that

$$\begin{aligned}
& \frac{1}{2}h(x) + \frac{1}{2}h(\xi) - h\left(\frac{x+\xi}{2}\right) \\
&= \frac{1}{2}x + (\rho/2)x^2 + \frac{1}{2}\gamma + (\rho/2)\xi^2 - \min\left\{\frac{x+\xi}{2}, \gamma\right\} - (\rho/2)\left(\frac{x+\xi}{2}\right)^2 \\
&= \frac{1}{2}x + \frac{1}{2}\gamma - \min\left\{\frac{x+\xi}{2}, \gamma\right\} + \frac{\rho}{4}(x-\xi)^2. \tag{B.5}
\end{aligned}$$

Substituting $x := \gamma - \epsilon$ and $\xi := \gamma + \epsilon$ for sufficiently small $\epsilon > 0$ to (B.5) yields

$$\begin{aligned}
\frac{1}{2}h(x) + \frac{1}{2}h(\xi) - h\left(\frac{x+\xi}{2}\right) &= \frac{1}{2}(\gamma - \epsilon) - \frac{1}{2}\gamma + \rho\epsilon^2 \\
&= \rho\epsilon^2 - \frac{1}{2}\epsilon, \tag{B.6}
\end{aligned}$$

which is negative for $\epsilon \in (0, 1/(2\rho))$. This implies that there exist $x, \xi \in \mathbb{R}$ for any $\rho > 0$ such that (B.6) is negative, and hence h is nonconvex for any $\rho > 0$. Thus, ϕ^{cap} is not weakly convex.

(iii) By definition in (2.54), it holds for any $x \leq \sqrt{2\mu}/\gamma$ that

$$\begin{aligned}
\phi_{\gamma, \mu}^{\text{CEL0}}(x) &= \mu - \frac{\gamma^2}{2}x^2 + \frac{\gamma^2}{2} \frac{2\sqrt{2\mu}}{\gamma} |x| - \frac{\gamma^2}{2} \frac{2\mu}{\gamma^2} \\
&= -\frac{\gamma^2}{2}x^2 + \gamma\sqrt{2\mu}|x| \\
&= \gamma\sqrt{2\mu} \left(|x| - \frac{\gamma}{2\sqrt{2\mu}}x^2 \right). \tag{B.7}
\end{aligned}$$

On the other hand, it holds for any $x > \sqrt{2\mu}/\gamma$ that

$$\phi_{\gamma, \mu}^{\text{CEL0}}(x) = \mu = \gamma\sqrt{2\mu} \frac{\sqrt{2\mu}}{2\gamma}. \tag{B.8}$$

Hence, (B.7) and (B.8) together with (2.45) yield that

$$\phi_{\gamma, \mu}^{\text{CEL0}}(x) = \gamma\sqrt{2\mu} \phi_{\sqrt{2\mu}/\gamma}^{\text{MC}}. \tag{B.9}$$

Since $\phi_{\sqrt{2\mu}/\gamma}^{\text{MC}}$ is $(\gamma/\sqrt{2\mu})$ -weakly convex, $\phi_{\gamma, \mu}^{\text{CEL0}}$ is γ^2 -weakly convex.

Appendix C

Forward-Backward-Based Primal-Dual Splitting Method

We consider the following optimization problem:

$$\min_{\mathbf{x} \in \mathbb{R}^n} F(\mathbf{x}) + G(\mathbf{B}\mathbf{x}), \quad (\text{C.1})$$

where $F \in \Gamma_0(\mathbb{R}^n)$ is differentiable on \mathbb{R}^n with its gradient Lipschitz continuous with constant $\beta > 0$, $G \in \Gamma_0(\mathbb{R}^m)$, and $\mathbf{B} \in \mathcal{B}(\mathbb{R}^n, \mathbb{R}^m)$. The set of minimizers of (C.1) is assumed nonempty. There are several algorithms to solve (C.1) based on the forward-backward approach: they combine a gradient decent step (forward step) with a computation step involving a proximity operator (backward step) [118, 150, 116, 151] (see [131] for a survey paper). This thesis utilizes the forward-backward based primal-dual splitting method proposed in [118], which is shown in Algorithm C.1. It is known that the conditions on the step sizes for the algorithm of [118] are less restrictive than that of [150]. The convergence conditions of Algorithm C.1 are given in the following fact.

Fact C.1 ([118, Theorem 3.5] and [131]). Let $(\mathbf{x}_k)_{k \in \mathbb{N}}$ and $(\mathbf{v}_k)_{k \in \mathbb{N}}$ be the sequences generated by Algorithm C.1, respectively. Suppose that

$$0 < \tau < \frac{2}{\beta}, \quad 0 < \varsigma \leq \frac{1}{\lambda_{\max}(\mathbf{B}^\top \mathbf{B})}. \quad (\text{C.2})$$

Then, the sequence $(\mathbf{x}_k)_{k \in \mathbb{N}}$ converges to a solution of (C.1), and $((\tau/\varsigma)\mathbf{v}_k)_{k \in \mathbb{N}}$ converges to a solution of the associated dual problem

$$\min_{\mathbf{y} \in \mathbb{R}^m} F^*(\mathbf{B}^\top \mathbf{y}) + G^*(\mathbf{y}). \quad (\text{C.3})$$

Algorithm C.1 : Forward-backward-based primal-dual splitting algorithm [118]

Set $(\mathbf{x}_0, \mathbf{v}_0) \in \mathbb{R}^n \times \mathbb{R}^m$, and $\tau, \varsigma > 0$

For $k = 0, 1, 2, \dots$

$$\mathbf{x}_{k+\frac{1}{2}} := \mathbf{x}_k - \tau \nabla F(\mathbf{x}_k)$$

$$\mathbf{v}_{k+1} := (\text{Id} - \text{Prox}_{(\tau/\varsigma)G})(\mathbf{B}\mathbf{x}_{k+\frac{1}{2}} + \varsigma \mathbf{B}\mathbf{B}^\top)\mathbf{v}_k$$

$$\mathbf{x}_{k+1} := \mathbf{x}_{k+\frac{1}{2}} - \varsigma \mathbf{B}^\top \mathbf{v}_{k+1}$$

Appendix D

Big Oh Notation

We define the big Oh notation (a.k.a. Landau's symbol) to express the computational cost for algorithms.

Definition 19 (Eventually positive function). A function¹ $f : \mathbb{N}^n \rightarrow \mathbb{R}$ is eventually positive if there is a constant $N \in \mathbb{N}$ such that $f(l_1, l_2, \dots, l_n)$ is positive for all $l_1, l_2, \dots, l_n \geq N$.

Definition 20 (Big Oh notation [152]). Let $g : \mathbb{N}^n \rightarrow \mathbb{R}$ be eventually positive. Then, the set $\mathcal{O}(g)$ is the set of all eventually positive functions $f : \mathbb{N} \rightarrow \mathbb{R}$ for which there exist $N, m \in \mathbb{N}$ such that

$$f(l_1, l_2, \dots, l_n) \leq mg(l_1, l_2, \dots, l_n) \quad (\text{D.1})$$

for all $l_1, l_2, \dots, l_n \geq N$.

In the literature, the big Oh notation is slightly abused as $f = \mathcal{O}(g)$ to express $f \in \mathcal{O}(g)$. Moreover, the notion $n^3 \in \mathcal{O}(n^3)$ or $n^3 = \mathcal{O}(n^3)$ is also used to express $f \in \mathcal{O}(g)$ for $f, g : \mathbb{N} \rightarrow \mathbb{N} : n \mapsto n^3$. For example, if $f : \mathbb{N} \rightarrow \mathbb{N} : n \mapsto 3n^2 + 4n + 1$, then $f(n) \in \mathcal{O}(n^2)$ since it holds for any $n \geq 3$ that

$$\begin{aligned} 4n^2 - (3n^2 + 4n + 1) &= n^2 - 4n - 1 \\ &= (n - 2)^2 - 5 \\ &\geq 0. \end{aligned} \quad (\text{D.2})$$

Note that $f(n) \in \mathcal{O}(n^3)$ and $f(n) \notin \mathcal{O}(n)$ for this case. In general, $f \in \mathcal{O}(g)$ does not imply that $g(n)$ is as accurate to $f(n)$ as possible. For any $f, g : \mathbb{N}^n \rightarrow \mathbb{R}$, the following hold:

1. $c\mathcal{O}(f) = \mathcal{O}(f)$ for any $c > 0$,
2. $\mathcal{O}(f + g) = \mathcal{O}(\max\{f, g\})$, where max is the pointwise maximum.

¹Eventually positive function can be defined for partial functions, i.e., $f : C \rightarrow \mathbb{R}$ for some $C \subset \mathbb{N}^n$.

Appendix E

Matrix Computation Methods

E.1 Power Methods

Assume that there exist $\mathbf{V} \in \mathbb{R}^{n \times n}$ and $\lambda_1 > \lambda_2 \geq \dots \geq \lambda_n$ for a given matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ such that

$$\mathbf{V}^{-1}\mathbf{A}\mathbf{V} = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n). \quad (\text{E.1.1})$$

The *power method* [153] shown in Algorithm E.1.1 is used to estimate the largest eigenvalue of \mathbf{A} and a corresponding eigenvector. The following fact shows the convergence properties of the power method for the case when \mathbf{A} is symmetric¹.

Fact E.1.1 ([153, Theorem 8.2.1]). Suppose that $\mathbf{A} \in \mathbb{R}^{n \times n}$ is symmetric and that

$$\mathbf{V}^T\mathbf{A}\mathbf{V} = \text{diag}(\lambda_1, \dots, \lambda_n), \quad (\text{E.1.2})$$

where $\mathbf{V} \in \mathbb{R}^{n \times n}$ is orthogonal and $\lambda_1 > \lambda_2 \geq \dots \geq \lambda_n$. Let the vectors $(\mathbf{q}^{(k)})_{k=1}^{\infty}$ be the sequence produced by Algorithm E.1.1, and define $\theta_k \in [0, \pi/2]$ by

$$\cos(\theta_k) = |\mathbf{v}_1^T \mathbf{q}_k|, \quad k = 0, 1, 2, \dots \quad (\text{E.1.3})$$

If $\cos(\theta_0) \neq 0$, then for $k = 0, 1, \dots$, we have

$$|\sin(\theta_k)| \leq \tan(\theta_0) \left| \frac{\lambda_2}{\lambda_1} \right|^k, \quad (\text{E.1.4})$$

and

$$|l_k - \lambda_1| \leq \max_{2 \leq i \leq n} |\lambda_1 - \lambda_i| \tan^2(\theta_0) \left| \frac{\lambda_2}{\lambda_1} \right|^{2k}. \quad (\text{E.1.5})$$

¹Unsymmetric case is also studied [153].

Algorithm E.1.1 : Power method

Set $\mathbf{q}_0 \in \mathbb{R}^n$ such that $\|\mathbf{q}_0\|_2 = 1$

For $k = 0, 1, 2, \dots$

$\mathbf{z}_k := \mathbf{A}\mathbf{q}_{k-1}$

$\mathbf{q}_k := \mathbf{z}_k / \|\mathbf{z}_k\|_2$

$l_k := \mathbf{q}_k^\top \mathbf{A}\mathbf{q}_k$

E.2 QR Decomposition

The QR decomposition (or QR factorization) [153] of any matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ is defined by

$$\mathbf{A} = \mathbf{Q}\mathbf{R}, \quad (\text{E.2.1})$$

where $\mathbf{Q} \in \mathbb{R}^{m \times m}$ is orthogonal and $\mathbf{R} \in \mathbb{R}^{m \times n}$ is upper triangle. There are several methods to obtain \mathbf{Q} and \mathbf{R} such as Householder, block Householder, and Givens transformations [153]. The existence of this factorization is shown by the following theorem.

Fact E.2.1 ([153, Theorem 5.2.1]). If $\mathbf{A} \in \mathbb{R}^{m \times n}$, there exists an orthogonal $\mathbf{Q} \in \mathbb{R}^{m \times m}$ and an upper triangular $\mathbf{R} \in \mathbb{R}^{m \times n}$ so that $\mathbf{A} = \mathbf{Q}\mathbf{R}$.

Appendix F

LiMES Model

F.1 Stable Outlier-Robust Regression

Let us consider when the observation vector is contaminated by sparse outliers as follows:

$$\mathbf{y} := \mathbf{A}\mathbf{x}_\star + \boldsymbol{\varepsilon}_\star + \mathbf{o}_\diamond, \quad (\text{F.1.1})$$

where \mathbf{x}_\star and $\boldsymbol{\varepsilon}$ are the coefficient and noise vectors, which follows i.i.d. zero-mean Gaussian distributions with variances $\sigma_{x_\star}^2$ and $\sigma_{\varepsilon_\star}^2$. As stated in Section 1.1.4, existing outlier-robust regression methods have two major issues: (i) existence of the tradeoff problem between robustness and global optimality, and (ii) no distinction of the statistical differences between noise and outliers.

The above two issues have been resolved by SORR [35]. The formulation of SORR is based on the MC loss function¹:

$$\min_{\mathbf{x} \in \mathbb{R}^n, \boldsymbol{\varepsilon} \in \mathbb{R}^m} \mu_{\text{SORR}} \Phi_\gamma^{\text{MC}}(\mathbf{y} - (\mathbf{A}\mathbf{x} + \boldsymbol{\varepsilon})) + \frac{1}{2} \|\mathbf{x}\|_2^2 + \frac{\rho_{\text{SORR}}}{2} \|\boldsymbol{\varepsilon}\|_2^2, \quad (\text{F.1.3})$$

where $\gamma, \mu_{\text{SORR}}, \rho_{\text{SORR}} > 0$. The MC loss induces the sparsity of the residual $\mathbf{y} - (\mathbf{A}\mathbf{x} + \boldsymbol{\varepsilon})$, which can be regarded as an estimate of the pure outliers with the Gaussian noise eliminated. In addition, the influence of huge outliers is completely annihilated since the derivative of the MC function vanishes at a certain level. By introducing the auxiliary vector $\boldsymbol{\varepsilon}$, which models the Gaussian noise, the statistical difference between noise and outliers is explicitly distinguished. When the noise power $\sigma_{\varepsilon_\star}^2$ is large, $\|\boldsymbol{\varepsilon}_\star\|_2^2$ becomes large as well. In this case, the inverse $\sigma_{\varepsilon_\star}^{-2}$ of the noise-power estimate would

¹The original formulation of SORR in [35] is

$$\min_{\mathbf{x} \in \mathbb{R}^n, \boldsymbol{\varepsilon} \in \mathbb{R}^m} \mu \Phi_\gamma^{\text{MC}}(\mathbf{y} - (\mathbf{A}\mathbf{x} + \boldsymbol{\varepsilon})) + \frac{\sigma_{x_\star}^{-2}}{2} \|\mathbf{x}\|_2^2 + \frac{\sigma_{\varepsilon_\star}^{-2}}{2} \|\boldsymbol{\varepsilon}\|_2^2, \quad (\text{F.1.2})$$

where $\sigma_{x_\star}^2 > 0$ and $\sigma_{\varepsilon_\star}^2 > 0$ are estimates of the variances of \mathbf{x} and $\boldsymbol{\varepsilon}$, respectively. Multiplying the cost function of (F.1.2) by $\sigma_{x_\star}^2$ and letting $\mu_{\text{SORR}} := \sigma_{x_\star}^2 \mu$ and $\rho_{\text{SORR}} := \sigma_{x_\star}^2 / \sigma_{\varepsilon_\star}^2$ yield the cost function of (F.1.3).

be small, and it allows $\|\varepsilon\|_2^2$ to be large so that ε mimics ε_* well, yielding efficient mitigation of the MC loss $\Phi_\gamma^{\text{MC}}(\mathbf{y} - (\mathbf{A}\mathbf{x} + \varepsilon))$. This leads to the “stability”, which means that the residuals are bounded to be proportional to the magnitude of noise [154]. Therefore, SORR achieves remarkable robustness against outliers and stability even in severely noisy environments while the overall convexity of the whole cost function is guaranteed for appropriate choices of μ_{SORR} and γ .

The following fact shows the convexity conditions of (F.1.3).

Fact F.1.1 ([35, Proposition 3]). The smooth part

$$\frac{\sigma_{x_*}^{-2}}{2}\|\mathbf{x}\|_2^2 + \frac{\sigma_{\varepsilon_*}^{-2}}{2}\|\varepsilon\|_2^2 - \mu^\gamma \|\cdot\|_1(\mathbf{y} - (\mathbf{A}\mathbf{x} + \varepsilon)) \quad (\text{F.1.4})$$

is convex in $(\mathbf{x}, \varepsilon) \in \mathbb{R}^n \times \mathbb{R}^m$ if and only if

$$\mu(\sigma_{\varepsilon_*}^2 + \sigma_{x_*}^2 \lambda_{\max}(A^\top A)) \leq \gamma. \quad (\text{F.1.5})$$

The formulation for SORR in (F.1.3) is a special case of the LiMES model to be presented in the following section.

F.2 LiMES Model: Convexity Conditions

Let \mathcal{X} , \mathcal{Y} , and \mathcal{Z} be finite-dimensional Hilbert spaces. Let $\mathcal{A}_1 : \mathcal{X} \rightarrow \mathcal{Y} : x \mapsto M_1x + c_1$ and $\mu > 0$, where $(0 \neq)M_1 : \mathcal{X} \rightarrow \mathcal{Y}$ is a bounded linear operator and $c_1 \in \mathcal{Y}$. Let $(0 \neq)\mathcal{L} : \mathcal{Z} \rightarrow \mathcal{Z}$ be a bounded linear operator, $D \succ 0 : \mathcal{Z} \rightarrow \mathcal{Z}$ be a diagonal positive-definite operator, and $\mathcal{A}_2 : \mathcal{X} \rightarrow \mathcal{Z} : x \mapsto M_2x + c_2$, where $(0 \neq)M_2 : \mathcal{X} \rightarrow \mathcal{Z}$ is a bounded linear operator and $c_2 \in \mathcal{Z}$. Let $\Psi \in \Gamma_0(\mathcal{Z})$, which is referred to as a seed function. The LiMES model [35] is defined as the minimization of the following function:

$$J_{\Psi_D^{\mathcal{L}} \circ \mathcal{A}_2}^{\mathcal{A}_1} : \mathcal{X} \rightarrow (-\infty, +\infty] : x \mapsto \frac{1}{2}\|\mathcal{A}_1x\|_{\mathcal{Y}}^2 + \mu\Psi_D^{\mathcal{L}}(\mathcal{A}_2x), \quad (\text{F.2.1})$$

where $\Psi_D^{\mathcal{L}} \circ \mathcal{A}_2 : \mathcal{X} \rightarrow (-\infty, +\infty]$ is the LiMES function with

$$\begin{aligned} &\Psi_D^{\mathcal{L}} : \mathcal{Z} \rightarrow (-\infty, +\infty] \\ &: z \mapsto \Psi(z) - \min_{v \in \mathcal{Z}} \left(\Psi(v) + \frac{1}{2}\|D(\mathcal{L}z - v)\|_{\mathcal{Z}}^2 \right). \end{aligned} \quad (\text{F.2.2})$$

Due to Moreau’s decomposition

$$\gamma f + \gamma^{-1}(f^*) \circ \gamma^{-1} \text{Id} = \frac{1}{2\gamma} \|\cdot\|_{\mathcal{Z}}^2, \quad (\text{F.2.3})$$

for any $f \in \Gamma_0(\mathcal{Z})$ and $\gamma > 0$ [96], the following smooth-nonsmooth separation can be verified under the nonsingularity of D :

$$\begin{aligned}
J_{\Psi \circ D}^{\mathcal{A}_1 \circ \mathcal{A}_2} &= \frac{1}{2} \|\mathcal{A}_1 \cdot\|_{\mathcal{Y}}^2 - \mu^1 (\Psi \circ D^{-1}) \circ D \mathcal{L} \mathcal{A}_2 + \mu \Psi \circ \mathcal{A}_2 \\
&= \underbrace{\frac{1}{2} \|\mathcal{A}_1 \cdot\|_{\mathcal{Y}}^2 - \frac{\mu}{2} \|D \mathcal{L} \mathcal{A}_2 \cdot\|_{\mathcal{Z}}^2}_{=: F \text{ (smooth)}} + \underbrace{\mu^1 (\Psi^* \circ D) \circ D \mathcal{L} \mathcal{A}_2 + \mu \Psi \circ \mathcal{A}_2}_{\text{(nonsmooth)}}.
\end{aligned} \tag{F.2.4}$$

Since the third term of F is convex by the convexity of the Moreau envelope [96], F is convex if the sum of the first two terms is convex. A result regarding the convexity condition of the smooth part F is given as follows.

Fact F.2.1 (Convexity condition for the smooth part [35, Proposition 5]).

(a) $F \in \Gamma_0(\mathcal{X})$ if

$$(\spadesuit) \quad M_1^* M_1 - \mu M_2^* \mathcal{L}^* D^2 \mathcal{L} M_2 \succeq 0.$$

(b) Let $\Psi := \sigma_C$ with a nonempty closed convex set $C \subset \mathcal{Z}$. Then, the following statements hold.

(i) Given any $x \in \mathcal{X}$, the following equivalence holds:

$$\begin{aligned}
F(x) &= \frac{1}{2} \|\mathcal{A}_1 x\|_{\mathcal{Y}}^2 - \frac{\mu}{2} \|D \mathcal{L} \mathcal{A}_2 x\|_{\mathcal{Z}}^2 \\
&\Leftrightarrow \quad {}^1 (\sigma_C \circ D^{-1}) (D \mathcal{L} \mathcal{A}_2 x) = \frac{1}{2} \|D \mathcal{L} \mathcal{A}_2 x\|_{\mathcal{Z}}^2 \\
&\Leftrightarrow \quad {}^1 (\sigma_C^* \circ D) (D \mathcal{L} \mathcal{A}_2 x) = 0 \\
&\Leftrightarrow \quad x \in K_C := \{x \in \mathcal{X} \mid D^2 \mathcal{L} \mathcal{A}_2 x \in C\}.
\end{aligned} \tag{F.2.5}$$

(ii) Assume that $\text{int } K_C \neq \emptyset$. Then, $F \in \Gamma_0(\mathcal{X})$ if and only if (\spadesuit) is satisfied.

Appendix G

Monotone Lipschitz-Gradient Denoiser

Let us consider the following iterate:

$$x_{k+1} := T(x_k - \mu \nabla f(x_k)), \quad k \in \mathbb{N}, \quad (\text{G.1})$$

where T is a nonlinear mapping (denoiser) from a real Hilbert space \mathcal{H} to \mathcal{H} , $f : \mathcal{H} \rightarrow \mathbb{R}$ is a smooth convex function, and $\mu > 0$ is the step size. In particular T is Moreau's proximity operator of a proper lower semicontinuous convex function g , (G.1) is the classical forward-backward splitting algorithm to minimize $\mu f + g$. In [139], weak convergence of (G.1) is studied when T is Monotone Lipschitz-gradient (MoL-Grad) denoiser, which is defined as follows:

Condition \diamond (MoL-Grad Denoiser) Denoiser $T : \mathcal{H} \rightarrow \mathcal{H}$ is a β^{-1} -Lipschitz continuous operator for $\beta \in (0, 1)$ such that $T = \nabla \psi$ for a Fréchet differentiable convex function ψ . In other words, T can be expressed as the gradient of a β^{-1} -smooth convex function.

Necessary and sufficient conditions for T to be a MoL-Grad denoiser is shown in the following fact.

Fact G.1 ([139, Theorem 1]). Let $T : \mathcal{H} \rightarrow \mathcal{H}$ with $\text{dom } T = \mathcal{H}$. Then, for $\beta \in (0, 1)$, the following statements are equivalent.

(C1) $T = \text{s-Prox}_\varphi$ for some $\varphi : \mathcal{H} \rightarrow (-\infty, +\infty]$ such that $\varphi + ((1 - \beta)/2) \|\cdot\|_2^2 \in \Gamma_0(\mathcal{H})$.

(C2) T satisfies condition \diamond , *i.e.*, the following hold jointly:

- 1) $T = \nabla \psi$ for some Fréchet differentiable convex function $\psi \in \Gamma_0(\mathcal{H})$.
- 2) T is β -cocoercive, or equivalently β^{-1} -Lipschitz continuous by the Baillon-Haddad theorem (see Fact 2.2).

In particular, the following statements hold.

- (a) Assume that (C1) is satisfied. Define $\check{\varphi} := \varphi + ((1-\beta)/2)\|\cdot\|^2 \in \Gamma_0(\mathcal{H})$. Then, it holds that

$$T = \text{s-Prox } \varphi = \nabla \underbrace{\left(\varphi + \frac{1}{2}\|\cdot\|^2 \right)}_{=\psi}^* = \nabla^\beta(\check{\varphi}^*), \quad (\text{G.2})$$

which is β -cocoercive (thus β^{-1} -Lipschitz continuous and maximally monotone).

- (b) Assume that (C2) is satisfied. Then, it holds that where

$$\varphi = \psi^* - \frac{1}{2}\|\cdot\|^2 \quad (\text{G.3})$$

is $(1-\beta)$ -weakly convex.

The following fact shows the conditions for the iterate (G.1) with a MoL-Grad denoiser T to converge weakly to a solution of $\mu f + g$.

Fact G.2 ([139, Theorem 2]). Let $f \in \Gamma_0(\mathcal{H})$ be a κ -smooth ρ -strongly-convex function with $\kappa > \rho > 0$. Assume that $T : \mathcal{H} \rightarrow \mathcal{H}$ satisfies condition \diamond for $\beta \in ((\kappa - \rho)/(\kappa + \rho), 1) \subsetneq (0, 1)$ so that the Lipschitz constant is bounded by $\beta^{-1} < (\kappa + \rho)/(\kappa - \rho)$. Set $\mu \in [(1 - \beta)/\rho, (1 + \beta)/\kappa]$. Then, the following hold.

- 1) Let $\hat{f} := f - [(1-\beta)/(2\mu)]\|\cdot\|^2 \in \Gamma_0(\mathcal{H})$ and $\check{\varphi} := \varphi + [(1-\beta)/2]\|\cdot\|^2 \in \Gamma_0(\mathcal{H})$ so that $\mu\hat{f} + \check{\varphi} = \mu f + \varphi$. Then, it holds that

$$T \circ (\text{Id} - \mu\nabla f) = \text{s-Prox } \beta^{-1}\check{\varphi} \circ (\text{Id} - \beta^{-1}\mu\nabla\hat{f}) \quad (\text{G.4})$$

with $\beta^{-1} \in (1, 2/L_{\mu\nabla\hat{f}})$, where $L_{\mu\nabla\hat{f}} := \mu\kappa - (1 - \beta) > 0$ is the Lipschitz constant of $\mu\nabla\hat{f}$.

- 2) Suppose that $\mu f + \varphi$ has a minimizer in \mathcal{H} . Then, for an arbitrary $x_0 \in \mathcal{H}$, the sequence $(x_k)_{k \in \mathbb{N}} \subset \mathcal{H}$ generated by (G.1) converges weakly to a minimizer \hat{x} of $\mu f + \varphi$. (In this case, φ is the implicit regularizer.)

Appendix H

Decomposition of Proximity Operator

In this section, we review some useful sufficient conditions for the following equality to hold:

$$\text{Prox}_{h_1} \circ \text{Prox}_{h_2} = \text{Prox}_{h_1+h_2}, \quad (\text{H.1})$$

where $h_1, h_2 \in \Gamma_0(\mathbb{R}^n)$ such that $h_1 + h_2$ is proper. The assumption based on (H.1) for certain functions is essential in the main results of this study.

Proposition H.1. [144, Proposition 3.13] Let $h_1, h_2 \in \Gamma_0(\mathbb{R}^n)$ such that $\text{dom } h_1 \cap \text{dom } h_2 \neq \emptyset$. Suppose that one of the following holds:

- (i). $(\forall \mathbf{x} \in \text{dom } \partial h_2) \partial h_2(\mathbf{x}) \subset \partial h_2(\text{Prox}_{h_1} \mathbf{x})$
- (ii). $(\forall (\mathbf{x}, \mathbf{u}) \in \text{gra } \partial h_1) \partial h_2(\mathbf{x} + \mathbf{u}) \subset \partial h_2(\mathbf{x})$.
- (iii). $0 \in \text{int}(\text{dom } h_1 - \text{dom } h_2)$ and $(\forall (\mathbf{x}, \mathbf{u}) \in \text{gra } \partial h_1) \partial h_2(\mathbf{x}) \subset \partial h_2(\mathbf{x} + \mathbf{u})$.

Then, (H.1) holds.

The condition $0 \in \text{int}(\text{dom } h_1 - \text{dom } h_2)$ in (iii) can be generalized to $0 \in \text{sri}(\text{dom } h_1 - \text{dom } h_2)$, where $\text{sri } C$ denotes the strong relative interior of a subset $C \subset \mathbb{R}^n$.

Proposition H.2. [144, Proposition 3.13] Let $h_1, h_2 \in \Gamma_0(\mathbb{R})$ such that $\text{dom } h_1 \cap \text{dom } h_2 \neq \emptyset$. Then, $\text{Prox}_{h_1} \circ \text{Prox}_{h_2}$ is the proximity operator of some function in $\Gamma_0(\mathbb{R})$.

Proposition H.3. [122, Proposition II.2] Let $h_1, h_2 \in \Gamma_0(\mathbb{R}^n)$. Assume the following:

1. h_1 is separable, *i.e.*, for any $\mathbf{x} \in \mathbb{R}^n$, $h_1(\mathbf{x}) = \sum_{i=1}^n h_0(x_i)$ for some function $h_0 \in \Gamma_0(\mathbb{R})$.

2. For any $\mathbf{x} \in \mathbb{R}^n$,

$$h_2(\mathbf{x}) = \sum_{i=1}^n \sum_{j=1}^n \sigma_{[a_{i,j}, b_{i,j}]}(x_i - x_j) \quad (\text{H.2})$$

for some $a_{i,j} \in \mathbb{R} \cup \{-\infty\}$ and $b_{i,j} \in \mathbb{R} \cup \{+\infty\}$ with $a_{i,j} \leq b_{i,j}$ for any $i, j = 1, 2, \dots, n$.

Then, (H.1) holds.

The following decomposition is studied in the context of proximal thresholds, which is defined as $T : \mathbb{R}^n \rightarrow \mathbb{R}^n$ on a nonempty closed convex subset $\Omega \subset \mathbb{R}^n$ if there exists a function $f \in \Gamma_0(\mathbb{R}^n)$ such that $T = \text{Prox}_f$ and $\{z \in \mathbb{R}^n \mid Tz = 0\} = \Omega$.

Proposition H.4 ([96, Proposition 24.54]). Let Ω be a nonempty closed interval in \mathbb{R} , let $h_1 \in \Gamma_0(\mathbb{R})$ be differentiable at 0 and such that $h_1'(0) = 0$, and set $h_2 := \sigma_\Omega$. Then, (H.1) holds.

Appendix I

Proof for Chapter 3

I.1 Proof of (3.11)

The following lemma is used in the proof of (3.11).

Lemma I.1.1. Let $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ and $((\mathcal{K}_i, \langle \cdot, \cdot \rangle_{\mathcal{K}_i}))_{i \in I}$ be finite-dimensional real Hilbert spaces for $I := \{1, 2, \dots, d\}$ with some $d \in \mathbb{N} \setminus \{0\}$. For every $i \in I$, suppose that $L_i \in \mathcal{B}(\mathcal{H}, \mathcal{K}_i) \setminus \{0\}$, $B_i \in \mathcal{K}_i$, and let $\beta > 0$ and $T_i : \mathcal{K}_i \rightarrow \mathcal{K}_i$ be β_i -cocoercive. Set

$$T := \sum_{i \in I} L_i^* \circ T_i \circ (B_i + L_i(\cdot)) \text{ and } \beta := \sum_{i \in I} \frac{\|L_i\|_{\mathcal{K}_i}^2}{\beta_i}. \quad (\text{I.1.1})$$

Then, T is β^{-1} -cocoercive.

Proof of Lemma I.1.1. For any $\Xi_1, \Xi_2 \in \mathcal{H}$,

$$\begin{aligned} & \langle \Xi_1 - \Xi_2, T\Xi_1 - T\Xi_2 \rangle_{\mathcal{H}} \\ &= \sum_{i \in I} \langle \Xi_1 - \Xi_2, L_i^* \circ T_i(B_i + L_i\Xi_1) - L_i^* \circ T_i(B_i + L_i\Xi_2) \rangle_{\mathcal{H}} \\ &= \sum_{i \in I} \langle B_i + L_i\Xi_1 - (B_i + L_i\Xi_2), T_i(B_i + L_i\Xi_1) - T_i(B_i + L_i\Xi_2) \rangle_{\mathcal{K}_i} \\ &\geq \sum_{i \in I} \beta_i \|T_i(B_i + L_i\Xi_1) - T_i(B_i + L_i\Xi_2)\|_{\mathcal{K}_i}^2 \\ &\geq \beta^{-1} \|T\Xi_1 - T\Xi_2\|_{\mathcal{H}}^2, \end{aligned} \quad (\text{I.1.2})$$

where the last inequality is proven in the same way as the proof of [96, Proposition 4.12]. \square

Proof of (3.11). For notational simplicity, we let $P_{\mathbf{L}} := \text{Prox}_{\|\cdot\|_{2,1} \circ \mathbf{L}^{-1/2}}$ and $P_{\mathbf{M}} := \text{Prox}_{\|\cdot\|_{2,1} \circ \mathbf{M}^{-1/2}}$. From the second equality of (I.3.5) and Facts 2.1

and 2.4(c), we obtain

$$\begin{aligned}
\nabla F(\mathbf{X}) &= \lambda_2 \mathbf{X} - \lambda_1 \mathbf{M} \mathbf{X} + \mathbf{A}^\top \mathbf{L} (\mathbf{Y} - \mathbf{A} \mathbf{X}) - \mathbf{A}^\top \mathbf{L}^{1/2} P_{\mathbf{L}} (\mathbf{L}^{1/2} (\mathbf{Y} - \mathbf{A} \mathbf{X})) \\
&\quad + \lambda_1 \mathbf{M}^{1/2} P_{\mathbf{M}} (\mathbf{M}^{1/2} \mathbf{X}) \\
&= [(\lambda_2 \mathbf{I}_n - \lambda_1 \mathbf{M} - \mathbf{A}^\top \mathbf{L} \mathbf{A}) + \mathbf{A}^\top \mathbf{L}^{1/2} P_{\mathbf{L}} \mathbf{L}^{1/2} \mathbf{A} \\
&\quad + \lambda_1 \mathbf{M}^{1/2} P_{\mathbf{M}} \mathbf{M}^{1/2}] \mathbf{X} + \mathbf{A}^\top \mathbf{L} \mathbf{Y} - \mathbf{A}^\top \mathbf{L}^{1/2} P_{\mathbf{L}} \mathbf{L}^{1/2} \mathbf{Y}. \tag{I.1.3}
\end{aligned}$$

Hence, by Fact 2.3(a) and Lemma I.1.1, we obtain ∇F is β^{-1} -cocoercive (i.e., $\beta^{-1} \nabla F$ is firmly nonexpansive) for

$$\begin{aligned}
\beta &:= \|[\lambda_2 \mathbf{I}_n - \lambda_1 \mathbf{M} - \mathbf{A}^\top \mathbf{L} \mathbf{A}]^{1/2}\|_2^2 + \|\mathbf{L}^{1/2} \mathbf{A}\|_2^2 + \lambda_1 \|\mathbf{M}^{1/2}\|_2^2 \\
&= \lambda_{\max}(\lambda_2 \mathbf{I}_n - \mathbf{A}^\top \text{diag}(l_1, \dots, l_m) \mathbf{A} - \lambda_1 \text{diag}(m_1, \dots, m_n)) \\
&\quad + \lambda_{\max}(\mathbf{A}^\top \text{diag}(l_1, \dots, l_m) \mathbf{A}) + \lambda_1 \max\{m_1, \dots, m_n\}, \tag{I.1.4}
\end{aligned}$$

where the square root $[\lambda_2 \mathbf{I}_n - \lambda_1 \mathbf{M} - \mathbf{A}^\top \mathbf{L} \mathbf{A}]^{1/2}$ exists since $\lambda_2 \mathbf{I}_n - \lambda_1 \mathbf{M} - \mathbf{A}^\top \mathbf{L} \mathbf{A} \succeq \mathbf{0}_{n \times n}$ due to Proposition 3.1 [97]. Since firm nonexpansivity implies nonexpansivity, ∇F is β -Lipschitz continuous. \square

I.2 Proof of Proposition 3.1

Define

$$f : \mathbb{R}^{n \times r} \rightarrow \mathbb{R} : \mathbf{X} \mapsto \frac{\lambda_2}{2} \|\mathbf{X}\|_{\mathbb{F}}^2 - \frac{1}{2} \|\mathbf{A} \mathbf{X}\|_{\mathbf{L}}^2 - \frac{\lambda_1}{2} \|\mathbf{X}\|_{\mathbf{M}}^2. \tag{I.2.1}$$

The following lemma is used in the proof of Proposition 3.1.

Lemma I.2.1. The function f is convex if and only if $f(\mathbf{X}) \geq 0$ for any $\mathbf{X} \in \mathbb{R}^{n \times r}$.

Proof. The function f is convex if and only if, for any $\mathbf{X} \in \mathbb{R}^{n \times r}$, $\mathbf{\Xi} \in \mathbb{R}^{n \times r}$, and any $\alpha \in (0, 1)$,

$$\begin{aligned}
&\frac{\alpha \lambda_2}{2} \|\mathbf{X}\|_{\mathbb{F}}^2 - \frac{\alpha}{2} \|\mathbf{A} \mathbf{X}\|_{\mathbf{L}}^2 - \frac{\alpha \lambda_1}{2} \|\mathbf{X}\|_{\mathbf{M}}^2 + \frac{(1-\alpha) \lambda_2}{2} \|\mathbf{\Xi}\|_{\mathbb{F}}^2 \\
&\quad - \frac{(1-\alpha)}{2} \|\mathbf{A} \mathbf{\Xi}\|_{\mathbf{L}}^2 - \frac{(1-\alpha) \lambda_1}{2} \|\mathbf{\Xi}\|_{\mathbf{M}}^2 - \frac{\lambda_2}{2} \|\alpha \mathbf{X} + (1-\alpha) \mathbf{\Xi}\|_{\mathbb{F}}^2 \\
&\quad + \frac{1}{2} \|\mathbf{A}(\alpha \mathbf{X} + (1-\alpha) \mathbf{\Xi})\|_{\mathbf{L}}^2 + \frac{\lambda_1}{2} \|\alpha \mathbf{X} + (1-\alpha) \mathbf{\Xi}\|_{\mathbf{M}}^2 \geq 0 \\
\Leftrightarrow &\frac{\alpha(1-\alpha) \lambda_2}{2} \|\mathbf{X} - \mathbf{\Xi}\|_{\mathbb{F}}^2 - \frac{\alpha(1-\alpha)}{2} \|\mathbf{A}(\mathbf{X} - \mathbf{\Xi})\|_{\mathbf{L}}^2 \\
&\quad - \frac{\alpha(1-\alpha) \lambda_1}{2} \|\mathbf{X} - \mathbf{\Xi}\|_{\mathbf{M}}^2 \geq 0 \\
\Leftrightarrow &\frac{\lambda_2}{2} \|\mathbf{X} - \mathbf{\Xi}\|_{\mathbb{F}}^2 - \frac{1}{2} \|\mathbf{A}(\mathbf{X} - \mathbf{\Xi})\|_{\mathbf{L}}^2 - \frac{\lambda_1}{2} \|\mathbf{X} - \mathbf{\Xi}\|_{\mathbf{M}}^2 \geq 0, \tag{I.2.2}
\end{aligned}$$

which verifies the claim. \square

Proof of Proposition 3.1. In the light of Lemma I.2.1, the following equivalence can readily be verified:

$$\begin{aligned}
& f \text{ is convex} \\
& \Leftrightarrow f(\mathbf{X}) \geq 0, \forall \mathbf{X} \in \mathbb{R}^{n \times r} \\
& \Leftrightarrow \text{Tr} \left[\mathbf{X}^\top \left(\lambda_2 \mathbf{I} - \mathbf{A}^\top \text{diag}(l_1, \dots, l_m) \mathbf{A} - \lambda_1 \text{diag}(m_1, \dots, m_n) \right) \mathbf{X} \right] \geq 0 \\
& \Leftrightarrow \lambda_2 \mathbf{I} - \mathbf{A}^\top \text{diag}(l_1, \dots, l_m) \mathbf{A} - \lambda_1 \text{diag}(m_1, \dots, m_n) \succeq 0 \\
& \Leftrightarrow \lambda_2 \geq \lambda_{\max} \{ \mathbf{A}^\top \text{diag}(l_1, \dots, l_m) \mathbf{A} + \lambda_1 \text{diag}(m_1, \dots, m_n) \}. \quad (\text{I.2.3})
\end{aligned}$$

The sufficiency has been proven since ${}^1(\iota_C \circ \mathbf{M}^{1/2})(\mathbf{M}^{1/2} \cdot)$ and ${}^1(\iota_C \circ \mathbf{L}^{1/2})(\mathbf{L}^{1/2}(\mathbf{Y} - \cdot \mathbf{A}))$ are convex by Fact 2.4(a). The strict convexity part can be proven in an analogous way. In the rest of the proof, we prove the necessity for the convexity part. Assume that (3.12) does not hold. In this case, by (I.2.3), f is nonconvex. Since the second-order derivative of a quadratic function is constant, f has a negative curvature in some direction everywhere. At any interior point \mathbf{X} of K , $\nabla {}^1(\iota_C \circ \mathbf{M}^{1/2})(\mathbf{M}^{1/2} \mathbf{X}) = \mathbf{0}$ since ${}^1(\iota_C \circ \mathbf{M}^{1/2})(\mathbf{M}^{1/2} \cdot)$ is constant in the neighborhood of the point \mathbf{X} . Hence, the second-order derivative of ${}^1(\iota_C \circ \mathbf{M}^{1/2})(\mathbf{M}^{1/2} \mathbf{X})$ is also zero at any interior point \mathbf{X} of K . In the same way, one can show that the second-order derivative of ${}^1(\iota_C \circ \mathbf{L}^{1/2})(\mathbf{L}^{1/2}(\mathbf{Y} - \mathbf{X} \mathbf{A}))$ is zero. This implies that the last two terms in F make no impact on the curvature in the interior of K , and hence F is nonconvex. \square

I.3 Proof of Proposition 3.3

The following fact and lemma are used in the proof of Proposition 3.3.

Fact I.3.1 ([155, Proposition 2]). It holds for any $\mathbf{X} \in \mathbb{R}^{n \times r}$ and $\mathbf{D} := \text{diag}(d_1, d_2, \dots, d_n) \in \mathbb{R}^{n \times n}$ that

$$\text{Prox}_{\|\cdot\|_{2,1} \circ \mathbf{D}}(\mathbf{X}) = \sum_{i=1}^n \max \left\{ 1 - \frac{d_i}{\|\mathbf{X}_{(i,:)}\|_2}, 0 \right\} \mathbf{e}_{n,i} \mathbf{X}_{(i,:)}^\top. \quad (\text{I.3.1})$$

Lemma I.3.1. It holds for any $\mathbf{X} \in \mathbb{R}^{n \times r}$ and $\mathbf{D} := \text{diag}(d_1, d_2, \dots, d_n) \in \mathbb{R}^{n \times n}$ that

$$\begin{aligned}
\mathbf{D}^{1/2} \text{Prox}_{\iota_C \circ \mathbf{D}^{1/2}}(\mathbf{D}^{1/2} \mathbf{X}) &= \mathbf{D}^{1/2} (\text{Id} - \text{Prox}_{\|\cdot\|_{2,1} \circ \mathbf{D}^{-1/2}})(\mathbf{D}^{1/2} \mathbf{X}) \\
&= \sum_{i=1}^n \min \left\{ \frac{1}{\|\mathbf{X}_{(i,:)}\|_2}, d_i \right\} \mathbf{e}_{n,i} \mathbf{X}_{(i,:)}^\top. \quad (\text{I.3.2})
\end{aligned}$$

Proof. The first equality is due to Facts 2.1 and 2.4(c). The second equality is proven as follows:

$$\begin{aligned}
& \mathbf{D}^{1/2}(\text{Id} - \text{Prox}_{\|\cdot\|_{2,1} \circ \mathbf{D}^{-1/2}})(\mathbf{D}^{1/2} \mathbf{X}) \\
&= \mathbf{D} \mathbf{X} - \mathbf{D}^{1/2} \text{Prox}_{\|\cdot\|_{2,1} \circ \mathbf{D}^{-1/2}}(\mathbf{D}^{1/2} \mathbf{X}) \\
&= \mathbf{D} \mathbf{X} - \mathbf{D}^{1/2} \sum_{i=1}^n \max \left\{ 1 - \frac{d_i^{-1/2}}{\|[\mathbf{D}^{1/2} \mathbf{X}]_{(i,:)}\|_2}, 0 \right\} \mathbf{e}_{n,i} [\mathbf{D}^{1/2} \mathbf{X}]_{(i,:)}^\top \\
&= \mathbf{D} \mathbf{X} - \sum_{i=1}^n \max \left\{ d_i - \frac{1}{\|\mathbf{X}_{(i,:)}\|_2}, 0 \right\} \mathbf{e}_{n,i} \mathbf{X}_{(i,:)}^\top \\
&= \sum_{i=1}^n \min \left\{ \frac{1}{\|\mathbf{X}_{(i,:)}\|_2}, d_i \right\} \mathbf{e}_{n,i} \mathbf{X}_{(i,:)}^\top, \tag{I.3.3}
\end{aligned}$$

where the second equality is due to Fact I.3.1. \square

Proof of Proposition 3.3.1. It holds for any $\mathbf{X} \in \mathbb{R}^{n \times r}$ that

$$\begin{aligned}
F(\mathbf{X}) &= \frac{\lambda_2}{2} \|\mathbf{X}\|_F^2 - \frac{1}{2} \|\mathbf{A} \mathbf{X}\|_L^2 - \frac{\lambda_1}{2} \|\mathbf{X}\|_M^2 + \langle \mathbf{A} \mathbf{X}, \mathbf{Y} \rangle_L \\
&\quad + {}^1(\iota_C \circ \mathbf{L}^{1/2})(\mathbf{L}^{1/2}(\mathbf{Y} - \mathbf{A} \mathbf{X})) \\
&\quad + \lambda_1 {}^1(\iota_C \circ \mathbf{M}^{1/2})(\mathbf{M}^{1/2} \mathbf{X}) \\
&= \frac{\lambda_2}{2} \langle \mathbf{X}, \mathbf{X} \rangle_F - \frac{1}{2} \langle \mathbf{X}, \mathbf{A}^\top \mathbf{L} \mathbf{A} \mathbf{X} \rangle_F - \frac{\lambda_1}{2} \langle \mathbf{M} \mathbf{X}, \mathbf{X} \rangle_F \\
&\quad + \langle \mathbf{X}, \mathbf{A}^\top \mathbf{L} \mathbf{Y} \rangle_F + {}^1(\iota_C \circ \mathbf{L}^{1/2})(\mathbf{L}^{1/2}(\mathbf{Y} - \mathbf{A} \mathbf{X})) \\
&\quad + \lambda_1 {}^1(\iota_C \circ \mathbf{M}^{1/2})(\mathbf{M}^{1/2} \mathbf{X}), \tag{I.3.4}
\end{aligned}$$

from which it follows that

$$\begin{aligned}
\nabla F(\mathbf{X}) &= \lambda_2 \mathbf{X} - \mathbf{A}^\top \mathbf{L} \mathbf{A} \mathbf{X} - \lambda_1 \mathbf{M} \mathbf{X} + \mathbf{A}^\top \mathbf{L} \mathbf{Y} \\
&\quad + (-\mathbf{L}^{1/2} \mathbf{A})^\top (\text{Id} - \text{Prox}_{\iota_C \circ \mathbf{L}^{1/2}})(\mathbf{L}^{1/2}(\mathbf{Y} - \mathbf{A} \mathbf{X})) \\
&\quad + \lambda_1 (\mathbf{M}^{1/2})^\top (\text{Id} - \text{Prox}_{\iota_C \circ \mathbf{M}^{1/2}})(\mathbf{M}^{1/2} \mathbf{X}) \\
&= \lambda_2 \mathbf{X} - \lambda_1 \mathbf{M} \mathbf{X} + \mathbf{A}^\top \mathbf{L}(\mathbf{Y} - \mathbf{A} \mathbf{X}) \\
&\quad - \mathbf{A}^\top \mathbf{L}^{1/2} (\text{Id} - \text{Prox}_{\iota_C \circ \mathbf{L}^{1/2}})(\mathbf{L}^{1/2}(\mathbf{Y} - \mathbf{A} \mathbf{X})) \\
&\quad + \lambda_1 \mathbf{M}^{1/2} (\text{Id} - \text{Prox}_{\iota_C \circ \mathbf{M}^{1/2}})(\mathbf{M}^{1/2} \mathbf{X}) \\
&= \lambda_2 \mathbf{X} + \mathbf{A}^\top \mathbf{L}^{1/2} \text{Prox}_{\iota_C \circ \mathbf{L}^{1/2}}(\mathbf{L}^{1/2}(\mathbf{Y} - \mathbf{A} \mathbf{X})) \\
&\quad - \lambda_1 \mathbf{M}^{1/2} \text{Prox}_{\iota_C \circ \mathbf{M}^{1/2}}(\mathbf{M}^{1/2} \mathbf{X}) \\
&= \lambda_2 \mathbf{X} + \mathbf{A}^\top \sum_{i=1}^m \min \left\{ \frac{1}{\|[\mathbf{Y} - \mathbf{A} \mathbf{X}]_{(i,:)}\|_2}, l_i \right\} \mathbf{e}_{m,i} [\mathbf{Y} - \mathbf{A} \mathbf{X}]_{(i,:)}^\top \\
&\quad - \lambda_1 \sum_{i=1}^n \min \left\{ \frac{1}{\|\mathbf{X}_{(i,:)}\|_2}, m_i \right\} \mathbf{e}_{n,i} \mathbf{X}_{(i,:)}^\top. \tag{I.3.5}
\end{aligned}$$

Here, the first and fourth equalities are verified by Fact 2.4(b) and Lemma I.3.1, respectively. \square

Proof of Proposition 3.3.2. Let $\tilde{\mathbf{Y}} := \begin{bmatrix} \mathbf{Y} \\ \mathbf{0}_{n \times r} \end{bmatrix}$. Using Facts 2.3(a) and 2.4(c), it holds for any matrix $\mathbf{Z} \in \mathbb{R}^{(m+n) \times r}$ that

$$\begin{aligned} \text{Prox}_{(\tau/\varsigma)G}(\mathbf{Z}) &= \text{Prox}_{(\tau/\varsigma)\|\cdot - \tilde{\mathbf{Y}}\|_{2,1}}(\mathbf{Z}) \\ &= \tilde{\mathbf{Y}} + \text{Prox}_{(\tau/\varsigma)\|\cdot\|_{2,1}}(\mathbf{Z} - \tilde{\mathbf{Y}}) \\ &= \tilde{\mathbf{Y}} + \sum_{i=1}^m \max \left\{ 1 - \frac{\tau/\varsigma}{\|[\mathbf{Z} - \mathbf{Y}]_{(i,:)}\|_2}, 0 \right\} \mathbf{e}_{m+n,i} [\mathbf{Z} - \mathbf{Y}]_{(i,:)}^\top \\ &\quad + \sum_{i=1}^n \max \left\{ 1 - \frac{\tau/\varsigma}{\|\mathbf{Z}_{(m+i,:)}\|_2}, 0 \right\} \mathbf{e}_{m+n,m+i} \mathbf{Z}_{(m+i,:)}^\top, \quad (\text{I.3.6}) \end{aligned}$$

where the last equality is due to Fact I.3.1. \square

Appendix J

Proof for Chapter 4

J.1 Proof of (4.15)

For any $\xi_1, \xi_2 \in \mathbb{R}^{n+m}$, it holds from (4.19) that

$$\begin{aligned}
& \|\nabla F(\xi_1) - \nabla F(\xi_2)\|_2 \\
&= \|(\Theta_3^{1/2})^\top \Theta_3^{1/2}(\xi_1 - \xi_2) + (1 - \alpha)(\mathbf{A}\Theta_2)^\top (\mathbf{A}\Theta_2)(\xi_1 - \xi_2) \\
&\quad - \mu_1\gamma_1^{-1}\Theta_1^\top[(\text{Id} - \text{Prox}_{\gamma_1\|\cdot\|_1})(\Theta_1\xi_1 - \mathbf{y}) - (\text{Id} - \text{Prox}_{\gamma_1\|\cdot\|_1})(\Theta_1\xi_2 - \mathbf{y})] \\
&\quad - \mu_2\gamma_2^{-1}\Theta_2^\top[(\text{Id} - \text{Prox}_{\gamma_2\|\cdot\|_1})(\Theta_2\xi_1) - (\text{Id} - \text{Prox}_{\gamma_2\|\cdot\|_1})(\Theta_2\xi_2)]\|_2 \\
&\leq \|(\Theta_3^{1/2})^\top \Theta_3^{1/2}\|_2\|\xi_1 - \xi_2\|_2 + (1 - \alpha)\|\mathbf{A}^\top \mathbf{A}\|_2\|\xi_1 - \xi_2\|_2 \\
&\quad + \mu_1\gamma_1^{-1}\|\Theta_1^\top\|_2\|(\Theta_1\xi_1 - \mathbf{y}) - (\Theta_1\xi_2 - \mathbf{y})\|_2 + \mu_2\gamma_2^{-1}\|\Theta_2^\top\|_2\|\Theta_2(\xi_1 - \xi_2)\|_2,
\end{aligned} \tag{J.1.1}$$

where the first inequality is due to Fact 2.4(b). The right-most side of (J.1.1) is bounded above by

$$\begin{aligned}
& (\lambda_{\max}(\Theta_3) + (1 - \alpha)\lambda_{\max}(\mathbf{A}^\top \mathbf{A}) + \mu_1\gamma_1^{-1}\lambda_{\max}(\Theta_1^\top \Theta_1) \\
&\quad + \mu_2\gamma_2^{-1}\lambda_{\max}(\Theta_2^\top \Theta_2))\|\xi_1 - \xi_2\|_2 \\
&= [\max\{\alpha, \rho\} + (1 - \alpha + \mu_1\gamma_1^{-1})\lambda_{\max}(\mathbf{A}^\top \mathbf{A}) + \mu_1\gamma_1^{-1} + \mu_2\gamma_2^{-1}]\|\xi_1 - \xi_2\|_2,
\end{aligned}$$

where the equality is verified as follows:

$$\begin{aligned}
\lambda_{\max}(\Theta_1^\top \Theta_1) &= \lambda_{\max}(\Theta_1 \Theta_1^\top) &= \lambda_{\max}(\mathbf{A}\mathbf{A}^\top + \mathbf{I}_m) \\
&= \lambda_{\max}(\mathbf{A}\mathbf{A}^\top) + 1 &= \lambda_{\max}(\mathbf{A}^\top \mathbf{A}) + 1.
\end{aligned}$$

J.2 Proof of Proposition 4.1

By applying Proposition 4.6(a) to F in (4.13), a sufficient condition for convexity is given by

$$\begin{aligned} & (\Theta_3^{1/2})^\top \Theta_3^{1/2} + (1 - \alpha)(\mathbf{A}\Theta_2)^\top (\mathbf{A}\Theta_2) - (\mu_1\gamma_1^{-1}\Theta_1^\top \Theta_1 + \mu_2\gamma_2^{-1}\Theta_2^\top \Theta_2) \\ & \succeq \mathbf{0}_{(n+m) \times (n+m)}, \end{aligned} \quad (\text{J.2.1})$$

which can be expressed equivalently as follows:

$$\begin{bmatrix} \tau_1 \mathbf{I}_n + \tau_2 \mathbf{A}^\top \mathbf{A} & -\mathbf{A}^\top \\ -\mathbf{A} & (\mu_1^{-1}\gamma_1\rho - 1)\mathbf{I}_m \end{bmatrix} \succeq \mathbf{0}_{(n+m) \times (n+m)}, \quad (\text{J.2.2})$$

where

$$\tau_1 := \mu_1^{-1}\gamma_1(\alpha - \mu_2\gamma_2^{-1}) \quad (\text{J.2.3})$$

and

$$\tau_2 := (1 - \alpha)\mu_1^{-1}\gamma_1 - 1. \quad (\text{J.2.4})$$

By [97, Theorem 7.7.9], (J.2.2) holds if and only if the following three conditions are satisfied:

- (i) $\tau_1 \mathbf{I}_n + \tau_2 \mathbf{A}^\top \mathbf{A} \succeq \mathbf{0}_{n \times n}$;
- (ii) $(\mu_1^{-1}\gamma_1\rho - 1)\mathbf{I}_m \succeq \mathbf{0}_{m \times m}$ ($\Leftrightarrow \mu_1^{-1}\gamma_1\rho - 1 \geq 0$);
- (iii) $-\mathbf{A}^\top = (\tau_1 \mathbf{I}_n + \tau_2 \mathbf{A}^\top \mathbf{A})^{1/2} \Upsilon ((\mu_1^{-1}\gamma_1\rho - 1)\mathbf{I}_m)^{1/2}$ for some $\Upsilon \in \mathbb{R}^{n \times m}$ with its largest singular value at most one.

If $\mathbf{A} = \mathbf{0}_{m \times n}$, condition (iii) holds trivially with $\Upsilon := \mathbf{0}_{n \times m}$, and conditions (i) and (ii) coincide under $\lambda_{\max}(\mathbf{A}^\top \mathbf{A}) = \lambda_{\min}(\mathbf{A}^\top \mathbf{A}) = 0$ with conditions (K-II) and (K-I), respectively. Assume that $\mathbf{A} \neq \mathbf{0}_{m \times n}$ in the following. Suppose that conditions (i)–(iii) are satisfied. Since $\mathbf{A} = \mathbf{0}_{m \times n}$ if $\tau_1 \mathbf{I}_n + \tau_2 \mathbf{A}^\top \mathbf{A} = \mathbf{0}_{n \times n}$ or $\mu_1^{-1}\gamma_1\rho - 1 = 0$ due to condition (iii), it holds by $\mathbf{A} \neq \mathbf{0}_{m \times n}$ that

$$\tau_1 \mathbf{I}_n + \tau_2 \mathbf{A}^\top \mathbf{A} \succ \mathbf{0}_{n \times n} \Leftrightarrow \tau_1 + \lambda_{\min}(\tau_2 \mathbf{A}^\top \mathbf{A}) > 0 \quad (\text{J.2.5})$$

and

$$\mu_1^{-1}\gamma_1\rho - 1 > 0. \quad (\text{J.2.6})$$

Let $\tau_3 := (\mu_1^{-1}\gamma_1\rho - 1)^{-1/2} \neq 0$. Then, the equality in condition (iii) can be rewritten as

$$\tau_3 \mathbf{A}^\top = (\tau_1 \mathbf{I}_n + \tau_2 \mathbf{A}^\top \mathbf{A})^{1/2} \tilde{\Upsilon}, \quad (\text{J.2.7})$$

where $\tilde{\Upsilon} := -\Upsilon$. Let $\mathbf{A} = \mathbf{V}\Sigma\mathbf{U}^\top$ be a singular value decomposition of \mathbf{A} , where $\mathbf{U} \in \mathbb{R}^{n \times n}$ and $\mathbf{V} \in \mathbb{R}^{m \times m}$ are orthogonal matrices, and the diagonal

entries of $\Sigma \in \mathbb{R}^{m \times n}$ satisfy $\varsigma_1 := \lambda_{\max}^{1/2}(\mathbf{A}^\top \mathbf{A}) \geq \varsigma_2 \geq \dots \geq \varsigma_{\min(n,m)} \geq 0$. Then, (J.2.7) can be rewritten as

$$\begin{aligned} \mathbf{U} \left(\tau_3 \Sigma^\top \right) \mathbf{V}^\top &= \mathbf{U} \left(\tau_1 \mathbf{I}_n + \tau_2 \Sigma^\top \Sigma \right)^{1/2} \mathbf{U}^\top \tilde{\mathbf{Y}} \\ \Leftrightarrow \tau_3 \Sigma^\top &= \left(\tau_1 \mathbf{I}_n + \tau_2 \Sigma^\top \Sigma \right)^{1/2} \mathbf{U}^\top \tilde{\mathbf{Y}} \mathbf{V}. \end{aligned} \quad (\text{J.2.8})$$

Let $\tilde{\mathbf{Y}} = \mathbf{U} \Xi \mathbf{V}^\top$ for some matrix $\Xi \in \mathbb{R}^{n \times m}$. Then, (J.2.8) reads

$$\tau_3 \Sigma^\top = \left(\tau_1 \mathbf{I}_n + \tau_2 \Sigma^\top \Sigma \right)^{1/2} \Xi. \quad (\text{J.2.9})$$

By $\varsigma_1 > 0$, one can verify from (J.2.9) that Ξ must be written in the following form:

$$\Xi = \begin{bmatrix} \varsigma_{1,\Upsilon} & \mathbf{0}_{m-1}^\top \\ \mathbf{0}_{n-1} & \Xi_{2,2} \end{bmatrix} \in \mathbb{R}^{n \times m}, \quad (\text{J.2.10})$$

where $\varsigma_{1,\Upsilon} > 0$ and $\Xi_{2,2} \in \mathbb{R}^{(n-1) \times (m-1)}$. Then, by the same arguments as in the proof of [35, Proposition 3], $\varsigma_{1,\Upsilon}$ can be shown to be a singular value of Υ , and by (J.2.9) and (J.2.10), condition (iii) implies that

$$\varsigma_{1,\Upsilon}^2 = \frac{\tau_3^2 \varsigma_1^2}{\tau_1 + \tau_2 \varsigma_1^2} \leq 1. \quad (\text{J.2.11})$$

Since it holds from (J.2.5) that

$$\tau_1 + \tau_2 \varsigma_1^2 \geq \tau_1 + \tau_2 \lambda_{\min}(\mathbf{A}^\top \mathbf{A}) > 0, \quad (\text{J.2.12})$$

it holds from (J.2.11) that

$$\begin{aligned} &(\tau_3^2 - \tau_2) \varsigma_1^2 \leq \tau_1 \\ \Leftrightarrow &\left(\frac{1}{\mu_1^{-1} \gamma_1 \rho - 1} - \tau_2 \right) \lambda_{\max}(\mathbf{A}^\top \mathbf{A}) \leq \mu_1^{-1} \gamma_1 (\alpha - \mu_2 \gamma_2^{-1}) \\ \Leftrightarrow &\alpha - \mu_2 \gamma_2^{-1} \geq \left(\frac{\mu_1 \gamma_1^{-1}}{\mu_1^{-1} \gamma_1 \rho - 1} - (1 - \alpha) + \mu_1 \gamma_1^{-1} \right) \lambda_{\max}(\mathbf{A}^\top \mathbf{A}) \\ \Leftrightarrow &\mu_2 \leq \gamma_2 \left[\alpha - \left(\frac{\mu_1 \gamma_1^{-1} \rho}{\rho - \mu_1 \gamma_1^{-1}} - (1 - \alpha) \right) \lambda_{\max}(\mathbf{A}^\top \mathbf{A}) \right]. \end{aligned} \quad (\text{J.2.13})$$

It also holds from (J.2.12) that

$$\begin{aligned} &\mu_1^{-1} \gamma_1 (\alpha - \mu_2 \gamma_2^{-1}) + [(1 - \alpha) \mu_1^{-1} \gamma_1 - 1] \lambda_{\max}(\mathbf{A}^\top [\mathbf{A}]) > 0 \\ \Leftrightarrow &\mu_2 \gamma_2^{-1} \leq \alpha + (1 - \alpha - \mu_1 \gamma_1^{-1}) \lambda_{\min}(\mathbf{A}^\top \mathbf{A}). \end{aligned} \quad (\text{J.2.14})$$

On the other hand, it holds by (J.2.6) that

$$\frac{\mu_1 \gamma_1^{-1} \rho}{\rho - \mu_1 \gamma_1^{-1}} \geq \mu_1 \gamma_1^{-1}, \quad (\text{J.2.15})$$

from which it follows that

$$\frac{\mu_1 \gamma_1^{-1} \rho}{\rho - \mu_1 \gamma_1^{-1}} - (1 - \alpha) \geq \mu_1 \gamma_1^{-1} - (1 - \alpha). \quad (\text{J.2.16})$$

Hence, (J.2.13) and (J.2.14) under (J.2.16) yields (K-II).

Conversely, suppose that condition in (4.16) holds. Then, condition (ii) holds and

$$\begin{aligned} \alpha + (\rho + 1 - \alpha) \lambda_{\max}(\mathbf{A}^\top \mathbf{A}) - (\alpha + (1 - \alpha) \lambda_{\max}(\mathbf{A}^\top \mathbf{A})) &= \rho \lambda_{\max}(\mathbf{A}^\top \mathbf{A}) \\ &\geq 0. \end{aligned} \quad (\text{J.2.17})$$

On the other hand, it holds from (4.16) and (J.2.16) that

$$\begin{aligned} \mu_2 \gamma_2^{-1} &\leq \alpha - (\mu_1 \gamma_1^{-1} - 1 + \alpha) \lambda_{\min}(\mathbf{A}^\top \mathbf{A}) \\ \Leftrightarrow \mu_1^{-1} \gamma_1 (\alpha - \mu_2 \gamma_2^{-1}) &\geq -((1 - \alpha) \mu_1^{-1} \gamma_1 - 1) \lambda_{\max}(\mathbf{A}^\top \mathbf{A}) \\ \Leftrightarrow \tau_1 + \tau_2 \lambda_{\max}(\mathbf{A}^\top \mathbf{A}) &\geq 0, \end{aligned} \quad (\text{J.2.18})$$

which is equivalent to assumption (i). Let

$$\varsigma_{i,\Upsilon} := \frac{\tau_3 \varsigma_i}{(\tau_1 + \tau_2 \varsigma_i^2)^{1/2}} \in (0, 1], \quad \forall \varsigma_i > 0, \quad (\text{J.2.19})$$

and $\varsigma_{i,\Upsilon} := 0$ for all $\varsigma_i = 0$ if any. Define $\Upsilon := \mathbf{U} \Sigma_\Upsilon \mathbf{V}^\top$ for the diagonal matrix $\Sigma_\Upsilon \in \mathbb{R}^{n \times m}$ with diagonal entries $\varsigma_{i,\Upsilon} \in [0, 1]$. One can then verify that this particular Υ satisfies (J.2.8) and thus condition (iii). It is now verified that the condition in (4.16) is equivalent to (J.2.1), and hence it is a sufficient condition for convexity of F in (18).

The necessity is shown in the following. By Proposition 4.6(b), (J.2.1) is necessary and sufficient for convexity when

$$\begin{aligned} \text{int}(K_2 \cap K_3) &= \text{int}(\{[\mathbf{x}^\top \ \boldsymbol{\varepsilon}^\top]^\top \in \mathbb{R}^{n+m} \mid \|\mathbf{y} - \mathbf{A}\mathbf{x} - \boldsymbol{\varepsilon}\|_\infty \leq \gamma_1\} \cap \\ &\quad \{[\mathbf{x}^\top \ \boldsymbol{\varepsilon}^\top]^\top \in \mathbb{R}^{n+m} \mid \|\mathbf{x}\|_\infty \leq \gamma_2\}) \\ &= \text{int} K \neq \emptyset. \end{aligned} \quad (\text{J.2.20})$$

Note here that the dual norm of the ℓ_1 norm is the ℓ_∞ norm [97].

J.3 An Alternative Proof of Proposition 4.1 (the Case of $\alpha = 1$)

Due to Moreau's decomposition (F.2.3), F in (4.13) can be rewritten as

$$\begin{aligned}
 F(\mathbf{x}, \boldsymbol{\varepsilon}) &= \frac{1}{2}\|\mathbf{x}\|_2^2 + \frac{\rho}{2}\|\boldsymbol{\varepsilon}\|_2^2 - \mu_1\gamma_1\|\cdot\|_1(\mathbf{A}\mathbf{x} + \boldsymbol{\varepsilon} - \mathbf{y}) - \mu_2\gamma_2\|\cdot\|_1(\mathbf{x}) \\
 &= \frac{1}{2}\|\mathbf{x}\|_2^2 + \frac{\rho}{2}\|\boldsymbol{\varepsilon}\|_2^2 - \frac{\mu_1}{2\gamma_1}\|\mathbf{A}\mathbf{x} + \boldsymbol{\varepsilon} - \mathbf{y}\|_2^2 \\
 &\quad + \mu_1\gamma_1^{-1}(\|\cdot\|_\infty)(\gamma_1^{-1}(\mathbf{A}\mathbf{x} + \boldsymbol{\varepsilon} - \mathbf{y})) - \frac{\mu_2}{2\gamma_2}\|\mathbf{x}\|_2^2 \\
 &\quad + \mu_2\gamma_2^{-1}(\|\cdot\|_\infty)(\gamma_2^{-1}\mathbf{x}). \tag{J.3.1}
 \end{aligned}$$

First, we consider the scalar case. Let $\mathbf{A} = a \in \mathbb{R}$. Assume that $[x, \varepsilon]^\top \in \text{int } K (\neq \emptyset)$. Then, by inspecting (J.7.11) in the case of F in (4.13), it follows from (J.3.1) that

$$\begin{aligned}
 2F(x, \varepsilon) &= -\mu_1\gamma_1^{-1}(ax + \varepsilon)^2 + \rho\varepsilon^2 + x^2 - \mu_2\gamma_2^{-1}x^2 \\
 &= (1 - \mu_1\gamma_1^{-1}a^2 - \mu_2\gamma_2^{-1})x^2 - 2\mu_1\gamma_1^{-1}ax\varepsilon + (\rho - \mu_1\gamma_1^{-1})\varepsilon^2. \tag{J.3.2}
 \end{aligned}$$

The right-hand side of (J.3.2) is convex if and only if the following set of conditions are satisfied:

$$\begin{cases} \rho - \mu_1\gamma_1^{-1} \geq 0, & \text{(J.3.3)} \\ 1 - \mu_1\gamma_1^{-1}a^2 - \mu_2\gamma_2^{-1} \geq 0, & \text{(J.3.4)} \\ (1 - \mu_1\gamma_1^{-1}a^2 - \mu_2\gamma_2^{-1})(\rho - \mu_1\gamma_1^{-1}) - \mu_1^2\gamma_1^{-2}a^2 \geq 0. & \text{(J.3.5)} \end{cases}$$

This set of conditions is clearly sufficient for F to be convex also outside of $\text{int } K$ since the only difference between the right-hand side of (J.3.2) and $2F(\mathbf{x}, \boldsymbol{\varepsilon})$ are some negative quadratic terms, as can be verified by (J.3.1).

For the general case, let $\mathbf{A} = \mathbf{V}\boldsymbol{\Sigma}\mathbf{U}^\top$ be a singular value decomposition of \mathbf{A} , where $\mathbf{V} \in \mathbb{R}^{m \times m}$ and $\mathbf{U} \in \mathbb{R}^{n \times n}$ are orthogonal matrices, and the diagonal entries of $\boldsymbol{\Sigma} \in \mathbb{R}^{m \times n}$ satisfy $\varsigma_1 := \lambda_{\max}^{1/2}(\mathbf{A}^\top \mathbf{A}) \geq \varsigma_2 \geq \dots \geq \varsigma_{\min\{n, m\}} \geq 0$. Assume that $[\mathbf{x}^\top, \boldsymbol{\varepsilon}^\top]^\top \in \text{int } K$. Then, by inspecting (J.7.11) in the case of F in (4.13), it follows from (J.3.1) that

$$2F(\mathbf{x}, \boldsymbol{\varepsilon}) = -\mu_1\gamma_1^{-1}\|\mathbf{A}\mathbf{x} + \boldsymbol{\varepsilon}\|_2^2 + \rho\|\boldsymbol{\varepsilon}\|_2^2 + \|\mathbf{x}\|_2^2 - \mu_2\gamma_2^{-1}\|\mathbf{x}\|_2^2. \tag{J.3.6}$$

Set $\tilde{\mathbf{x}} := \mathbf{U}\mathbf{x}$ and $\tilde{\boldsymbol{\varepsilon}} := \mathbf{V}\boldsymbol{\varepsilon}$. Since $\|\tilde{\mathbf{x}}\|_2 = \|\mathbf{x}\|_2$ and $\|\tilde{\boldsymbol{\varepsilon}}\|_2 = \|\boldsymbol{\varepsilon}\|_2$, the right-hand side of (J.3.6) can be expressed equivalently as

$$\begin{aligned}
 & -\mu_1\gamma_1^{-1}\|\boldsymbol{\Sigma}\tilde{\mathbf{x}} + \tilde{\boldsymbol{\varepsilon}}\|_2^2 + \rho\|\tilde{\boldsymbol{\varepsilon}}\|_2^2 + \|\tilde{\mathbf{x}}\|_2^2 - \mu_2\gamma_2^{-1}\|\tilde{\mathbf{x}}\|_2^2 \\
 &= \sum_{j=1}^{\min\{m, n\}} \left[(1 - \mu_1\gamma_1^{-1}\varsigma_j^2 - \mu_2\gamma_2^{-1})\tilde{x}_j^2 - 2\mu_1\gamma_1^{-1}\varsigma_j\tilde{x}_j\tilde{\varepsilon}_j + (\rho - \mu_1\gamma_1^{-1})\tilde{\varepsilon}_j^2 \right]. \tag{J.3.7}
 \end{aligned}$$

This sum is convex if and only if each summand is convex. This happens if and only if (J.3.3), (J.3.4), and (J.3.5) are satisfied for each $a = \varsigma_j$ for $j = 1, 2, \dots, \min\{m, n\}$ since ς_j in (J.3.7) plays the same role as a in (J.3.2). Since the left-hand sides of (J.3.4) and (J.3.5) are monotonically decreasing for a and $\varsigma_1 \geq \varsigma_2 \geq \dots \geq \varsigma_{\min\{m, n\}}$, it is sufficient to verify that if (J.3.3), (J.3.4), and (J.3.5) hold with $a = \varsigma_1$. Since $\varsigma_1^2 = \lambda_{\max}(\mathbf{A}^\top \mathbf{A})$, the conditions for convexity of F are given by

$$\begin{cases} \rho - \mu_1 \gamma_1^{-1} \geq 0, & \text{(J.3.8)} \\ 1 - \mu_1 \gamma_1^{-1} \lambda_{\max}(\mathbf{A}^\top \mathbf{A}) - \mu_2 \gamma_2^{-1} \geq 0, & \text{(J.3.9)} \\ \mu_1^2 \gamma_1^{-2} \lambda_{\max}(\mathbf{A}^\top \mathbf{A}) \\ \leq \left(1 - \mu_1 \gamma_1^{-1} \lambda_{\max}(\mathbf{A}^\top \mathbf{A}) - \mu_2 \gamma_2^{-1}\right) (\rho - \mu_1 \gamma_1^{-1}). & \text{(J.3.10)} \end{cases}$$

The rest of the proof will be devoted to the equivalence “(J.3.8), (J.3.9), and (J.3.10)” \Leftrightarrow (4.3.2).

- Case of $\lambda_{\max}(\mathbf{A}^\top \mathbf{A}) = 0$:

It clearly holds that

$$(J.3.8) \Leftrightarrow (K-I), \quad (J.3.11)$$

$$(J.3.9) \Leftrightarrow (K-II). \quad (J.3.12)$$

Since (J.3.8) and (J.3.9) imply (J.3.10), it holds that

$$\text{“(J.3.8), (J.3.9), and (J.3.10)”} \Leftrightarrow \text{“(J.3.8) and (J.3.9)”}. \quad (J.3.13)$$

Hence, (J.3.12), (J.3.11), and (J.3.13) verify the equivalence

$$\text{“(J.3.8), (J.3.9), and (J.3.10)”} \Leftrightarrow (4.3.2). \quad (J.3.14)$$

- Case of $\lambda_{\max}(\mathbf{A}^\top \mathbf{A}) > 0$:

Assume that (J.3.8), (J.3.9), and (J.3.10) are satisfied. Then, since

$$\begin{aligned} 0 &< \mu_1^2 \gamma_1^{-2} \lambda_{\max}(\mathbf{A}^\top \mathbf{A}) \\ &\leq \left(1 - \mu_1 \gamma_1^{-1} \lambda_{\max}(\mathbf{A}^\top \mathbf{A}) - \mu_2 \gamma_2^{-1}\right) (\rho - \mu_1 \gamma_1^{-1}), \end{aligned} \quad (J.3.15)$$

(J.3.8) and (J.3.9) hold with strict inequality, *i.e.*,

$$1 - \mu_1 \gamma_1^{-1} \lambda_{\max}(\mathbf{A}^\top \mathbf{A}) - \mu_2 \gamma_2^{-1} > 0, \quad (J.3.16)$$

$$\rho - \mu_1 \gamma_1^{-1} > 0. \quad (J.3.17)$$

Since

$$\begin{aligned} (J.3.10) &\Leftrightarrow -\mu_1^2 \gamma_1^{-2} \lambda_{\max}(\mathbf{A}^\top \mathbf{A}) \leq (\rho - \mu_1 \gamma_1^{-1}) \mu_1 \gamma_1^{-1} \lambda_{\max}(\mathbf{A}^\top \mathbf{A}) \\ &\quad + (1 - \mu_2 \gamma_2^{-1}) (\rho - \mu_1 \gamma_1^{-1}) \\ &\Leftrightarrow \rho \mu_1 \gamma_1^{-1} \lambda_{\max}(\mathbf{A}^\top \mathbf{A}) \leq (1 - \mu_2 \gamma_2^{-1}) (\rho - \mu_1 \gamma_1^{-1}), \end{aligned} \quad (J.3.18)$$

from which together with (J.3.17) it follows that

$$1 - \mu_2 \gamma_2^{-1} \geq \frac{\rho \mu_1 \gamma_1^{-1} \lambda_{\max}(\mathbf{A}^\top \mathbf{A})}{\rho - \mu_1 \gamma_1^{-1}} \Leftrightarrow \mu_2 \leq \gamma_2 \left[1 - \frac{\mu_1 \rho \lambda_{\max}(\mathbf{A}^\top \mathbf{A})}{\gamma_1 \rho - \mu_1} \right]. \quad (\text{J.3.19})$$

Since $\mu_2 \geq 0$ and $\gamma_2 > 0$, it follows from (J.3.17) and (J.3.19) that

$$\begin{aligned} 1 - \frac{\mu_1 \rho \lambda_{\max}(\mathbf{A}^\top \mathbf{A})}{\gamma_1 \rho - \mu_1} \geq 0 &\Leftrightarrow \gamma_1 \rho - \mu_1 \geq \mu_1 \rho \lambda_{\max}(\mathbf{A}^\top \mathbf{A}) \\ &\Leftrightarrow \mu_1 \leq \gamma_1 \rho (1 + \rho \lambda_{\max}(\mathbf{A}^\top \mathbf{A}))^{-1} \\ &\Leftrightarrow \mu_2 \geq 0. \end{aligned} \quad (\text{J.3.20})$$

Therefore, it holds from (J.3.19) and (J.3.20) that

$$\text{“(J.3.8), (J.3.9), and (J.3.10)”} \Rightarrow (4.3.2). \quad (\text{J.3.21})$$

Conversely, suppose that (4.3.2) is satisfied. Then, it holds from $\lambda_{\max}(\mathbf{A}^\top \mathbf{A}) > 0$ that

$$\mu_1 < \gamma_1 \rho, \quad (\text{J.3.22})$$

which is equivalent to (J.3.17), and hence (J.3.8) holds. It also holds from $-1/(\gamma_1 \rho - \mu_1) > -1/(\gamma_1 \rho)$ that

$$\begin{aligned} \mu_2 &\leq \gamma_2 \left[1 - \frac{\mu_1 \rho \lambda_{\max}(\mathbf{A}^\top \mathbf{A})}{\gamma_1 \rho - \mu_1} \right] \\ &< \gamma_2 \left(1 - \mu_1 \gamma_1^{-1} \lambda_{\max}(\mathbf{A}^\top \mathbf{A}) \right), \end{aligned} \quad (\text{J.3.23})$$

which is equivalent to (J.3.16), and hence (J.3.9) holds. Finally, combining (J.3.18) and (J.3.19) in light of (J.3.22) verifies

$$(4.3.2) \Rightarrow (J.3.10). \quad (\text{J.3.24})$$

Therefore, we obtain

$$(4.3.2) \Rightarrow \text{“(J.3.8), (J.3.9), and (J.3.10)”}. \quad (\text{J.3.25})$$

The proof is completed by (J.3.21) and (J.3.25).

Remark 5. In this alternative proof, the discussion of the positive semidefiniteness of F is simplified by rewriting the quadratic parts of F as the separable sum of functions across each variable as in (J.3.7) through a singular value decomposition of \mathbf{A} . This is the essential difference from the original proof in Appendix J.2.

J.4 Proof of Proposition 4.3

Proof of Proposition 4.3(a). Let $\mathbf{a}_i = \mathbf{a}_j =: \mathbf{a}$. Since $|\hat{x}_i| > \gamma$ and $|\hat{x}_j| > \gamma$, we have $|\hat{x}_i| + |\hat{x}_j| > 2\gamma$, which yields $\gamma/(|\hat{x}_i| + |\hat{x}_j|) < 1/2$. By (4.27), it holds that

$$\begin{aligned} \hat{x}_i^* \mathbf{a}_i + \hat{x}_j^* \mathbf{a}_j &= (\hat{x}_i^* + \hat{x}_j^*) \mathbf{a} \\ &= ((\hat{x}_i + \hat{x}_j)\alpha + (\hat{x}_i + \hat{x}_j)(1 - \alpha)) \mathbf{a} \\ &= (\hat{x}_i + \hat{x}_j) \mathbf{a} = \hat{x}_i \mathbf{a}_i + \hat{x}_j \mathbf{a}_j, \end{aligned} \quad (\text{J.4.1})$$

from which it follows that $\mathbf{A}\hat{\mathbf{x}} = \mathbf{A}\hat{\mathbf{x}}^*$, and hence

$$H(\mathbf{A}\hat{\mathbf{x}}) = H(\mathbf{A}\hat{\mathbf{x}}^*). \quad (\text{J.4.2})$$

Since $\alpha \in \left[\frac{\gamma}{|\hat{x}_i| + |\hat{x}_j|}, \frac{1}{2} \right]$, we have

$$\begin{aligned} |\hat{x}_i + \hat{x}_j|(1 - \alpha) &\geq |\hat{x}_i + \hat{x}_j|\alpha \\ &\geq |\hat{x}_i + \hat{x}_j| \frac{\gamma}{|\hat{x}_i| + |\hat{x}_j|} = \gamma, \end{aligned} \quad (\text{J.4.3})$$

where the equality is due to $|\hat{x}_i + \hat{x}_j| = |\hat{x}_i| + |\hat{x}_j|$ since $\hat{x}_i \hat{x}_j > 0$. Since $\phi_\gamma^{\text{MC}}(x) = \gamma/2$ for any x satisfying $|x| \geq \gamma$ by (2.45), we obtain

$$\begin{aligned} \phi_\gamma^{\text{MC}}(\hat{x}_i^*) + \phi_\gamma^{\text{MC}}(\hat{x}_j^*) &= \phi_\gamma^{\text{MC}}((\hat{x}_i + \hat{x}_j)\alpha) + \phi_\gamma^{\text{MC}}((\hat{x}_i + \hat{x}_j)(1 - \alpha)) \\ &= \gamma/2 + \gamma/2 = \phi_\gamma^{\text{MC}}(\hat{x}_i) + \phi_\gamma^{\text{MC}}(\hat{x}_j), \end{aligned} \quad (\text{J.4.4})$$

which yields

$$\Phi_\gamma^{\text{MC}}(\hat{\mathbf{x}}) = \Phi_\gamma^{\text{MC}}(\hat{\mathbf{x}}^*). \quad (\text{J.4.5})$$

By (J.4.2) and (J.4.5), we obtain $J(\hat{\mathbf{x}}^*) = J(\hat{\mathbf{x}})$. \square

Proof of Proposition 4.3(b). We will derive contradiction by assuming that $\hat{x}_i \neq \hat{x}_j$. Define $\hat{\mathbf{x}}^*$ as (4.27) for $\alpha = \frac{1}{2}$ so that $\hat{x}_i^* = \hat{x}_j^* = \frac{\hat{x}_i + \hat{x}_j}{2}$. Then, we have $H(\mathbf{A}\hat{\mathbf{x}}) = H(\mathbf{A}\hat{\mathbf{x}}^*)$ and $\Phi_\gamma^{\text{MC}}(\hat{\mathbf{x}}) = \Phi_\gamma^{\text{MC}}(\hat{\mathbf{x}}^*)$ in the same way as the proof of Proposition 4.3(a). However, the strict convexity of the quadratic function under $\hat{x}_i \neq \hat{x}_j$ implies that

$$\frac{1}{2} \|\hat{x}_i^*\|_2^2 + \frac{1}{2} \|\hat{x}_j^*\|_2^2 = 2 \cdot \frac{1}{2} \left\| \frac{\hat{x}_i + \hat{x}_j}{2} \right\|_2^2 < \frac{1}{2} \|\hat{x}_i\|_2^2 + \frac{1}{2} \|\hat{x}_j\|_2^2, \quad (\text{J.4.6})$$

from which it follows that $\frac{1}{2} \|\hat{\mathbf{x}}^*\|_2^2 < \frac{1}{2} \|\hat{\mathbf{x}}\|_2^2$. Hence, we have $J(\hat{\mathbf{x}}^*) < J(\hat{\mathbf{x}})$, which contradicts the assumption that $\hat{\mathbf{x}}$ is a minimizer. \square

J.5 Proof of Proposition 4.4

The following lemma is used in the proof.

Lemma J.5.1 ([156, Propositions 4.2.4 and 4.2.5]). Let $\phi_1 : \mathbb{R}^n \rightarrow \mathbb{R}$ and $\phi_2 : \mathbb{R}^m \rightarrow \mathbb{R}$ be convex functions, and $\mathbf{A} \in \mathbb{R}^{m \times n}$. Then, $\partial(\phi_1 + \phi_2 \circ \mathbf{A}) = \partial\phi_1 + \mathbf{A}^\top \circ (\partial\phi_2) \circ \mathbf{A}$.

By (J.2.18), (K-II) implies assumption (a). Let $\tilde{\mathbf{y}} := \mathbf{y} + \boldsymbol{\varepsilon}$. Since $\hat{\mathbf{x}} \in \operatorname{argmin}_{\mathbf{x} \in \mathbb{R}^n} J(\mathbf{x})$, it holds that

$$\mathbf{0} \in \partial J(\hat{\mathbf{x}}) = \partial \left(\mu_1 \Phi_{\gamma_1}^{\text{MC}}(\mathbf{A} \cdot -\tilde{\mathbf{y}}) + \frac{1-\alpha}{2} \|\mathbf{A} \cdot -\mathbf{y}\|_2^2 + \mu_2 \Phi_{\gamma_2}^{\text{MC}} + \frac{\alpha}{2} \|\cdot\|_2^2 \right) (\hat{\mathbf{x}}). \quad (\text{J.5.1})$$

By (2.46) and (F.2.3), it holds for any $\mathbf{x} \in \mathbb{R}^n$ that

$$\Phi_{\gamma_1}^{\text{MC}}(\mathbf{x}) = \|\mathbf{x}\|_1 + \gamma_1^{-1} (\|\cdot\|_1^*)(\gamma_1^{-1} \mathbf{x}) - \frac{1}{2\gamma_1} \|\mathbf{x}\|_2^2, \quad (\text{J.5.2})$$

from which it follows that the right side in (J.5.1) becomes

$$\begin{aligned} & \partial \left(\mu_1 \|\mathbf{A} \cdot -\tilde{\mathbf{y}}\|_1 + \mu_1 \gamma_1^{-1} (\|\cdot\|_1^*)(\gamma_1^{-1}(\mathbf{A} \cdot -\tilde{\mathbf{y}})) \right. \\ & \quad - \frac{\mu_1 \gamma_1^{-1}}{2} \|\mathbf{A} \cdot -\mathbf{y}\|_2^2 + \frac{1-\alpha}{2} \|\mathbf{A} \cdot -\tilde{\mathbf{y}}\|_2^2 + \frac{\alpha}{2} \|\cdot\|_2^2 + \mu_2 \|\cdot\|_1 \\ & \quad \left. + \mu_2 \gamma_2^{-1} (\|\cdot\|_1^*)(\gamma_2^{-1} \cdot) - \frac{\mu_2 \gamma_2^{-1}}{2} \|\cdot\|_2^2 \right) (\hat{\mathbf{x}}) \\ &= \partial \left(\mu_1 \|\mathbf{A} \cdot -\tilde{\mathbf{y}}\|_1 + \mu_1 \gamma_1^{-1} (\|\cdot\|_1^*)(\gamma_1^{-1}(\mathbf{A} \cdot -\tilde{\mathbf{y}})) + \mu_2 \|\cdot\|_1 \right. \\ & \quad \left. + \mu_2 \gamma_2^{-1} (\|\cdot\|_1^*)(\gamma_2^{-1} \cdot) \right) (\hat{\mathbf{x}}) + \alpha \hat{\mathbf{x}} - \mu_1 \gamma_1^{-1} \mathbf{A}^\top (\mathbf{A} \hat{\mathbf{x}} - \tilde{\mathbf{y}}) \\ & \quad + (1-\alpha) \mathbf{A}^\top (\mathbf{A} \hat{\mathbf{x}} - \mathbf{y}) - \mu_2 \gamma_2^{-1} \hat{\mathbf{x}} \\ &= \mu_1 \partial(\|\mathbf{A} \cdot -\tilde{\mathbf{y}}\|_1) (\hat{\mathbf{x}}) + \mu_1 \partial(\gamma_1^{-1} (\|\cdot\|_1^*)(\gamma_1^{-1}(\mathbf{A} \cdot -\tilde{\mathbf{y}}))) (\hat{\mathbf{x}}) \\ & \quad + \mu_2 \partial(\|\cdot\|_1) (\hat{\mathbf{x}}) + \mu_2 \partial(\gamma_2^{-1} (\|\cdot\|_1^*)(\gamma_2^{-1} \cdot)) (\hat{\mathbf{x}}) + \alpha \hat{\mathbf{x}} - \mu_1 \gamma_1^{-1} \mathbf{A}^\top (\mathbf{A} \hat{\mathbf{x}} - \tilde{\mathbf{y}}) \\ & \quad + (1-\alpha) \mathbf{A}^\top (\mathbf{A} \hat{\mathbf{x}} - \mathbf{y}) - \mu_2 \gamma_2^{-1} \hat{\mathbf{x}}. \end{aligned} \quad (\text{J.5.3})$$

Here, the first equality is verified by Lemma J.5.1 and assumption (a), which is sufficient to guarantee the convexity of $-\mu_1 \gamma_1^{-1} \|\mathbf{A} \cdot -\tilde{\mathbf{y}}\|_2^2/2 + (1-\alpha) \|\mathbf{A} \cdot -\mathbf{y}\|_2^2/2 + \alpha \|\cdot\|_2^2/2 - \mu_2 \gamma_2^{-1} \|\cdot\|_2^2/2$ by the equivalence

$$(a) \Leftrightarrow (\mu_1^{-1} \gamma_1 (1-\alpha) - 1) \mathbf{A}^\top \mathbf{A} + \mu_1^{-1} \gamma_1 (\alpha - \mu_2 \gamma_2^{-1}) \mathbf{I}_n \succeq \mathbf{0}_{n \times n}. \quad (\text{J.5.4})$$

The second equality of (J.5.3) is verified by Lemma J.5.1. By virtue of identity $\text{Id} = \text{Prox}_{\gamma f} + \gamma \text{Prox}_{f^*/\gamma} \circ \gamma^{-1} \text{Id}$ [96] together with $\nabla(\gamma f) =$

$\gamma^{-1}(\text{Id} - \text{Prox}_{\gamma f})$ and Lemma J.5.1, the first and second terms of the right-most side of (J.5.3) reduce to

$$\begin{aligned} & \mu_1 \mathbf{A}^\top \partial \|\cdot\|_1(\mathbf{A}\hat{\mathbf{x}} - \tilde{\mathbf{y}}) + \mu_1 \mathbf{A}^\top (\text{Id} - \text{Prox}_{\gamma_1^{-1}\|\cdot\|_1^*})(\gamma_1^{-1}(\mathbf{A}\hat{\mathbf{x}} - \tilde{\mathbf{y}})) \\ &= \mu_1 \mathbf{A}^\top \partial \|\cdot\|_1(\mathbf{A}\hat{\mathbf{x}} - \tilde{\mathbf{y}}) + \mu_1 \gamma_1^{-1} \mathbf{A}^\top \text{Prox}_{\gamma_1\|\cdot\|_1}(\mathbf{A}\hat{\mathbf{x}} - \tilde{\mathbf{y}}) \\ &= \mu_1 \mathbf{A}^\top \partial \|\cdot\|_1(\mathbf{A}\hat{\mathbf{x}} - \tilde{\mathbf{y}}) + \mu_1 \gamma_1^{-1} \mathbf{A}^\top \text{Soft}_{\gamma_1}(\mathbf{A}\hat{\mathbf{x}} - \tilde{\mathbf{y}}). \end{aligned} \quad (\text{J.5.5})$$

In a similar way, the third and fourth terms reduce to

$$\partial \left(\mu_2 \|\cdot\|_1 + \mu_2 \gamma_2^{-1} (\|\cdot\|_1^*)(\gamma_2^{-1}\cdot) \right) (\hat{\mathbf{x}}) = \mu_2 \partial \|\cdot\|_1(\hat{\mathbf{x}}) + \mu_2 \gamma_2^{-1} \text{Soft}_{\gamma_2}(\hat{\mathbf{x}}). \quad (\text{J.5.6})$$

Combining (J.5.1), (J.5.3), (J.5.5), and (J.5.6) yields

$$\begin{aligned} \mathbf{0} &\in \mu_1 \mathbf{A}^\top \partial \|\cdot\|_1(\mathbf{A}\hat{\mathbf{x}} - \tilde{\mathbf{y}}) + \mu_1 \gamma_1^{-1} \mathbf{A}^\top \text{Soft}_{\gamma_1}(\mathbf{A}\hat{\mathbf{x}} - \tilde{\mathbf{y}}) \\ &\quad + \mu_2 \partial \|\cdot\|_1(\hat{\mathbf{x}}) + \mu_2 \gamma_2^{-1} \text{Soft}_{\gamma_2}(\hat{\mathbf{x}}) \\ &\quad + \alpha \hat{\mathbf{x}} - \mu_1 \gamma_1^{-1} \mathbf{A}^\top (\mathbf{A}\hat{\mathbf{x}} - \tilde{\mathbf{y}}) + (1 - \alpha) \mathbf{A}^\top (\mathbf{A}\hat{\mathbf{x}} - \mathbf{y}) - \mu_2 \gamma_2^{-1} \hat{\mathbf{x}}. \end{aligned} \quad (\text{J.5.7})$$

Considering the i th and j th components of both sides of (J.5.7), there exist some $\mathbf{s} := [s_1, s_2, \dots, s_m]^\top \in \partial \|\cdot\|_1(\mathbf{A}\hat{\mathbf{x}} - \tilde{\mathbf{y}})$ and $\mathbf{t} := [t_1, t_2, \dots, t_n]^\top \in \partial \|\cdot\|_1(\hat{\mathbf{x}})$ such that

$$\begin{aligned} 0 &= \mu_1 \mathbf{a}_i^\top (\gamma_1^{-1}(\text{Soft}_{\gamma_1} - \text{Id})(\mathbf{A}\hat{\mathbf{x}} - \tilde{\mathbf{y}}) + \mathbf{s}) + \alpha \hat{x}_i \\ &\quad + \mu_2 (\gamma_2^{-1}(\text{soft}_{\gamma_2} - 1)(\hat{x}_i) + t_i) + (1 - \alpha) \mathbf{a}_i^\top (\mathbf{A}\hat{\mathbf{x}} - \mathbf{y}), \end{aligned} \quad (\text{J.5.8})$$

$$\begin{aligned} 0 &= \mu_1 \mathbf{a}_j^\top (\gamma_1^{-1}(\text{Soft}_{\gamma_1} - \text{Id})(\mathbf{A}\hat{\mathbf{x}} - \tilde{\mathbf{y}}) + \mathbf{s}) + \alpha \hat{x}_j \\ &\quad + \mu_2 (\gamma_2^{-1}(\text{soft}_{\gamma_2} - 1)(\hat{x}_j) + t_j) + (1 - \alpha) \mathbf{a}_j^\top (\mathbf{A}\hat{\mathbf{x}} - \mathbf{y}). \end{aligned} \quad (\text{J.5.9})$$

From (J.5.8) and (J.5.9), we obtain

$$\begin{aligned} \hat{x}_i - \hat{x}_j &= \alpha^{-1} \left[\mu_1 (\mathbf{a}_i^\top - \mathbf{a}_j^\top) (\gamma_1^{-1}(\text{Id} - \text{Soft}_{\gamma_1})(\mathbf{A}\hat{\mathbf{x}} - \tilde{\mathbf{y}}) - \mathbf{s}) \right. \\ &\quad \left. + \mu_2 (\gamma_2^{-1}(1 - \text{soft}_{\gamma_2})(\hat{x}_i) - t_i) - \mu_2 (\gamma_2^{-1}(1 - \text{soft}_{\gamma_2})(\hat{x}_j) - t_j) \right] \\ &\quad - \alpha^{-1} (1 - \alpha) (\mathbf{a}_i^\top - \mathbf{a}_j^\top) (\mathbf{A}\hat{\mathbf{x}} - \mathbf{y}). \end{aligned} \quad (\text{J.5.10})$$

To evaluate $\gamma_1^{-1}(\text{Id} - \text{Soft}_{\gamma_1})(\mathbf{A}\hat{\mathbf{x}} - \tilde{\mathbf{y}}) - \mathbf{s}$, let us first consider the case when $\mathbf{a}_i^\top \hat{\mathbf{x}} - \tilde{y}_i \neq 0$ for $i = 1, 2, \dots, m$. In this case, it holds that $s_i = \text{sign}(\mathbf{a}_i^\top \hat{\mathbf{x}} - \tilde{y}_i)$, and the i th component of $\gamma_1^{-1}(\text{Id} - \text{Soft}_{\gamma_1})(\mathbf{A}^\top \hat{\mathbf{x}} - \tilde{\mathbf{y}})$ is given by

$$\gamma_1^{-1}(1 - \text{soft}_{\gamma_1})(\mathbf{a}_i^\top \hat{\mathbf{x}} - \tilde{y}_i) = \begin{cases} \gamma_1^{-1}(\mathbf{a}_i^\top \hat{\mathbf{x}} - \tilde{y}_i), & \text{if } |\mathbf{a}_i^\top \hat{\mathbf{x}} - \tilde{y}_i| \leq \gamma_1, \\ \text{sign}(\mathbf{a}_i^\top \hat{\mathbf{x}} - \tilde{y}_i), & \text{if } |\mathbf{a}_i^\top \hat{\mathbf{x}} - \tilde{y}_i| > \gamma_1. \end{cases} \quad (\text{J.5.11})$$

Hence, it holds that

$$\begin{aligned} & \gamma_1^{-1}(1 - \text{soft}_{\gamma_1})(\mathbf{a}_i^\top \hat{\mathbf{x}} - \tilde{y}_i) - s_i \\ &= \begin{cases} \gamma_1^{-1}(\mathbf{a}_i^\top \hat{\mathbf{x}} - \tilde{y}_i) - 1, & \text{if } \gamma_1^{-1}(\mathbf{a}_i^\top \hat{\mathbf{x}} - \tilde{y}_i) \in (0, 1], \\ \gamma_1^{-1}(\mathbf{a}_i^\top \hat{\mathbf{x}} - \tilde{y}_i) + 1, & \text{if } \gamma_1^{-1}(\mathbf{a}_i^\top \hat{\mathbf{x}} - \tilde{y}_i) \in [-1, 0), \\ 0, & \text{if } \gamma_1^{-1}|\mathbf{a}_i^\top \hat{\mathbf{x}} - \tilde{y}_i| > 1, \end{cases} \end{aligned} \quad (\text{J.5.12})$$

from which it follows that

$$\gamma_1^{-1}(1 - \text{soft}_{\gamma_1})(\mathbf{a}_i^\top \hat{\mathbf{x}} - \tilde{y}_i) - s_i \in (-1, 1). \quad (\text{J.5.13})$$

Let us now consider the case when $\mathbf{a}_i^\top \hat{\mathbf{x}} - \tilde{y}_i = 0$. Then, it holds by (J.5.11) that

$$\gamma_1^{-1}(1 - \text{soft}_{\gamma_1})(\mathbf{a}_i^\top \hat{\mathbf{x}} - \tilde{y}_i) - s_i = -s_i \in [-1, 1]. \quad (\text{J.5.14})$$

By (J.5.13) and (J.5.14), it follows that $|\gamma_1^{-1}(1 - \text{soft}_{\gamma_1})(\mathbf{a}_i^\top \hat{\mathbf{x}} - \tilde{y}_i) - s_i| \leq 1$, and hence

$$\|\gamma_1^{-1}(\text{Id} - \text{Soft}_{\gamma_1})(\mathbf{A}\hat{\mathbf{x}} - \tilde{\mathbf{y}}) - \mathbf{s}\|_2 \leq \sqrt{m}. \quad (\text{J.5.15})$$

In a similar way, it holds under assumption (b) that

$$\begin{aligned} \gamma_2^{-1}(1 - \text{soft}_{\gamma_2})(\hat{x}_i) - t_i &= \begin{cases} \gamma_2^{-1}\hat{x}_i - 1, & \text{if } \gamma_2^{-1}\hat{x}_i \in (0, 1], \\ \gamma_2^{-1}\hat{x}_i + 1, & \text{if } \gamma_2^{-1}\hat{x}_i \in [-1, 0), \\ 0, & \text{if } \gamma_2^{-1}|\hat{x}_i| > 1 \end{cases} \\ &= -(\phi_{\gamma_2}^{\text{MC}})'(\hat{x}_i). \end{aligned} \quad (\text{J.5.16})$$

where the derivative exists everywhere but the origin. Since $J(\mathbf{0}) \geq J(\hat{\mathbf{x}})$ (see (4.26)), it holds that

$$\begin{aligned} \frac{1-\alpha}{2} \|\mathbf{A}\hat{\mathbf{x}} - \mathbf{y}\|_2^2 &\leq \mu_1 \Phi_{\gamma_1}^{\text{MC}}(\mathbf{A}\hat{\mathbf{x}} - \tilde{\mathbf{y}}) + \frac{1-\alpha}{2} \|\mathbf{A}\hat{\mathbf{x}} - \mathbf{y}\|_2^2 + \mu_2 \Phi_{\gamma_2}^{\text{MC}}(\hat{\mathbf{x}}) + \frac{\alpha}{2} \|\hat{\mathbf{x}}\|_2^2 \\ &\leq \mu_1 \Phi_{\gamma_1}^{\text{MC}}(\tilde{\mathbf{y}}) + \frac{1-\alpha}{2} \|\mathbf{y}\|_2^2 \\ &\leq \frac{\mu_1 \gamma_1}{2} + \frac{1-\alpha}{2} \|\mathbf{y}\|_2^2, \end{aligned} \quad (\text{J.5.17})$$

from which it follows that

$$(1-\alpha) \|\mathbf{A}\hat{\mathbf{x}} - \tilde{\mathbf{y}}\|_2 \leq \sqrt{(1-\alpha)\mu_1\gamma_1 + (1-\alpha)^2 \|\mathbf{y}\|_2^2}, \quad (\text{J.5.18})$$

Equations (J.5.10), (J.5.15), (J.5.16), and (J.5.18) yield

$$\begin{aligned}
|\hat{x}_i - \hat{x}_j| &\leq \alpha^{-1} \left[\mu_1 |(\mathbf{a}_i^\top - \mathbf{a}_j^\top)(\gamma_1^{-1}(\text{Id} - \text{Soft}_{\gamma_1})(\mathbf{A}\hat{\mathbf{x}} - \tilde{\mathbf{y}}) - \mathbf{s})| \right. \\
&\quad + \mu_2 |(\gamma_2^{-1}(1 - \text{soft}_{\gamma_2})(\hat{x}_i) - t_i) \\
&\quad \left. - (\gamma_2^{-1}(1 - \text{soft}_{\gamma_2})(\hat{x}_j) - t_j) \right] \\
&\quad + \alpha^{-1}(1 - \alpha) |(\mathbf{a}_i^\top - \mathbf{a}_j^\top)(\mathbf{A}\hat{\mathbf{x}} - \mathbf{y})| \\
&\leq \alpha^{-1} \left[\mu_1 \|\mathbf{a}_i - \mathbf{a}_j\|_2 \|\gamma_1^{-1}(\text{Id} - \text{Soft}_{\gamma_1})(\mathbf{A}\hat{\mathbf{x}} - \tilde{\mathbf{y}}) - \mathbf{s}\|_2 \right. \\
&\quad + \mu_2 |(\phi_{\gamma_2}^{\text{MC}})'(\hat{x}_i) - (\phi_{\gamma_2}^{\text{MC}})'(\hat{x}_j)| \left. \right] \\
&\quad + \alpha^{-1}(1 - \alpha) \|\mathbf{a}_i^\top - \mathbf{a}_j^\top\|_2 \|\mathbf{A}\hat{\mathbf{x}} - \mathbf{y}\|_2 \\
&\leq \alpha^{-1} \left[\sqrt{2(1 - \mathbf{a}_i^\top \mathbf{a}_j)(\mu_1^2 m + (1 - \alpha)\mu_1 \gamma_1 + (1 - \alpha)^2 \|\mathbf{y}\|_2^2)} \right. \\
&\quad \left. + \mu_2 |(\phi_{\gamma_2}^{\text{MC}})'(\hat{x}_i) - (\phi_{\gamma_2}^{\text{MC}})'(\hat{x}_j)| \right], \tag{J.5.19}
\end{aligned}$$

where the second inequality is due to the Cauchy-Schwarz inequality, and the last equality holds since $\|\mathbf{a}_i - \mathbf{a}_j\|_2^2 = 2(1 - \mathbf{a}_i^\top \mathbf{a}_j)$ by $\|\mathbf{a}_i\|_2 = \|\mathbf{a}_j\|_2 = 1$.

J.6 Proof of Proposition 4.5

We begin with some preliminaries and lemmas used in the proof. For any locally Lipschitz function $\phi : \mathbb{R}^n \rightarrow \mathbb{R}$, its generalized differential at $\mathbf{x} \in \mathbb{R}^n$ is defined by [157]:

$$\partial\phi(\mathbf{x}) = \{ \phi'(\mathbf{x}) : \langle \phi'(\mathbf{x}), \mathbf{d} \rangle \leq D_\phi(\mathbf{x}; \mathbf{d}), \forall \mathbf{d} \in \mathbb{R}^n \}, \tag{J.6.1}$$

where

$$D_\phi(\mathbf{x}; \mathbf{d}) := \limsup_{\mathbf{y} \rightarrow \mathbf{x}, \theta \downarrow 0} \frac{\phi(\mathbf{y} + \theta \mathbf{d}) - \phi(\mathbf{y})}{\theta} \tag{J.6.2}$$

is the generalized directional derivative. When ϕ is convex, $\partial\phi(\mathbf{x})$ is the subdifferential of ϕ at \mathbf{x} . A function ϕ is regular at \mathbf{x} if the directional derivative $D_\phi(\mathbf{x}; \boldsymbol{\zeta})$ exists for all $\boldsymbol{\zeta} \in \mathbb{R}^n$. If this holds for all $\mathbf{x} \in \mathbb{R}^n$, ϕ is a regular function. For a regular function ϕ , $\hat{\mathbf{x}}$ is a stationary point if and only if the directional derivative is nonnegative in all directions *i.e.*, $\mathbf{0} \in \partial\phi(\hat{\mathbf{x}})$. For example, convex functions are regular. The following lemma is used in the proof.

Lemma J.6.1 ([157]). For any locally Lipschitz function $\phi_1, \phi_2 : \mathbb{R}^n \rightarrow \mathbb{R}$, it holds that $\partial(\phi_1 + \phi_2)(\mathbf{x}) \subset \partial\phi_1(\mathbf{x}) + \partial\phi_2(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^n$.

We remark that the reverse inclusion $\partial\phi_1(\mathbf{x}) + \partial\phi_2(\mathbf{x}) \subset \partial(\phi_1 + \phi_2)(\mathbf{x})$ always holds for any convex function $\phi_1, \phi_2 : \mathbb{R}^n \rightarrow \mathbb{R}$.

Proof of Proposition 5. Let $\tilde{\mathbf{y}} := \mathbf{y} + \boldsymbol{\varepsilon}$. Then, by Lemma J.6.1, it holds that

$$\begin{aligned} \mathbf{0} &\in \partial J(\hat{\mathbf{x}}) \\ &= \partial \left(\mu_1 \Phi_{\gamma_1}^{\text{MC}}(\mathbf{A} \cdot -\tilde{\mathbf{y}}) + \mu_2 \Phi_{\gamma_2}^{\text{MC}} + \frac{1}{2} \|\cdot\|_2^2 \right) (\hat{\mathbf{x}}) \\ &\subset \partial \left(\mu_1 \Phi_{\gamma_1}^{\text{MC}}(\mathbf{A} \cdot -\tilde{\mathbf{y}}) + \frac{1}{2} \|\cdot\|_2^2 \right) (\hat{\mathbf{x}}) + \partial(\mu_2 \Phi_{\gamma_2}^{\text{MC}})(\hat{\mathbf{x}}). \end{aligned} \quad (\text{J.6.3})$$

By (J.5.2), the first term of the right side in (J.6.3) becomes

$$\begin{aligned} &\partial \left(\mu_1 \|\mathbf{A} \cdot -\tilde{\mathbf{y}}\|_1 + \mu_1 \gamma_1^{-1} (\|\cdot\|_1^*) (\gamma_1^{-1}(\mathbf{A} \cdot -\tilde{\mathbf{y}})) - \frac{\mu_1}{2\gamma_1} \|\mathbf{A} \cdot -\tilde{\mathbf{y}}\|_2^2 + \frac{1}{2} \|\cdot\|_2^2 \right) (\hat{\mathbf{x}}) \\ &= \mu_1 \partial(\|\mathbf{A} \cdot -\tilde{\mathbf{y}}\|_1)(\hat{\mathbf{x}}) + \mu_1 \partial(\gamma_1^{-1}(\|\cdot\|_1^*)(\gamma_1^{-1}(\mathbf{A} \cdot -\tilde{\mathbf{y}})))(\hat{\mathbf{x}}) + \hat{\mathbf{x}} \\ &\quad - \mu_1 \gamma_1^{-1} \mathbf{A}^\top (\mathbf{A} \hat{\mathbf{x}} - \tilde{\mathbf{y}}). \end{aligned} \quad (\text{J.6.4})$$

Here, the equality is verified by Lemma J.5.1 together with assumption (a) and the convexity of the Moreau envelope [96]. Combining (J.6.3), (J.6.4), and (J.5.5) yields

$$\begin{aligned} \mathbf{0} &\in \mu_1 \mathbf{A}^\top \partial \|\cdot\|_1(\mathbf{A} \hat{\mathbf{x}} - \tilde{\mathbf{y}}) + \mu_1 \gamma_1^{-1} \mathbf{A}^\top \text{Soft}_{\gamma_1}(\mathbf{A} \hat{\mathbf{x}} - \tilde{\mathbf{y}}) \\ &\quad + \hat{\mathbf{x}} - \mu_1 \gamma_1^{-1} \mathbf{A}^\top (\mathbf{A} \hat{\mathbf{x}} - \tilde{\mathbf{y}}) + \partial(\mu_2 \Phi_{\gamma_2}^{\text{MC}})(\hat{\mathbf{x}}). \end{aligned} \quad (\text{J.6.5})$$

Noting that $\phi_{\gamma_2}^{\text{MC}}$ is constant over $[\gamma_2, +\infty)$, assumption (b) implies that $D_{\phi_{\gamma_2}^{\text{MC}}}(\hat{x}_i; d) = D_{\phi_{\gamma_2}^{\text{MC}}}(\hat{x}_j; d) = 0$ for any $d \in \mathbb{R}$, and hence

$$\partial \phi_{\gamma_2}^{\text{MC}}(\hat{x}_i) = \partial \phi_{\gamma_2}^{\text{MC}}(\hat{x}_j) = \{0\}, \quad (\text{J.6.6})$$

from which it follows that

$$\partial(\mu_2 \phi_{\gamma_2}^{\text{MC}})(\hat{x}_i) = \partial(\mu_2 \phi_{\gamma_2}^{\text{MC}})(\hat{x}_j) = \{0\}. \quad (\text{J.6.7})$$

Considering the i th and j th components of both sides of (J.6.5) under (J.6.7), there exists some $\mathbf{s} := [s_1, s_2, \dots, s_m]^\top \in \partial \|\cdot\|_1(\mathbf{A} \hat{\mathbf{x}} - \tilde{\mathbf{y}})$ such that

$$0 = \mu_1 \mathbf{a}_i^\top (\gamma_1^{-1}(\text{Soft}_{\gamma_1} - \text{Id})(\mathbf{A} \hat{\mathbf{x}} - \tilde{\mathbf{y}}) + \mathbf{s}) + \hat{x}_i, \quad (\text{J.6.8})$$

$$0 = \mu_1 \mathbf{a}_j^\top (\gamma_1^{-1}(\text{Soft}_{\gamma_1} - \text{Id})(\mathbf{A} \hat{\mathbf{x}} - \tilde{\mathbf{y}}) + \mathbf{s}) + \hat{x}_j, \quad (\text{J.6.9})$$

from which we obtain

$$\hat{x}_i - \hat{x}_j = \mu_1 (\mathbf{a}_i^\top - \mathbf{a}_j^\top) (\gamma_1^{-1}(\text{Id} - \text{Soft}_{\gamma_1})(\mathbf{A} \hat{\mathbf{x}} - \tilde{\mathbf{y}}) - \mathbf{s}). \quad (\text{J.6.10})$$

Equations (J.6.10) and (J.5.15) yield

$$\begin{aligned} |\hat{x}_i - \hat{x}_j| &= \mu_1 |(\mathbf{a}_i^\top - \mathbf{a}_j^\top) (\gamma_1^{-1}(\text{Id} - \text{Soft}_{\gamma_1})(\mathbf{A} \hat{\mathbf{x}} - \tilde{\mathbf{y}}) - \mathbf{s})| \\ &\leq \mu_1 \|\mathbf{a}_i - \mathbf{a}_j\|_2 \|\gamma_1^{-1}(\text{Id} - \text{Soft}_{\gamma_1})(\mathbf{A} \hat{\mathbf{x}} - \tilde{\mathbf{y}}) - \mathbf{s}\|_2 \\ &\leq \mu_1 \sqrt{2m(1 - \mathbf{a}_i^\top \mathbf{a}_j)}, \end{aligned} \quad (\text{J.6.11})$$

where the first inequality is due to the Cauchy-Schwarz inequality, and the last inequality holds since $\|\mathbf{a}_i - \mathbf{a}_j\|_2^2 = 2(1 - \mathbf{a}_i^\top \mathbf{a}_j)$ by $\|\mathbf{a}_i\|_2 = \|\mathbf{a}_j\|_2 = 1$. \square

J.7 Proof of Proposition 4.6

Proof of Proposition 4.6(a). For any $z = (z_2, \dots, z_{Q+1}) \in \mathcal{Z}$, it holds that

$$\begin{aligned} {}^1(\Psi \circ D^{-1})(z) &= \min_{w \in \mathcal{Z}} \left[\Psi \circ D^{-1}(w) + \frac{1}{2} \|w - z\|_{\mathcal{Z}}^2 \right] \\ &= \min_{w \in \mathcal{Z}} \sum_{i=2}^{Q+1} \left[\nu_i \Psi_i(\nu_i^{-1/2} D_i^{-1} w_i) + \frac{1}{2} \|w_i - z_i\|_{\mathcal{Z}_i}^2 \right] \\ &= \sum_{i=2}^{Q+1} {}^1(\nu_i \Psi_i \circ (\nu_i^{-1/2} D_i^{-1}))(z_i). \end{aligned} \quad (\text{J.7.1})$$

It holds from (J.7.1) that

$${}^1(\Psi \circ D^{-1}) \circ D_{\mathcal{A}} = \sum_{i=2}^{Q+1} {}^1(\nu_i \Psi_i \circ (\nu_i^{-1/2} D_i^{-1})) \circ (\nu_i^{1/2} D_i \mathcal{A}_i), \quad (\text{J.7.2})$$

from which it follows that

$$F = \frac{1}{2} \|\mathcal{A}_1 \cdot\|_{\mathcal{Z}_1}^2 - \sum_{i=2}^{Q+1} {}^1(\nu_i \Psi_i \circ (\nu_i^{-1/2} D_i^{-1})) \circ (\nu_i^{1/2} D_i \mathcal{A}_i). \quad (\text{J.7.3})$$

For the linear operators defined in (4.42) and (4.46), it holds that

$$M^* D^2 M = \sum_{i=2}^{Q+1} \nu_i M_i^* D_i^2 M_i. \quad (\text{J.7.4})$$

Here, the adjoint operator of M is given by $M^* : \mathcal{Z} \rightarrow \mathcal{X} : z \mapsto \sum_{i=2}^{Q+1} M_i^* z_i$ because it holds, for any $x \in \mathcal{X}$ and $z \in \mathcal{Z}$, that

$$\begin{aligned} \langle Mx, z \rangle_{\mathcal{Z}} &= \langle (M_2 x, \dots, M_{Q+1} x), (z_2, \dots, z_{Q+1}) \rangle_{\mathcal{Z}} \\ &= \sum_{i=2}^{Q+1} \langle x, M_i^* z_i \rangle_{\mathcal{X}} = \left\langle x, \sum_{i=2}^{Q+1} M_i^* z_i \right\rangle_{\mathcal{X}}. \end{aligned} \quad (\text{J.7.5})$$

Substituting $M_2 := M$ and $\mathcal{L} := \text{Id}$ into convexity condition (\spadesuit) of LiMES and using (J.7.4) yield (\clubsuit), and hence the sufficiency is verified by Proposition F.2.1(a). \square

The following lemma is used in the proof of Proposition 4.6(b).

Lemma J.7.1. For any finite-dimensional real Hilbert space \mathcal{Z} , let $B_{\mathcal{Z}}(0, r) := \{\zeta \in \mathcal{Z} \mid \|\zeta\|_{\mathcal{Z},*} \leq r\}$ denote a closed ball of radius $r > 0$ centered at $0 \in \mathcal{Z}$. Then, it holds, for any $\nu > 0$, that

$$\nu \sigma_{B_{\mathcal{Z}}(0,1)} = \sigma_{B_{\mathcal{Z}}(0,\nu)}. \quad (\text{J.7.6})$$

Proof of Lemma J.7.1. By $\sigma_{B_{\mathcal{Z}}(0,1)}^* = \iota_{B_{\mathcal{Z}}(0,1)}$, it can be verified, for any $z \in \mathcal{Z}$, that

$$\begin{aligned} (\nu\sigma_{B_{\mathcal{Z}}(0,1)})^*(z) &= \nu\sigma_{B_{\mathcal{Z}}(0,1)}^*(\nu^{-1}z) = \nu\iota_{B_{\mathcal{Z}}(0,1)}(\nu^{-1}z) \\ &= \begin{cases} 0, & \text{if } \nu^{-1}z \in B_{\mathcal{Z}}(0,1) \ (\Leftrightarrow \|z\|_{\mathcal{Z},*} \leq \nu), \\ +\infty, & \text{otherwise} \end{cases} \\ &= \iota_{B_{\mathcal{Z}}(0,\nu)}. \end{aligned} \quad (\text{J.7.7})$$

By (J.7.7), it follows that

$$\nu\sigma_{B_{\mathcal{Z}}(0,1)} = (\nu\sigma_{B_{\mathcal{Z}}(0,1)})^{**} = \iota_{B_{\mathcal{Z}}(0,\nu)}^* = \sigma_{B_{\mathcal{Z}}(0,\nu)}. \quad (\text{J.7.8})$$

□

Proof of Proposition 4.6(b). For each $i = 2, 3, \dots, Q+1$, let $\Psi := \sigma_{B_{\mathcal{Z}_i}(0,\nu_i)} (= \nu_i\|\cdot\|_{\mathcal{Z}_i}$ due to Lemma J.7.1 and $\|\cdot\|_{\mathcal{Z}_i} = \sigma_{B(0,1)})$, $\mathcal{L} := \text{Id}$, $D := \nu_i^{1/2}D_i$, and $\mathcal{A}_2 := \mathcal{A}_i$ in Proposition F.2.1(b.i). Then, regarding the second term of (J.7.3), the following equivalence holds for any $x \in \mathcal{X}$:

$$\begin{aligned} &{}^1(\sigma_{B_{\mathcal{Z}_i}(0,\nu_i)} \circ \nu_i^{-1/2}D_i^{-1})(\nu_i^{1/2}D_i\mathcal{A}_i x) = \frac{\nu_i}{2}\|D_i\mathcal{A}_i x\|_{\mathcal{Z}_i}^2 \\ \Leftrightarrow &{}^1(\sigma_{B_{\mathcal{Z}_i}(0,\nu_i)}^* \circ \nu_i^{1/2}D_i)(\nu_i^{1/2}D_i\mathcal{A}_i x) = 0 \quad (\text{J.7.9}) \\ \Leftrightarrow &x \in K_{B_{\mathcal{Z}_i}(0,\nu_i)} := \{x \in \mathcal{X} \mid \nu_i D_i^2 \mathcal{A}_i x \in B_{\mathcal{Z}_i}(0,\nu_i)\} \\ &= \{x \in \mathcal{X} \mid D_i^2 \mathcal{A}_i x \in B_{\mathcal{Z}_i}(0,1)\} \\ &= K_i. \end{aligned} \quad (\text{J.7.10})$$

By (J.7.3) and (J.7.10), for any x in the neighborhood of some interior point $\hat{x} \in \text{int}\left(\bigcap_{i=2}^{Q+1} K_i\right) \neq \emptyset$, it holds that

$$F(x) = \frac{1}{2}\|\mathcal{A}_1 \hat{x}\|_{\mathcal{Z}}^2 - \sum_{i=2}^{Q+1} \frac{\nu_i}{2}\|D_i \mathcal{A}_i \hat{x}\|_{\mathcal{Z}_i}^2. \quad (\text{J.7.11})$$

This implies that the condition $(\nabla \nabla F(\hat{x}) =) M_1^* M_1 - \sum_{i=2}^{Q+1} \nu_i M_i^* D_i^2 M_i \succeq 0$ is necessary for $F \in \Gamma_0(\mathcal{X})$ by [96, Proposition 17.7]. (Sufficiency has been proven already in (a).)

□

J.8 Proof of (4.51)

It holds from (4.52) that

$$\langle \bar{\zeta}^\perp, \bar{z}_1 \rangle_2 = 0. \quad (\text{J.8.1})$$

Since $\bar{\zeta}^\perp$ and \bar{z}_1 are centered, $\mathbf{z}_2 = \bar{z}_2$ due to (4.53). Hence, it holds that

$$\begin{aligned}\|\bar{z}_2\|_2^2 &= \|\bar{\zeta}^\perp\|_2^2 + \frac{\|\bar{\zeta}^\perp\|_2^2}{\tan^2(\arccos(0.7))} \\ &= \left(1 + \frac{1}{\tan^2(\arccos(0.7))}\right) \|\bar{\zeta}^\perp\|_2^2 \\ &= \frac{1}{\sin^2(\arccos(0.7))} \|\bar{\zeta}^\perp\|_2^2,\end{aligned}\tag{J.8.2}$$

where the first equality is due to (J.8.1). On the other hand, it holds from (4.52) and (4.53) that

$$\begin{aligned}\langle \bar{z}_1, \bar{z}_2 \rangle_2 &= \langle \bar{z}_1, \bar{\zeta}^\perp \rangle_2 + \frac{\|\bar{\zeta}^\perp\|_2}{\tan(\arccos(0.7))} \|\bar{z}_1\|_2 \\ &= \frac{\|\bar{\zeta}^\perp\|_2}{\tan(\arccos(0.7))} \|\bar{z}_1\|_2.\end{aligned}\tag{J.8.3}$$

Hence, it holds from (J.8.2) and (J.8.3) that

$$\begin{aligned}\frac{|\bar{z}_1^\top \bar{z}_2|}{\|\bar{z}_1\|_2 \|\bar{z}_2\|_2} &= \frac{1}{\tan(\arccos(0.7))} / \frac{1}{\sin(\arccos(0.7))} \\ &= \cos(\arccos(0.7)) \\ &= 0.7.\end{aligned}\tag{J.8.4}$$

Appendix K

Proof for Chapter 5

K.1 Proof of Proposition 5.1

The following lemmas are used in the proof.

Lemma K.1.1. For any $n \in \mathbb{N}^*$, $\lambda_1, \lambda_2 > 0$, and $\mathbf{x} \in \mathbb{R}^n$, it holds that

$$2\gamma^{-1}(\Omega_{\lambda_1, \lambda_2}^{\text{OSCAR}})_{\gamma^{-1/2}\mathbf{I}_n}(\mathbf{x}) = \|\mathbf{w}\|_2^2 - \|\mathbf{w} - \gamma^{-1}(|\mathbf{x}|_{\downarrow} - P_{\mathcal{K}_{\geq 0}^n}(|\mathbf{x}|_{\downarrow} - \gamma\mathbf{w}))\|_2^2. \quad (\text{K.1.1})$$

Proof. By definition of the Moreau envelope, it holds that

$$\gamma\Omega_{\lambda_1, \lambda_2}^{\text{OSCAR}}(\mathbf{x}) = \Omega_{\lambda_1, \lambda_2}^{\text{OSCAR}}(\text{Prox}_{\gamma\Omega_{\lambda_1, \lambda_2}^{\text{OSCAR}}}(\mathbf{x})) + \frac{1}{2\gamma}\|\mathbf{x} - \text{Prox}_{\gamma\Omega_{\lambda_1, \lambda_2}^{\text{OSCAR}}}(\mathbf{x})\|_2^2. \quad (\text{K.1.2})$$

By (2.70) and (2.72), the first term of the right-hand side of (K.1.2) can be expressed as

$$\begin{aligned} \Omega_{\lambda_1, \lambda_2}^{\text{OSCAR}}(\text{Prox}_{\gamma\Omega_{\lambda_1, \lambda_2}^{\text{OSCAR}}}(\mathbf{x})) &= \left\langle \mathbf{w}, \left| \text{Prox}_{\gamma\Omega_{\lambda_1, \lambda_2}^{\text{OSCAR}}} \Big|_{\downarrow} \right\rangle_2 \\ &= \langle \mathbf{w}, P_{\mathcal{K}_{\geq 0}^n}(|\mathbf{x}|_{\downarrow} - \gamma\mathbf{w}) \rangle_2, \end{aligned} \quad (\text{K.1.3})$$

where the second equality is due to

$$\left| \text{Sign}(\mathbf{x}) \odot \mathbf{P}(|\mathbf{x}|)^{\text{T}} P_{\mathcal{K}_{\geq 0}^n}(|\mathbf{x}|_{\downarrow} - \mathbf{w}) \Big|_{\downarrow} = P_{\mathcal{K}_{\geq 0}^n}(|\mathbf{x}|_{\downarrow} - \mathbf{w}). \quad (\text{K.1.4})$$

In a similar way, by (2.72), the second term of the right-hand side of (K.1.2) can be expressed as

$$\begin{aligned} &\frac{1}{2\gamma}\|\mathbf{x} - \text{Prox}_{\gamma\Omega_{\lambda_1, \lambda_2}^{\text{OSCAR}}}(\mathbf{x})\|_2^2 \\ &= \frac{1}{2\gamma}\left\| \text{Sign}(\mathbf{x}) \odot \mathbf{P}(|\mathbf{x}|)^{\text{T}} \left(|\mathbf{x}|_{\downarrow} - P_{\mathcal{K}_{\geq 0}^n}(|\mathbf{x}|_{\downarrow} - \gamma\mathbf{w}) \right) \right\|_2^2 \\ &= \frac{1}{2\gamma}\| |\mathbf{x}|_{\downarrow} - P_{\mathcal{K}_{\geq 0}^n}(|\mathbf{x}|_{\downarrow} - \gamma\mathbf{w}) \|_2^2. \end{aligned} \quad (\text{K.1.5})$$

Hence, substituting (K.1.3) and (K.1.5) into (K.1.2) yields that

$$\gamma \Omega_{\lambda_1, \lambda_2}^{\text{OSCAR}}(\mathbf{x}) = \langle \mathbf{w}, P_{\mathcal{K}_{\geq 0}^n}(|\mathbf{x}|_{\downarrow} - \gamma \mathbf{w}) \rangle_2 + \frac{1}{2\gamma} \| |\mathbf{x}|_{\downarrow} - P_{\mathcal{K}_{\geq 0}^n}(|\mathbf{x}|_{\downarrow} - \gamma \mathbf{w}) \|_2^2. \quad (\text{K.1.6})$$

Hence, it holds from (2.70) and (K.1.6) that

$$\begin{aligned} (\Omega_{\lambda_1, \lambda_2}^{\text{OSCAR}})_{\gamma^{-1/2} \mathbf{I}_n}(\mathbf{x}) &= \Omega_{\lambda_1, \lambda_2}^{\text{OSCAR}}(\mathbf{x}) - \gamma \Omega_{\lambda_1, \lambda_2}^{\text{OSCAR}}(\mathbf{x}) \\ &= \langle \mathbf{w}, |\mathbf{x}|_{\downarrow} - P_{\mathcal{K}_{\geq 0}^n}(|\mathbf{x}|_{\downarrow} - \gamma \mathbf{w}) \rangle_2 \\ &\quad - \frac{1}{2\gamma} \| |\mathbf{x}|_{\downarrow} - P_{\mathcal{K}_{\geq 0}^n}(|\mathbf{x}|_{\downarrow} - \gamma \mathbf{w}) \|_2^2, \end{aligned} \quad (\text{K.1.7})$$

which is equivalent to (K.1.1). \square

Lemma K.1.2. Let $\mathbf{d} \in \mathbb{R}^n$ such that $d_1 \leq d_2 \leq \dots \leq d_n$. Then,

$$\sum_{i=1}^n \left(d_i - \max \left\{ \frac{1}{n} \sum_{j=1}^n d_j, 0 \right\} \right)^2 \leq \sum_{i=1}^n (d_i - z_i)^2, \quad \forall \mathbf{z} \in \mathcal{K}_{\geq 0}^n. \quad (\text{K.1.8})$$

In other words,

$$P_{\mathcal{K}_{\geq 0}^n}(\mathbf{d}) = \begin{cases} \left(\frac{1}{n} \sum_{j=1}^n d_j \right) \mathbf{1}_n, & \text{if } \sum_{j=1}^n d_j \geq 0, \\ \mathbf{0}_n, & \text{otherwise.} \end{cases} \quad (\text{K.1.9})$$

Proof. Let $\mathbf{w} \in \mathcal{K}_{\geq 0}^n$ such that

$$\sum_{i=1}^n (d_i - w_i)^2 \leq \sum_{i=1}^n (d_i - z_i)^2, \quad \forall \mathbf{z} \in \mathcal{K}_{\geq 0}^n. \quad (\text{K.1.10})$$

Suppose that there exists $i \in \{1, 2, \dots, n-1\}$ such that $w_i > w_{i+1}$. Then, it holds that

$$\begin{aligned} & (d_i - w_i)^2 + (d_{i+1} - w_{i+1})^2 - \left(d_i - \frac{w_i + w_{i+1}}{2} \right)^2 - \left(d_{i+1} - \frac{w_i + w_{i+1}}{2} \right)^2 \\ &= -2d_i w_i + w_i^2 - 2d_{i+1} w_{i+1} + w_{i+1}^2 + d_i(w_i + w_{i+1}) - 2 \left(\frac{w_i + w_{i+1}}{2} \right)^2 \\ &\quad + d_{i+1}(w_i + w_{i+1}) \\ &> (d_{i+1} - d_i)(w_i - w_{i+1}) \\ &\geq 0, \end{aligned} \quad (\text{K.1.11})$$

where the first inequality is due to the strong convexity of the quadratic function. Hence, it holds that

$$\begin{aligned}
\sum_{k=1}^n (d_k - w_k)^2 &= \sum_{k \neq i, i+1} (d_k - w_k)^2 + (d_i - w_i)^2 + (d_{i+1} - w_i)^2 \\
&> \sum_{k \neq i, i+1} (d_k - w_k)^2 + \left(d_i - \frac{w_i + w_{i+1}}{2} \right)^2 \\
&\quad + \left(d_{i+1} - \frac{w_i + w_{i+1}}{2} \right)^2 \\
&= \sum_{k=1}^n (d_k - \tilde{w}_k)^2, \tag{K.1.12}
\end{aligned}$$

where, for $k = 1, 2, \dots, n$,

$$\tilde{w}_k := \begin{cases} w_k, & \text{if } k \notin \{i, i+1\}, \\ \frac{w_i + w_{i+1}}{2}, & \text{if } k \in \{i, i+1\}. \end{cases} \tag{K.1.13}$$

This contradicts (K.1.10) due to $\tilde{w} \in \mathcal{K}_{\geq 0}^n$. Hence, it holds that $w_1 = w_2 = \dots = w_n =: w$. It holds for any $z \geq 0$ that

$$\begin{aligned}
\sum_{i=1}^n (d_i - z)^2 &= nz^2 - 2 \sum_{i=1}^n d_i z + \sum_{i=1}^n d_i^2 \\
&= n \left(z - \frac{1}{n} \sum_{i=1}^n d_i \right)^2 - n \left(\frac{1}{n} \sum_{i=1}^n d_i \right)^2 + \sum_{i=1}^n d_i^2, \tag{K.1.14}
\end{aligned}$$

which is minimized at

$$z = \begin{cases} \frac{1}{n} \sum_{i=1}^n d_i, & \text{if } \frac{1}{n} \sum_{i=1}^n d_i \geq 0, \\ 0, & \text{if } \frac{1}{n} \sum_{i=1}^n d_i < 0. \end{cases} \tag{K.1.15}$$

Hence, it holds that

$$w = \max \left\{ \frac{1}{n} \sum_{i=1}^n d_i, 0 \right\}. \tag{K.1.16}$$

□

Proof of Proposition 5.1.

(I) Case of $\mathbf{x} = \mathbf{0}$ (case 4).

In this case, it holds from Lemma K.1.1 that

$$2\gamma^{-1}(\Omega_{\lambda_1, \lambda_2}^{\text{OSCAR}})_{\gamma^{-1/2}\mathbf{I}_n}(\mathbf{x}) = \|\mathbf{w}\|_2^2 - \|\mathbf{w}\|_2^2 = 0. \quad (\text{K.1.17})$$

(II) Case of $\mathbf{x} \in \mathbb{R}_{++}^n \cap \mathcal{K}_{>}^n$ (case 1).

In this case, there exists a sufficiently small $\gamma > 0$ such that

$$x_1 - \gamma w_1 > x_2 - \gamma w_2 > \dots > x_n - \gamma w_n > 0 \Leftrightarrow \mathbf{x} - \gamma \mathbf{w} \in \mathcal{K}_{\geq 0}^n. \quad (\text{K.1.18})$$

Hence, it holds that

$$P_{\mathcal{K}_{\geq 0}^n}(\mathbf{x} - \gamma \mathbf{w}) = \mathbf{x} - \gamma \mathbf{w}, \quad (\text{K.1.19})$$

from which and Lemma K.1.1 it follows that

$$\begin{aligned} 2\gamma^{-1}(\Omega_{\lambda_1, \lambda_2}^{\text{OSCAR}})_{\gamma^{-1/2}\mathbf{I}_n}(\mathbf{x}) &= \|\mathbf{w}\|_2^2 - \|\mathbf{w} - \gamma^{-1}(\mathbf{x} - (\mathbf{x} - \gamma \mathbf{w}))\|_2^2 \\ &= \|\mathbf{w}\|_2^2. \end{aligned} \quad (\text{K.1.20})$$

(III) Case of $\mathbf{x} \in (\mathbb{R}_{++}^n)^c \cap \mathcal{K}_{>}^n$ (case 3a).

In this case, since $x_n = 0$, there exists a sufficiently small $\gamma > 0$ such that

$$x_1 - \gamma w_1 > x_2 - \gamma w_2 > \dots > x_{n-1} - \gamma w_{n-1} > 0 > x_n - \gamma w_n. \quad (\text{K.1.21})$$

Hence, it holds for $i = 1, 2, \dots, n$ that

$$[P_{\mathcal{K}_{\geq 0}^n}(\mathbf{x} - \gamma \mathbf{w})]_i = \begin{cases} x_i - \gamma w_i, & \text{if } i = 1, 2, \dots, n-1, \\ 0, & \text{if } i = n, \end{cases} \quad (\text{K.1.22})$$

from which and Lemma K.1.1 it follows that

$$\begin{aligned} &2\gamma^{-1}(\Omega_{\lambda_1, \lambda_2}^{\text{OSCAR}})_{\gamma^{-1/2}\mathbf{I}_n}(\mathbf{x}) \\ &= \|\mathbf{w}\|_2^2 - \left\| \mathbf{w} - \gamma^{-1} \left(\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n-1} \\ 0 \end{bmatrix} - \begin{bmatrix} x_1 - \gamma w_1 \\ x_2 - \gamma w_2 \\ \vdots \\ x_{n-1} - \gamma w_{n-1} \\ 0 \end{bmatrix} \right) \right\|_2^2 \\ &= \|\mathbf{w}\|_2^2 - w_n^2. \end{aligned} \quad (\text{K.1.23})$$

(IV) Case of $\mathbf{x} \in \mathcal{K}_{\geq 0}^n \cap (\mathcal{K}_{>}^n \cup \{\mathbf{0}\})^c$ (cases 2 and 3b).

In this case, it holds that $q \geq 1$. Let $\boldsymbol{\xi} \in \mathbb{R}^n$ such that

$$\xi_i := \begin{cases} \max \left\{ x_i - \frac{\gamma \sum_{j \in S_l} w_j}{\text{card}(S_l)}, 0 \right\}, & \text{if } i \in S_l \text{ for some } l \in \{1, 2, \dots, q\}, \\ 0, & \text{if } i = n \text{ and } x_n = 0, \\ x_i - \gamma w_i, & \text{otherwise,} \end{cases} \quad (\text{K.1.24})$$

for $i = 1, 2, \dots, n$. Since $x_i = 0$ implies that $i \in S_q$ or “ $i = n$ and $x_n = 0$ ”, it holds that $x_i > 0$ for all i in the third case of (K.1.24). Hence, there exists a sufficiently small $\gamma > 0$ such that

$$\xi_i \geq 0, \quad \forall i = 1, 2, \dots, n. \quad (\text{K.1.25})$$

In what follows, we first show that

$$\boldsymbol{\xi} \in \mathcal{K}_{\geq 0}^n \quad (\text{K.1.26})$$

for cases 2 and 3b. We then show that

$$P_{\mathcal{K}_{\geq 0}^n}(\mathbf{x} - \gamma \mathbf{w}) = \boldsymbol{\xi} \quad (\text{K.1.27})$$

for each case individually.

Fix $i \in \{1, 2, \dots, n-1\}$ arbitrarily. By (K.1.25), it suffices to show that $\xi_i \geq \xi_{i+1}$ for verifying (K.1.26). We first consider the case that $\xi_i > 0$ and $x_i > x_{i+1}$. Then, by (K.1.24), it holds that

$$\xi_i \geq \min \left\{ x_i - \frac{\gamma \sum_{j \in S_l} w_j}{\text{card}(S_l)}, x_i - \gamma w_i \right\}, \quad \text{and} \quad (\text{K.1.28})$$

$$\xi_{i+1} \leq x_{i+1}. \quad (\text{K.1.29})$$

Hence, it holds that

$$\begin{aligned} \xi_i - \xi_{i+1} &\geq \min \left\{ x_i - \frac{\gamma \sum_{j \in S_l} w_j}{\text{card}(S_l)}, x_i - \gamma w_i \right\} - x_{i+1} \\ &= x_i - x_{i+1} - \gamma \max \left\{ \frac{\sum_{j \in S_l} w_j}{\text{card}(S_l)}, w_i \right\} \\ &> 0, \quad \text{for some } \gamma > 0, \end{aligned} \quad (\text{K.1.30})$$

where the last inequality is due to $x_i - x_{i+1} > 0$.

Assume now that $\xi_i = 0$. Then, by (K.1.24), it holds that $x_i = 0$ for a sufficiently small $\gamma > 0$. Hence, $\xi_i = \xi_{i+1} = 0$ since $x_i = x_{i+1} = 0$, *i.e.*, $x_i, x_{i+1} \in S_q$. Assume, on the other hand, that $x_i = x_{i+1}$, *i.e.*, $x_i, x_{i+1} \in S_l$. Then, it holds from (K.1.24) that $\xi_i = \xi_{i+1}$. Hence, (K.1.26) holds.

Now we prove (K.1.27). For $l = 1, 2, \dots, q$, let $j_{S_l}^{\min} := \min_{j \in S_l} j$ and $j_{S_l}^{\max} := \max_{j \in S_l} j$. Then, there exists a sufficiently small $\gamma > 0$ such that

$$x_{j_{S_l}^{\min}} - \gamma w_{i_{j_{S_l}^{\min}}} \leq x_{j_{S_l}^{\min}+1} - \gamma w_{j_{S_l}^{\min}+1} \leq \dots \leq x_{j_{S_l}^{\max}} - \gamma w_{j_{S_l}^{\max}}. \quad (\text{K.1.31})$$

Hence, for any $\{z_j\}_{j=j_{S_l}^{\min}}^{j_{S_l}^{\max}} \subset \mathbb{R}^{\text{card}(S_l)}$ such that $z_{j_{S_l}^{\min}} \geq z_{j_{S_l}^{\min}+1} \geq \dots \geq z_{j_{S_l}^{\max}} \geq 0$, it holds due to Lemma K.1.2 that

$$\sum_{j \in S_l} ((x_j - \gamma w_j) - z_j)^2 \geq \sum_{j \in S_l} \left((x_j - \gamma w_j) - \max \left\{ x_j - \frac{\gamma \sum_{k \in S_l} w_k}{\text{card}(S_l)}, 0 \right\} \right)^2. \quad (\text{K.1.32})$$

(i) Case of $\mathbf{x} \in \mathbb{R}_{++}^n \cap (\mathcal{K}_{>}^n)^c$ (case 2).

In this case, due to $\mathbf{x} \in \mathbb{R}_{++}^n$, it holds from (K.1.24) that

$$\xi_i := \begin{cases} x_i - \frac{\gamma \sum_{j \in S_l} w_j}{\text{card}(S_l)}, & \text{if } i \in S_l \text{ for some } l \in \{1, 2, \dots, q\}, \\ x_i - \gamma w_i, & \text{otherwise,} \end{cases} \quad (\text{K.1.33})$$

for $i = 1, 2, \dots, n$. For $i \notin \cup_{l=1}^q S_l$, it holds that

$$x_i - \gamma w_i - \xi_i = 0. \quad (\text{K.1.34})$$

Hence, it holds for any $\mathbf{z} \in \mathcal{K}_{\geq 0}^n$ that

$$\begin{aligned} \sum_{i=1}^n ((x_i - \gamma w_i) - z_i)^2 &\geq \sum_{l=1}^q \sum_{j \in S_l} ((x_j - \gamma w_j) - z_j)^2 \\ &\geq \sum_{l=1}^q \sum_{j \in S_l} \left((x_j - \gamma w_j) - \max \left\{ x_j - \frac{\gamma \sum_{k \in S_l} w_k}{\text{card}(S_l)}, 0 \right\} \right)^2 \\ &= \sum_{l=1}^q \sum_{j \in S_l} \left((x_j - \gamma w_j) - \left(x_j - \frac{\gamma \sum_{k \in S_l} w_k}{\text{card}(S_l)} \right) \right)^2 \\ &= \sum_{i=1}^n ((x_i - \gamma w_i) - \xi_i)^2, \end{aligned} \quad (\text{K.1.35})$$

where the second inequality is due to (K.1.32), and the first equality is due to $\mathbf{x} \in \mathbb{R}_{++}^n$, and the last equality is due to (K.1.33) and (K.1.34). Hence,

(K.1.27) holds. Thus, it holds for any $\mathbf{x} \in \mathcal{K}_{\geq 0}^n$ that

$$\begin{aligned}
& 2\gamma^{-1}(\Omega_{\lambda_1, \lambda_2}^{\text{OSCAR}})_{\gamma^{-1/2}\mathbf{I}_n}(\mathbf{x}) \\
&= \|\mathbf{w}\|_2^2 - \|\mathbf{w} - \gamma^{-1}(\mathbf{x} - \boldsymbol{\xi})\|_2^2 \\
&= \|\mathbf{w}\|_2^2 - \sum_{i=1}^n (w_i - \gamma^{-1}(x_i - \xi_i))^2 \\
&= \|\mathbf{w}\|_2^2 - \sum_{l=1}^q \sum_{j \in S_l} \left(w_j - \gamma^{-1} \left(x_j - \left(x_j - \frac{\gamma \sum_{k \in S_l} w_k}{\text{card}(S_l)} \right) \right) \right)^2 \\
&= \|\mathbf{w}\|_2^2 - \sum_{l=1}^q \sum_{j \in S_l} \left(w_j - \frac{\sum_{k \in S_l} w_k}{\text{card}(S_l)} \right)^2, \tag{K.1.36}
\end{aligned}$$

where the first equality is due to Lemma K.1.1 and (K.1.27), and the third equality is due to (K.1.33) and (K.1.34).

(ii) Case of $\mathbf{x} \in \mathcal{K}_{\geq 0}^n \cap (\mathbb{R}_{++}^n \cup \mathcal{K}_{>}^n \cup \{\mathbf{0}\})^c$ (case 3b).

In this case, $x_n = 0$. Assume first that $x_{n-1} > 0$. Then, it holds that $n \notin S_q$ since $x_{n-1} > x_n = 0$. Hence, it holds from (K.1.24) that

$$\xi_i := \begin{cases} x_i - \frac{\gamma \sum_{j \in S_l} w_j}{\text{card}(S_l)}, & \text{if } k \in S_l \text{ for some } l \in \{1, 2, \dots, q\}, \\ 0, & \text{if } k = n, \\ x_i - \gamma w_i, & \text{otherwise,} \end{cases} \tag{K.1.37}$$

for $i = 1, 2, \dots, n$. By (K.1.37), it holds that

$$x_n - \gamma w_n - \xi_n = -\gamma w_n. \tag{K.1.38}$$

Hence, it holds for any $\mathbf{z} \in \mathcal{K}_{\geq 0}^n$ that

$$\begin{aligned}
& \sum_{i=1}^n ((x_i - \gamma w_i) - z_i)^2 \\
& \geq \sum_{l=1}^q \sum_{j \in S_l} ((x_j - \gamma w_j) - z_j)^2 + (\gamma w_n + z_n)^2 \\
& \geq \sum_{l=1}^q \sum_{j \in S_l} \left((x_j - \gamma w_j) - \max \left\{ x_j - \frac{\gamma \sum_{k \in S_l} w_k}{\text{card}(S_l)}, 0 \right\} \right)^2 + (\gamma w_n)^2 \\
& = \sum_{l=1}^q \sum_{j \in S_l} \left((x_j - \gamma w_j) - \left(x_j - \frac{\gamma \sum_{k \in S_l} w_k}{\text{card}(S_l)} \right) \right)^2 + (\gamma w_n)^2 \\
& = \sum_{i=1}^n ((x_i - \gamma w_i) - \xi_i)^2, \tag{K.1.39}
\end{aligned}$$

where the first inequality is due to $x_n = 0$, the second inequality is due to (K.1.32), the first equality is due to $x_n > 0$, and the last equality is due to (K.1.34), (K.1.37), and (K.1.38). Hence, (K.1.27) holds. Thus, in a similar way to (K.1.36), it holds for any $\mathbf{x} \in \mathcal{K}_{\geq 0}^n$ that

$$\begin{aligned}
& 2\gamma^{-1}(\Omega_{\lambda_1, \lambda_2}^{\text{OSCAR}})_{\gamma^{-1/2} \mathbf{I}_n}(\mathbf{x}) \\
&= \|\mathbf{w}\|_2^2 - \|\mathbf{w} - \gamma^{-1}(\mathbf{x} - \boldsymbol{\xi})\|_2^2 \\
&= \|\mathbf{w}\|_2^2 - \sum_{i=1}^n (w_i - \gamma^{-1}(x_i - \xi_i))^2 \\
&= \|\mathbf{w}\|_2^2 - \sum_{l=1}^q \sum_{j \in S_l} \left(w_j - \gamma^{-1} \left(x_j - \left(x_j - \frac{\gamma \sum_{k \in S_l} w_k}{\text{card}(S_l)} \right) \right) \right)^2 - w_n^2 \\
&= \|\mathbf{w}\|_2^2 - \sum_{l=1}^q \sum_{j \in S_l} \left(w_j - \frac{\sum_{k \in S_l} w_k}{\text{card}(S_l)} \right)^2 - w_n^2, \tag{K.1.40}
\end{aligned}$$

where the first equality is due to Lemma K.1.1 and (K.1.27), and the third equality is due to (K.1.34), (K.1.37), and (K.1.38).

Now we assume that $x_{n-1} = 0$. Then, it holds from (K.1.24) that

$$\xi_i := \begin{cases} x_i - \frac{\gamma \sum_{j \in S_l} w_j}{\text{card}(S_l)}, & \text{if } k \in S_l \text{ for some } l \in \{1, 2, \dots, q-1\}, \\ 0, & \text{if } k \in S_q, \\ x_i - \gamma w_i, & \text{otherwise,} \end{cases} \tag{K.1.41}$$

for $i = 1, 2, \dots, n$. It holds for any $\mathbf{z} \in \mathcal{K}_{\geq 0}^n$ that

$$\begin{aligned}
\sum_{i=1}^n ((x_i - \gamma w_i) - z_i)^2 &\geq \sum_{l=1}^q \sum_{j \in S_l} ((x_j - \gamma w_j) - z_j)^2 \\
&\geq \sum_{l=1}^q \sum_{j \in S_l} \left((x_j - \gamma w_j) - \max \left\{ x_j - \frac{\gamma \sum_{k \in S_l} w_k}{\text{card}(S_l)}, 0 \right\} \right)^2 \\
&= \sum_{l=1}^{q-1} \sum_{j \in S_l} \left((x_j - \gamma w_j) - \left(x_j - \frac{\gamma \sum_{k \in S_l} w_k}{\text{card}(S_l)} \right) \right)^2 \\
&\quad + \sum_{j \in S_q} (x_j - \gamma w_j)^2 \\
&= \sum_{i=1}^n ((x_i - \gamma w_i) - \xi_i)^2, \tag{K.1.42}
\end{aligned}$$

where the first inequality is due to (K.1.32), the first equality is due to $x_i > 0$ for $i \in \cup_{l=1}^{q-1} S_l$, and the last equality is due to (K.1.34), (K.1.37), and

(K.1.41). Hence, (K.1.27) holds. Thus, in a similar way to (K.1.36), it holds for any $\mathbf{x} \in \mathcal{K}_{\geq 0}^n$ that

$$\begin{aligned}
& 2\gamma^{-1}(\Omega_{\lambda_1, \lambda_2}^{\text{OSCAR}})_{\gamma^{-1/2}\mathbf{I}_n}(\mathbf{x}) \\
&= \|\mathbf{w}\|_2^2 - \|\mathbf{w} - \gamma^{-1}(\mathbf{x} - \boldsymbol{\xi})\|_2^2 \\
&= \|\mathbf{w}\|_2^2 - \sum_{i=1}^n (w_i - \gamma^{-1}(x_i - \xi_i))^2 \\
&= \|\mathbf{w}\|_2^2 - \sum_{l=1}^{q-1} \sum_{j \in S_l} \left(w_j - \gamma^{-1} \left(x_j - \left(x_j - \frac{\gamma \sum_{k \in S_l} w_k}{\text{card}(S_l)} \right) \right) \right)^2 - \sum_{j \in S_q} w_j^2 \\
&= \|\mathbf{w}\|_2^2 - \sum_{l=1}^{q-1} \sum_{j \in S_l} \left(w_j - \frac{\sum_{k \in S_l} w_k}{\text{card}(S_l)} \right)^2 - \sum_{j \in S_q} w_j^2. \tag{K.1.43}
\end{aligned}$$

The expressions (K.1.40) and (K.1.43) can be unified into the following form:

$$\begin{aligned}
& 2\gamma^{-1}(\Omega_{\lambda_1, \lambda_2}^{\text{OSCAR}})_{\gamma^{-1/2}\mathbf{I}_n}(\mathbf{x}) \\
&= \|\mathbf{w}\|_2^2 - \sum_{l=1}^q \sum_{j \in S_l} \left(w_j - \frac{\sum_{k \in S_l} w_k}{\text{card}(S_l)} \chi_{\mathbb{R}_{++}}(x_j) \right)^2 - w_n^2 \chi_{\mathbb{R}_{++}}(x_{n-1}). \tag{K.1.44}
\end{aligned}$$

□

K.2 Proof of Proposition 5.7

The following fact will be used in the proof.

Fact K.2.1. [100, Theorem 2] Let $f \in \Gamma_0(\mathbb{R}^n)$ be smooth and ρ -strongly convex with κ -Lipschitz continuous gradient operator ∇f for $\kappa, \rho > 0$ such that $\kappa > \rho$. Let $\mu > 0$ such that $\mu < 2/(\kappa + \rho)$, which implies $\mu < 1/\rho$. Assume that (i) $T = \nabla \psi$ for some Fréchet differentiable convex function $\psi \in \Gamma_0(\mathbb{R}^n)$, and (ii) T is β^{-1} -Lipschitz continuous for $\beta := 1 - \mu\rho \in (0, 1)$, or equivalently T is β -cocoercive. Suppose that $\mu f + \varphi$ has a minimizer in \mathbb{R}^n . Then, for an arbitrary $x_0 \in \mathbb{R}^n$, the sequence $(x_k)_{k \in \mathbb{N}} \subset \mathbb{R}^n$ generated by

$$x_{k+1} := T(x_k - \mu \nabla f(x_k)), k \in \mathbb{N} \tag{K.2.1}$$

converges weakly to a minimizer \hat{x} of $\mu f + \varphi$; i.e., $\lim_{k \rightarrow \infty} \langle x_k - \hat{x}, y \rangle = 0$ for every fixed $y \in \mathbb{R}^n$.

Proof of Proposition 5.7. By assumption, $f + \rho \|P_{\mathcal{M}^\perp} \cdot\|_2^2/2$ is smooth and ρ -strongly convex with κ -Lipschitz continuous gradient. By Proposition 5.6, $\Delta_\omega = \nabla \psi_\omega$ is ω -Lipschitz continuous. Hence, the convergence is guaranteed by Fact K.2.1. □

K.3 Proof of Proposition 5.8

(a) By assumption assumption (C-2), it holds that

$$\begin{aligned} \text{Prox}_{g_1} - \text{Prox}_{g_2} &= \text{Prox}_{g_1} - \text{Prox}_{g_2-g_1} \circ \text{Prox}_{g_1} \\ &= (\text{Id} - \text{Prox}_{g_2-g_1}) \circ \text{Prox}_{g_1} \\ &= \text{Prox}_{(g_2-g_1)^*} \circ \text{Prox}_{g_1}, \end{aligned} \quad (\text{K.3.1})$$

where the third equality is due to Fact 2.4(c) under assumption (C-1). It holds by definition of Δ_ω given in (5.1) that

$$\begin{aligned} \Delta_\omega &= \omega(\text{Prox}_{g_1} - \text{Prox}_{g_2}) + \text{Prox}_{g_2} \\ &= \omega \text{Prox}_{(g_2-g_1)^*} \circ \text{Prox}_{g_1} + \text{Prox}_{g_2}. \end{aligned} \quad (\text{K.3.2})$$

The operator Prox_{g_2} is firmly nonexpansive [96, Proposition 12.28], and hence it is monotone by (2.33). By definition of monotone operator (2.32), $\alpha_1 B_1 + \alpha_2 B_2$ is monotone for any $\alpha_1, \alpha_2 \geq 0$ if $B_1, B_2 : \mathbb{R}^n \rightarrow \mathbb{R}^n$ are monotone. Hence, Δ_ω ($= \nabla \psi_\omega$ by Proposition 5.6) is monotone due to (K.3.2) under $\omega > 1$ and assumption (C-3), which implies that ψ_ω is convex [96, Proposition 17.7].

(b) (i) \Rightarrow (C-3): Since $(g_2 - g_1)^* + g_1 \in \Gamma_0(\mathbb{R}^n)$, $\text{Prox}_{(g_2-g_1)^*+g_1}$ is firmly nonexpansive, and hence it is monotone.

(ii) \Rightarrow (C-3): Since g_1, g_2 are separable, the corresponding proximity operators are also separable [96, Proposition 24.11], it holds for any $\mathbf{x} \in \mathbb{R}^n$ that

$$\text{Prox}_{g_1}(\mathbf{x}) = (\text{Prox}_{\tilde{g}_1}(x_i))_{i=1}^n, \quad (\text{K.3.3})$$

and

$$\begin{aligned} \text{Prox}_{(g_2-g_1)^*}(\mathbf{x}) &= \mathbf{x} - \text{Prox}_{g_2-g_1}(\mathbf{x}) \\ &= (x_i - \text{Prox}_{\tilde{g}_2-\tilde{g}_1}(x_i))_{i=1}^n \\ &= (\text{Prox}_{(\tilde{g}_2-\tilde{g}_1)^*}(x_i))_{i=1}^n. \end{aligned} \quad (\text{K.3.4})$$

Hence, we obtain

$$\text{Prox}_{(g_2-g_1)^*} \circ \text{Prox}_{g_1}(\mathbf{x}) = (\text{Prox}_{(\tilde{g}_2-\tilde{g}_1)^*}(\text{Prox}_{\tilde{g}_1}(x_i)))_{i=1}^n, \quad (\text{K.3.5})$$

implies which, together with Proposition H.2 in Appendix H, that there exists $\Gamma_0(\mathbb{R}^n) \ni h : \mathbf{x} \mapsto \sum_{i=1}^n \tilde{h}(x_i)$, where $\tilde{h} \in \Gamma_0(\mathbb{R})$ such that $\text{Prox}_{(\tilde{g}_2-\tilde{g}_1)^*} \circ \text{Prox}_{\tilde{g}_1} = \text{Prox}_{\tilde{h}}$, namely

$$\text{Prox}_{(g_2-g_1)^*} \circ \text{Prox}_{g_1}(\mathbf{x}) = \text{Prox}_h(\mathbf{x}). \quad (\text{K.3.6})$$

Since Prox_h is firmly nonexpansive, it is monotone.

K.4 Proof of Proposition 5.9

In the following, we shall represent ψ_ω in two different ways (see (K.4.3) and (K.4.4) below), and we then obtain $\psi_\omega^* \circ \Delta_\omega$ (see (K.4.7) below) by the combinations of two expressions of ψ_ω and the Fenchel-Young inequality.

Fix a vector $\mathbf{x} \in \mathbb{R}^n$ arbitrarily. Since the conjugate of the support function is the indicator function (see Section 2.1.2), it holds that

$$\begin{aligned}
 {}^1(\sigma_{C_1}^*)(\mathbf{x}) &= {}^1(\iota_{C_1})(\mathbf{x}) \\
 &= \iota_{C_1}(\text{Prox}_{\iota_{C_1}}(\mathbf{x})) + \frac{1}{2}\|\mathbf{x} - \text{Prox}_{\iota_{C_1}}(\mathbf{x})\|_2^2 \\
 &= \iota_{C_1}(P_{C_1}(\mathbf{x})) + \frac{1}{2}\|\text{Prox}_{\sigma_{C_1}}(\mathbf{x})\|_2^2 \\
 &= \frac{1}{2}\|\text{Prox}_{\sigma_{C_1}}(\mathbf{x})\|_2^2, \tag{K.4.1}
 \end{aligned}$$

where the second equality is due to the definition of the Moreau envelope (2.21), the third equality is due to Fact 2.4(c) and the fact that the proximity operator of the indicator function corresponds to the projection operator [96, Example 12.25]. In the same way, we have

$${}^1(\sigma_{C_2}^*)(\mathbf{x}) = \frac{1}{2}\|\text{Prox}_{\sigma_{C_2}}(\mathbf{x})\|_2^2. \tag{K.4.2}$$

Hence, substituting (K.4.1) and (K.4.2) into (5.29) yields that

$$\psi_\omega(\mathbf{x}) = \frac{\omega}{2}\|\text{Prox}_{\sigma_{C_1}}(\mathbf{x})\|_2^2 - \frac{\omega-1}{2}\|\text{Prox}_{\sigma_{C_2}}(\mathbf{x})\|_2^2. \tag{K.4.3}$$

On the other hand, it holds from the definition in (5.29) that

$$\begin{aligned}
\psi_\omega(\mathbf{x}) &= \frac{1}{2}\|\mathbf{x}\|_2^2 - \omega \sigma_{C_1}(\mathbf{x}) + (\omega - 1) \sigma_{C_2}(\mathbf{x}) \\
&= \frac{1}{2}\|\mathbf{x}\|_2^2 - \omega \left(\sigma_{C_1}(\text{Prox}_{\sigma_{C_1}}(\mathbf{x})) + \frac{1}{2}\|\mathbf{x} - \text{Prox}_{\sigma_{C_1}}(\mathbf{x})\|_2^2 \right) \\
&\quad + (\omega - 1) \left(\sigma_{C_2}(\text{Prox}_{\sigma_{C_2}}(\mathbf{x})) + \frac{1}{2}\|\mathbf{x} - \text{Prox}_{\sigma_{C_2}}(\mathbf{x})\|_2^2 \right) \\
&= \frac{1}{2}\|\mathbf{x}\|_2^2 - \omega \sigma_{C_1}(\text{Prox}_{\sigma_{C_1}}(\mathbf{x})) - \frac{\omega}{2}\|\mathbf{x}\|_2^2 - \frac{\omega}{2}\|\text{Prox}_{\sigma_{C_2}}(\mathbf{x})\|_2^2 \\
&\quad + \omega \langle \mathbf{x}, \text{Prox}_{\sigma_{C_1}}(\mathbf{x}) \rangle_2 + (\omega - 1) \sigma_{C_2}(\text{Prox}_{\sigma_{C_2}}(\mathbf{x})) \\
&\quad + (\omega - 1) \frac{1}{2}\|\mathbf{x}\|_2^2 + \frac{\omega - 1}{2}\|\text{Prox}_{\sigma_{C_2}}(\mathbf{x})\|_2^2 \\
&\quad - (\omega - 1) \langle \mathbf{x}, \text{Prox}_{\sigma_{C_2}}(\mathbf{x}) \rangle_2 \\
&= (\omega - 1) \sigma_{C_2}(\text{Prox}_{\sigma_{C_2}}(\mathbf{x})) - \omega \sigma_{C_1}(\text{Prox}_{\sigma_{C_1}}(\mathbf{x})) \\
&\quad + \underbrace{\langle \mathbf{x}, \omega \text{Prox}_{\sigma_{C_1}}(\mathbf{x}) - (\omega - 1) \text{Prox}_{\sigma_{C_2}}(\mathbf{x}) \rangle_2}_{=\Delta_\omega(\mathbf{x})} \\
&\quad + \underbrace{\frac{\omega - 1}{2}\|\text{Prox}_{\sigma_{C_2}}(\mathbf{x})\|_2^2 - \frac{\omega}{2}\|\text{Prox}_{\sigma_{C_1}}(\mathbf{x})\|_2^2}_{=-\psi_\omega(\mathbf{x})}, \tag{K.4.4}
\end{aligned}$$

where the first equality is due to Fact 2.4(c), and the second equality is due to the definition of the Moreau envelope (2.21). Hence, substituting (K.4.3) to (K.4.4) yields that

$$\begin{aligned}
\psi_\omega(\mathbf{x}) &= (\omega - 1) \sigma_{C_2}(\text{Prox}_{\sigma_{C_2}}(\mathbf{x})) - \omega \sigma_{C_1}(\text{Prox}_{\sigma_{C_1}}(\mathbf{x})) \\
&\quad + \langle \mathbf{x}, \Delta_\omega(\mathbf{x}) \rangle_2 - \psi_\omega(\mathbf{x}). \tag{K.4.5}
\end{aligned}$$

Since ψ_ω is convex by Proposition 5.8, we have $\Delta_\omega(\mathbf{x}) = \nabla \psi_\omega(\mathbf{x})$, hence the Fenchel-Young inequality holds with equality [96, Proposition 16.10], *i.e.*,

$$\psi_\omega(\mathbf{x}) + \psi_\omega^*(\Delta_\omega(\mathbf{x})) = \langle \mathbf{x}, \Delta_\omega(\mathbf{x}) \rangle_2. \tag{K.4.6}$$

Hence, substituting (K.4.6) to (K.4.5) yields that

$$\begin{aligned}
\psi_\omega^*(\Delta_\omega(\mathbf{x})) &= \psi_\omega(\mathbf{x}) - (\omega - 1) \sigma_{C_2}(\text{Prox}_{\sigma_{C_2}}(\mathbf{x})) + \omega \sigma_{C_1}(\text{Prox}_{\sigma_{C_1}}(\mathbf{x})) \\
&= \omega \left(\sigma_{C_1}(\text{Prox}_{\sigma_{C_1}}(\mathbf{x})) + \frac{1}{2}\|\text{Prox}_{\sigma_{C_1}}(\mathbf{x})\|_2^2 \right) \\
&\quad - (\omega - 1) \left(\sigma_{C_2}(\text{Prox}_{\sigma_{C_2}}(\mathbf{x})) + \frac{1}{2}\|\text{Prox}_{\sigma_{C_2}}(\mathbf{x})\|_2^2 \right), \tag{K.4.7}
\end{aligned}$$

where the last equality is due to (K.4.3).

In the following, we shall express $\text{Prox}_{\sigma_{C_1}}$ and $\text{Prox}_{\sigma_{C_2}}$ using Δ_ω (see (K.4.9) and (K.4.12) below), and we then substitute those expressions to (K.4.7), to obtain ψ_ω^* (see (K.4.14) below), which will lead to (5.52). It holds from the definition of Δ_ω in (5.1) that

$$\begin{aligned}
\Delta_\omega &= \omega \text{Prox}_{\sigma_{C_1}} - (\omega - 1) \text{Prox}_{\sigma_{C_2}} \\
&= \omega \text{Prox}_{\sigma_{C_1}} - (\omega - 1) \text{Prox}_{\sigma_{C_2} - \sigma_{C_1}} \circ \text{Prox}_{\sigma_{C_1}} \\
&= \text{Prox}_{\sigma_{C_1}} + (\omega - 1) \text{Prox}_{\sigma_{C_1}} - (\omega - 1) \text{Prox}_{\sigma_{C_2} - \sigma_{C_1}} \circ \text{Prox}_{\sigma_{C_1}} \\
&= [\text{Id} + (\omega - 1)(\text{Id} - \text{Prox}_{\sigma_{C_2} - \sigma_{C_1}})] \circ \text{Prox}_{\sigma_{C_1}} \\
&= [\text{Id} + (\omega - 1)\nabla(\text{}^1(\sigma_{C_2} - \sigma_{C_1}))] \circ \text{Prox}_{\sigma_{C_1}}, \tag{K.4.8}
\end{aligned}$$

where the second equality is due to assumption (C-2), and the last equality is due to Fact 2.4(b) under assumption (C-1). Then, it follows from (K.4.8) and (2.35) that

$$\begin{aligned}
\text{Prox}_{\sigma_{C_1}} &= [\text{Id} + (\omega - 1)\nabla(\text{}^1(\sigma_{C_2} - \sigma_{C_1}))]^{-1} \circ \Delta_\omega \\
&= \text{Prox}_{(\omega-1)\text{}^1(\sigma_{C_2} - \sigma_{C_1})} \circ \Delta_\omega. \tag{K.4.9}
\end{aligned}$$

On the other hand, it holds from [96, Proposition 12.22] that $\omega(\sigma_{C_2} - \sigma_{C_1}) = (\omega-1)(\text{}^1(\sigma_{C_2} - \sigma_{C_1}))$. Taking gradients of both sides and applying Fact 2.4(b) yields that

$$\begin{aligned}
\omega^{-1}(\text{Id} - \text{Prox}_{\omega(\sigma_{C_2} - \sigma_{C_1})}) &= (\omega - 1)^{-1}(\text{Id} - \text{Prox}_{(\omega-1)\text{}^1(\sigma_{C_2} - \sigma_{C_1})}) \\
\Leftrightarrow \text{Prox}_{(\omega-1)\text{}^1(\sigma_{C_2} - \sigma_{C_1})} &= \text{Id} - (\omega - 1)\omega^{-1}(\text{Id} - \text{Prox}_{\omega(\sigma_{C_2} - \sigma_{C_1})}). \tag{K.4.10}
\end{aligned}$$

Substituting (K.4.10) to (K.4.9) yields that

$$\text{Prox}_{\sigma_{C_1}} = [\text{Id} - (\omega - 1)\omega^{-1}(\text{Id} - \text{Prox}_{\omega(\sigma_{C_2} - \sigma_{C_1})})] \circ \Delta_\omega. \tag{K.4.11}$$

By substituting (K.4.11) to the definition of Δ_ω in (5.1), we obtain

$$\begin{aligned}
\text{Prox}_{\sigma_{C_2}} &= \omega(\omega - 1)^{-1} \text{Prox}_{\sigma_{C_1}} - (\omega - 1)^{-1} \Delta_\omega \\
&= [\omega(\omega - 1)^{-1} \text{Id} - (\text{Id} - \text{Prox}_{\omega(\sigma_{C_2} - \sigma_{C_1})})] \circ \Delta_\omega - (\omega - 1)^{-1} \Delta_\omega \\
&= \text{Prox}_{\omega(\sigma_{C_2} - \sigma_{C_1})} \circ \Delta_\omega. \tag{K.4.12}
\end{aligned}$$

Hence, substituting (K.4.11) and (K.4.12) to (K.4.7) yields that

$$\begin{aligned}
\psi_\omega^*(\Delta_\omega(\mathbf{x})) &= \omega \left(\sigma_{C_1} + \frac{1}{2} \|\cdot\|_2^2 \right) (\text{Id} \\
&\quad - (\omega - 1)\omega^{-1}(\text{Id} - \text{Prox}_{\omega(\sigma_{C_2} - \sigma_{C_1})})) (\Delta_\omega(\mathbf{x})) \\
&\quad - (\omega - 1) \left(\sigma_{C_2} + \frac{1}{2} \|\cdot\|_2^2 \right) (\text{Prox}_{\omega(\sigma_{C_2} - \sigma_{C_1})}(\Delta_\omega(\mathbf{x}))). \tag{K.4.13}
\end{aligned}$$

By setting $\mathbf{z} := \Delta_\omega(\mathbf{x})$, we obtain

$$\begin{aligned} \psi_\omega^*(\mathbf{z}) &= \omega \left(\sigma_{C_1} + \frac{1}{2} \|\cdot\|_2^2 \right) \left(\mathbf{z} - (1 - \omega^{-1})(\text{Id} - \text{Prox}_{\omega(\sigma_{C_2} - \sigma_{C_1})})(\mathbf{z}) \right) \\ &\quad - (\omega - 1) \left(\sigma_{C_2} + \frac{1}{2} \|\cdot\|_2^2 \right) (\text{Prox}_{\omega(\sigma_{C_2} - \sigma_{C_1})}(\mathbf{z})). \end{aligned} \quad (\text{K.4.14})$$

Since $\text{range } \Delta_\omega = \mathbb{R}^n$ by assumption, (K.4.14) holds for any $\mathbf{z} \in \mathbb{R}^n$. Here, by simple manipulations, it can be verified that

$$\begin{aligned} &\frac{\omega}{2} \|\mathbf{z} - (1 - \omega^{-1})(\text{Id} - \text{Prox}_{\omega(\sigma_{C_2} - \sigma_{C_1})})(\mathbf{z})\|_2^2 \\ &= \frac{\omega}{2} \|\omega^{-1}\mathbf{z} + (1 - \omega^{-1})\text{Prox}_{\omega(\sigma_{C_2} - \sigma_{C_1})}(\mathbf{z})\|_2^2 \\ &= \frac{\omega^{-1}}{2} \|\mathbf{z}\|_2^2 + (1 - \omega^{-1}) \langle \mathbf{z}, \text{Prox}_{\omega(\sigma_{C_2} - \sigma_{C_1})}(\mathbf{z}) \rangle_2 \\ &\quad + \frac{(\omega - 1)(1 - \omega^{-1})}{2} \|\text{Prox}_{\omega(\sigma_{C_2} - \sigma_{C_1})}(\mathbf{z})\|_2^2, \end{aligned} \quad (\text{K.4.15})$$

from which it follows that

$$\begin{aligned} &\frac{\omega}{2} \|\mathbf{z} - (1 - \omega^{-1})(\text{Id} - \text{Prox}_{\omega(\sigma_{C_2} - \sigma_{C_1})})(\mathbf{z})\|_2^2 - \frac{\omega - 1}{2} \|\text{Prox}_{\omega(\sigma_{C_2} - \sigma_{C_1})}(\mathbf{z})\|_2^2 \\ &\quad - \frac{1}{2} \|\mathbf{z}\|_2^2 \\ &= -\frac{1 - \omega^{-1}}{2} \|\mathbf{z}\|_2^2 + (1 - \omega^{-1}) \langle \mathbf{z}, \text{Prox}_{\omega(\sigma_{C_2} - \sigma_{C_1})}(\mathbf{z}) \rangle_2 \\ &\quad - \frac{1 - \omega^{-1}}{2} \|\text{Prox}_{\omega(\sigma_{C_2} - \sigma_{C_1})}(\mathbf{z})\|_2^2 \\ &= -\frac{1 - \omega^{-1}}{2} \|\mathbf{z} - \text{Prox}_{\omega(\sigma_{C_2} - \sigma_{C_1})}(\mathbf{z})\|_2^2. \end{aligned} \quad (\text{K.4.16})$$

Since $\varphi_\omega := \psi_\omega^* - \|\cdot\|_2^2/2$ (see (5.31)), substituting (K.4.16) to (K.4.14) and subtracting $\|\mathbf{z}\|_2^2$ from both sides of the equation yields that

$$\begin{aligned} \varphi_\omega(\mathbf{z}) &= \sigma_{C_1} \left(\mathbf{z} + (\omega - 1)\text{Prox}_{\omega(\sigma_{C_2} - \sigma_{C_1})}(\mathbf{z}) \right) \\ &\quad - (\omega - 1)\sigma_{C_2}(\text{Prox}_{\omega(\sigma_{C_2} - \sigma_{C_1})}(\mathbf{z})) \\ &\quad - \frac{1 - \omega^{-1}}{2} \|\mathbf{z} - \text{Prox}_{\omega(\sigma_{C_2} - \sigma_{C_1})}(\mathbf{z})\|_2^2, \end{aligned} \quad (\text{K.4.17})$$

where σ_{C_1} is positively homogeneous by definition of the support function in

Section 2.1. Regarding the second and third terms of (K.4.17), it holds that

$$\begin{aligned}
& (\omega - 1)\sigma_{C_2}(\text{Prox}_{\omega(\sigma_{C_2}-\sigma_{C_1})}(\mathbf{z})) + \frac{1 - \omega^{-1}}{2} \|\mathbf{z} - \text{Prox}_{\omega(\sigma_{C_2}-\sigma_{C_1})}(\mathbf{z})\|_2^2 \\
&= (\omega - 1)(\sigma_{C_2} - \sigma_{C_1})(\text{Prox}_{\omega(\sigma_{C_2}-\sigma_{C_1})}(\mathbf{z})) \\
&\quad + \frac{1 - \omega^{-1}}{2} \|\mathbf{z} - \text{Prox}_{\omega(\sigma_{C_2}-\sigma_{C_1})}(\mathbf{z})\|_2^2 \\
&\quad + (\omega - 1)\sigma_{C_1}(\text{Prox}_{\omega(\sigma_{C_2}-\sigma_{C_1})}(\mathbf{z})) \\
&= (1 - \omega^{-1})^1(\omega(\sigma_{C_2} - \sigma_{C_1}))(\mathbf{z}) + (\omega - 1)\sigma_{C_1}(\text{Prox}_{\omega(\sigma_{C_2}-\sigma_{C_1})}(\mathbf{z})).
\end{aligned} \tag{K.4.18}$$

Finally, substituting (K.4.18) to (K.4.17) yields (5.52).

K.5 Proof of Proposition 5.10

Since $\Omega_{\lambda_1, \lambda_2}^{\text{OSCAR}}(\mathbf{z}) = \Omega_{\lambda_1, \lambda_2}^{\text{OSCAR}}(\mathbf{P}(|\mathbf{x}|) \text{Sign}(\mathbf{x}) \odot \mathbf{z})$ for any $\mathbf{z} \in \mathbb{R}^n$, it holds for any $\mathbf{x} \in \mathbb{R}^n$ that

$$\begin{aligned}
& \Omega_{\lambda_1, \lambda_2}^{\text{OSCAR}}(\mathbf{x} + (\omega - 1) \text{Prox}_{\omega(\eta-1)\Omega_{\lambda_1, \lambda_2}^{\text{OSCAR}}}(\mathbf{x})) \\
&= \Omega_{\lambda_1, \lambda_2}^{\text{OSCAR}}(|\mathbf{x}|_{\downarrow} + (\omega - 1)P_{\mathcal{K}_{\geq 0}^n}(|\mathbf{x}|_{\downarrow} - \omega(\eta - 1)\mathbf{w})) \\
&= \langle \mathbf{w}, |\mathbf{x}|_{\downarrow} + (\omega - 1)P_{\mathcal{K}_{\geq 0}^n}(|\mathbf{x}|_{\downarrow} - \omega(\eta - 1)\mathbf{w}) \rangle_2 \\
&= \langle \mathbf{w}, |\mathbf{x}|_{\downarrow} + (\omega - 1)P_{\mathcal{K}_{\geq 0}^n}(|\mathbf{x}|_{\downarrow} - \omega(\eta - 1)\mathbf{w}) \rangle_2 \\
&= \langle \mathbf{w}, |\mathbf{x}|_{\downarrow} \rangle_2 + (\omega - 1)\langle \mathbf{w}, P_{\mathcal{K}_{\geq 0}^n}(|\mathbf{x}|_{\downarrow} - \omega(\eta - 1)\mathbf{w}) \rangle_2 \\
&= \Omega_{\lambda_1, \lambda_2}^{\text{OSCAR}}(\mathbf{x}) + (\omega - 1)\Omega_{\lambda_1, \lambda_2}^{\text{OSCAR}}(\text{Prox}_{\omega(\eta-1)\Omega_{\lambda_1, \lambda_2}^{\text{OSCAR}}}(\mathbf{x})),
\end{aligned} \tag{K.5.1}$$

where the second and last equalities are due to (2.70). Hence, (K.5.1) and (5.52) yield (5.59). Moreover, it holds that

$$\begin{aligned}
(5.59) &= \Omega_{\lambda_1, \lambda_2}^{\text{OSCAR}} - abb^{-1}1(b\Omega_{\lambda_1, \lambda_2}^{\text{OSCAR}}) \\
&= \Omega_{\lambda_1, \lambda_2}^{\text{OSCAR}} - ab^b(\Omega_{\lambda_1, \lambda_2}^{\text{OSCAR}}),
\end{aligned} \tag{K.5.2}$$

where the last equality is due to Lemma 5.1. Hence, (5.60) holds if $ab = 1$.

Bibliography

- [1] D. L. Donoho, “Compressed sensing,” *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.
- [2] E. J. Candes and T. Tao, “Near-optimal signal recovery from random projections: Universal encoding strategies?” *IEEE Trans. Inf. Theory*, vol. 52, no. 12, pp. 5406–5425, 2006.
- [3] Y. C. Eldar and G. Kutyniok, *Compressed Sensing: Theory and Applications*. U.K.: Cambridge Univ. Press, 2012.
- [4] S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing*. Birkhäuser New York, NY, 2013.
- [5] S. Mohamed, K. Heller, and Z. Ghahramani, “Bayesian and L1 approaches to sparse unsupervised learning,” in *Proc. Int. Conf. Mach. Learn.*, 2012, pp. 751–758.
- [6] J. Gui, Z. Sun, S. Ji, D. Tao, and T. Tan, “Feature selection based on structured sparsity: A comprehensive study,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 7, pp. 1490–1507, 2016.
- [7] H. Zhang, J. Wang, Z. Sun, J. M. Zurada, and N. R. Pal, “Feature selection for neural networks using group lasso regularization,” *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 4, pp. 659–673, 2019.
- [8] J. Fan and R. Li, “Statistical challenges with high dimensionality,” in *Proc. Int. Congr. Mathematicians*, 2006, pp. 595–622.
- [9] F. Nie, H. Huang, X. Cai, and C. H. Ding, “Efficient and robust feature selection via joint $\ell_{2,1}$ -norms minimization,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 1813–1821.
- [10] S. Xiang, F. Nie, G. Meng, C. Pan, and C. Zhang, “Discriminative least squares regression for multiclass classification and feature selection,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 11, pp. 1738–1754, Nov. 2012.

- [11] R. Jenatton, J.-Y. Audibert, and F. Bach, "Structured variable selection with sparsity-inducing norms," *J. Mach. Learn. Res.*, vol. 12, pp. 2777–2824, 2011.
- [12] M. Lustig, D. Donoho, and J. M. Pauly, "Sparse MRI: The application of compressed sensing for rapid MR imaging," *Magn. Reson. Med.*, vol. 58, no. 6, pp. 1182–1195, 2007.
- [13] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2861–2873, 2010.
- [14] P. S. Rao, P. K. Jana, and H. Banka, "A particle swarm optimization based energy efficient cluster head selection algorithm for wireless sensor networks," *Wireless Networks*, vol. 23, no. 7, pp. 2005–2020, 2017.
- [15] B. K. Natarajan, "Sparse approximate solutions to linear systems," *SIAM J. Comput.*, vol. 24, no. 2, pp. 227–234, 1995.
- [16] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3397–3415, 1993.
- [17] Y. C. Pati, R. Rezaifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Proc. Asilomar Conf. Signals Syst. Comput.*, 1993, pp. 40–44.
- [18] T. Blumensath and M. E. Davies, "Iterative hard thresholding for compressed sensing," *Appl. Comput. Harmon. Anal.*, vol. 27, no. 3, pp. 265–274, 2009.
- [19] ———, "Normalized iterative hard thresholding: Guaranteed stability and performance," *IEEE J. Select. Topics Signal Process.*, vol. 4, no. 2, pp. 298–309, 2010.
- [20] D. Needell and J. A. Tropp, "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples," *Appl. Comput. Harmon. Anal.*, vol. 26, no. 3, pp. 301–321, 2009.
- [21] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Stat. Soc. B*, vol. 58, no. 1, pp. 267–288, 1996.
- [22] A. M. Bruckstein, D. L. Donoho, and M. Elad, "From sparse solutions of systems of equations to sparse modeling of signals and images," *SIAM Rev.*, vol. 51, no. 1, pp. 34–81, 2009.

- [23] S. Osher, F. Ruan, J. Xiong, Y. Yao, and W. Yin, “Sparse recovery via differential inclusions,” *Appl. Comput. Harmon. Anal.*, vol. 41, no. 2, pp. 436–469, 2016.
- [24] C. H. Zhang, “Nearly unbiased variable selection under minimax concave penalty,” *The Ann. Statist.*, vol. 38, no. 2, pp. 894–942, Apr. 2010.
- [25] I. Selesnick, “Sparse regularization via convex analysis,” *IEEE Trans. Signal Process.*, vol. 65, no. 17, pp. 4481–4494, Sep. 2017.
- [26] J. Fan and R. Li, “Variable selection via nonconcave penalized likelihood and its oracle properties,” *J. Amer. Stat. Assoc.*, vol. 96, no. 456, pp. 1348–1360, 2001.
- [27] R. Chartrand, “Exact reconstruction of sparse signals via nonconvex minimization,” *IEEE Signal Process. Lett.*, vol. 14, no. 10, pp. 707–710, Oct. 2007.
- [28] K. Jeong, M. Yukawa, and S. Amari, “Can critical-point paths under ℓ_p -regularization ($0 < p < 1$) reach the sparsest least squares solutions?” *IEEE Trans. Inf. Theory*, vol. 60, no. 5, pp. 2960–2968, May 2014.
- [29] M. Yukawa and S. Amari, “ ℓ_p -regularized least squares ($0 < p < 1$) and critical path,” *IEEE Trans. Inf. Theory*, vol. 62, no. 1, pp. 488–502, Jan. 2016.
- [30] E. Soubies, L. Blanc-Féraud, and G. Aubert, “A continuous exact ℓ_0 penalty (CEL0) for least squares regularized problem,” *SIAM J. Imag. Sci.*, vol. 8, no. 3, pp. 1607–1639, 2015.
- [31] D. Peleg and R. Meir, “A bilinear formulation for vector sparsity optimization,” *Signal Process.*, vol. 88, no. 2, pp. 375–389, 2008.
- [32] E. J. Candes, M. B. Wakin, and S. P. Boyd, “Enhancing sparsity by reweighted ℓ_1 minimization,” *J. Fourier Anal. Appl.*, vol. 14, no. 5, pp. 877–905, Oct. 2008.
- [33] Z. Zhang and B. Tu, “Nonconvex penalization using laplace exponents and concave conjugates,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2012.
- [34] J. Abe, M. Yamagishi, and I. Yamada, “Linearly involved generalized Moreau enhanced models and their proximal splitting algorithm under overall convexity condition,” *Inverse Problems*, vol. 36, no. 3, pp. 1–36, Feb. 2020.

- [35] M. Yukawa, H. Kaneko, K. Suzuki, and I. Yamada, "Linearly-involved Moreau-enhanced-over-subspace model: Debiased sparse modeling and stable outlier-robust regression," *IEEE Trans. Signal Process.*, vol. 71, pp. 1232–1247, 2023.
- [36] A. Lanza, S. Morigi, I. W. Selesnick, and F. Sgallari, "Sparsity-inducing nonconvex nonseparable regularization for convex image processing," *SIAM J. Imag. Sci.*, vol. 12, no. 2, pp. 1099–1134, 2019.
- [37] A. Parekh and I. W. Selesnick, "Enhanced low-rank matrix approximation," *IEEE Signal Process. Lett.*, vol. 23, no. 4, pp. 493–497, 2016.
- [38] L. Jacob, G. Obozinski, and J.-P. Vert, "Group lasso with overlap and graph lasso," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 433–440.
- [39] T. Hastie, R. Tibshirani, and M. Wainwright, *Statistical Learning with Sparsity: the Lasso and Generalizations*. CRC Press, 2015.
- [40] I. F. Gorodnitsky and B. D. Rao, "Sparse signal reconstruction from limited data using focuss: A re-weighted minimum norm algorithm," *IEEE Trans. Signal Process.*, vol. 45, no. 3, pp. 600–616, 1997.
- [41] P. Stoica and R. Moses, *Introduction to Spectral Analysis*. Upper Saddle River, NJ: Prentice hall, 1997.
- [42] S. Erickson and C. Sabatti, "Empirical Bayes estimation of a sparse vector of gene expression changes," *Statist. Appl. Genetics Mol. Biol.*, vol. 4, no. 1, pp. 1–35, 2005.
- [43] S. F. Cotter and B. D. Rao, "Sparse channel estimation via matching pursuit with application to equalization," *IEEE Trans. Commun.*, vol. 50, no. 3, pp. 374–377, 2002.
- [44] I. F. Gorodnitsky, J. S. George, and B. D. Rao, "Neuromagnetic source imaging with FOCUSS: a recursive weighted minimum norm algorithm," *Electroencephalogr. Clin. Neurophysiol.*, vol. 95, no. 4, pp. 231–251, 1995.
- [45] D. M. Malioutov, M. Cetin, and A. S. Willsky, "Source localization by enforcing sparsity through a Laplacian prior: an SVD-based approach," in *Proc. Workshop Stat. Signal Process.* IEEE, 2003, pp. 573–576.
- [46] Y. C. Eldar and H. Rauhut, "Average case analysis of multichannel sparse recovery using convex relaxation," *IEEE Trans. Inf. Theory*, vol. 56, no. 1, pp. 505–519, 2009.

- [47] Y. C. Eldar, *Sampling Theory: Beyond Bandlimited Systems*. Cambridge Univ. Press, 2015.
- [48] M. E. Davies and Y. C. Eldar, "Rank awareness in joint sparse recovery," *IEEE Trans. Inf. Theory*, vol. 58, no. 2, pp. 1135–1146, 2012.
- [49] J. Chen and X. Huo, "Theoretical results on sparse representations of multiple-measurement vectors," *IEEE Trans. Signal Process.*, vol. 54, no. 12, pp. 4634–4643, 2006.
- [50] J. A. Tropp, A. C. Gilbert, and M. J. Strauss, "Algorithms for simultaneous sparse approximation. part i: Greedy pursuit," *Signal Processing*, vol. 86, no. 3, pp. 572–588, 2006.
- [51] J. D. Blanchard, M. Cermak, D. Hanle, and Y. Jing, "Greedy algorithms for joint sparse recovery," *IEEE Trans. Signal Process.*, vol. 62, no. 7, pp. 1694–1704, 2014.
- [52] R. O. Schmidt, "A signal subspace approach to multiple emitter location and spectral estimation." Ph.D. dissertation, Stanford Univ., Stanford, CA, Nov. 1981.
- [53] K. Lee, Y. Bresler, and M. Junge, "Subspace methods for joint sparse recovery," *IEEE Trans. Inf. Theory*, vol. 58, no. 6, pp. 3613–3641, 2012.
- [54] J. M. Kim, O. K. Lee, and J. C. Ye, "Compressive music: Revisiting the link between compressive sensing and array signal processing," *IEEE Trans. Inf. Theory*, vol. 58, no. 1, pp. 278–301, 2012.
- [55] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. Roy. Stat. Soc. Ser. B*, vol. 68, no. 1, pp. 49–67, 2006.
- [56] M. Carlsson, J.-Y. Tournier, and H. Wendt, "Unbiased group-sparsity sensing using quadratic envelopes," in *Proc. Int. Workshop Comput. Adv. Multi-Sensor Adaptive Process.* IEEE, 2019, pp. 425–429.
- [57] P. Sprechmann, I. Ramirez, G. Sapiro, and Y. C. Eldar, "Chilasso: A collaborative hierarchical sparse modeling framework," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4183–4198, 2011.
- [58] S. Chatterjee, A. Banerjee, S. Chatterjee, and A. R. Ganguly, "Sparse group lasso for regression on land climate variables," in *Proc. Int. Conf. Data Mining Workshops.* IEEE, 2011, pp. 1–8.
- [59] M. Yukawa and H. Kagami, "Supervised nonnegative matrix factorization via minimization of regularized Moreau-envelope of divergence function with application to music transcription," *J. Franklin Institute*, vol. 355, no. 4, pp. 2041–2066, 2018.

- [60] L. Wang, G. Chen, and H. Li, “Group scad regression analysis for microarray time course gene expression data,” *Bioinformatics*, vol. 23, no. 12, pp. 1486–1494, 2007.
- [61] J. Huang, P. Breheny, and S. Ma, “A selective review of group selection in high-dimensional models,” *Stat. Sci.*, vol. 27, no. 4, 2012.
- [62] P. Breheny and J. Huang, “Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors,” *Stat. and Comput.*, vol. 25, pp. 173–187, 2015.
- [63] Y. Chen, M. Yamagishi, and I. Yamada, “A Generalized Moreau Enhancement of $\ell_{2,1}$ -norm and Its Application to Group Sparse Classification,” in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, 2021, pp. 2134–2138.
- [64] P. J. Huber and E. M. Ronchetti, *Robust Statistics*, 2nd ed. Wiley, 2009.
- [65] R. A. Maronna, R. D. Martin, V. J. Yohai, and M. Salibián-Barrera, *Robust Statistics: Theory and Methods (with R)*, 2nd ed. Wiley, 2019.
- [66] A. M. Zoubir, V. Koivunen, E. Ollila, and M. Muma, *Robust Statistics for Signal Processing*. Cambridge Univ. Press, 2018.
- [67] T. Andrysiak, “Sparse representation and overcomplete dictionary learning for anomaly detection in electrocardiograms,” *Neural Comput. Appl.*, vol. 32, no. 5, pp. 1269–1285, 2020.
- [68] R. He, T. Tan, L. Wang, and W.-S. Zheng, “ $\ell_{2,1}$ regularized correntropy for robust feature selection,” in *Conf. Comput. Vis. Pattern Recognit.* IEEE, 2012, pp. 2504–2511.
- [69] D. Mo and Z. Lai, “Robust jointly sparse regression for image feature selection,” in *IAPR Asian Conf. Pattern Recognit. (ACPR)*. IEEE, 2017, pp. 477–482.
- [70] C. Hou, F. Nie, D. Yi, and Y. Wu, “Feature selection via joint embedding learning and sparse regression,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 1324–1329.
- [71] R. E. Carrillo, A. B. Ramirez, G. R. Arce, K. E. Barner, and B. M. Sadler, “Robust compressive sensing of sparse signals: a review,” *EURASIP J. Advances Signal Process.*, vol. 108, pp. 1–17, 2016.
- [72] C. Yang, X. Shen, H. Ma, B. Chen, Y. Gu, and H. C. So, “Weakly convex regularized robust sparse recovery methods with theoretical guarantees,” *IEEE Trans. Signal Process.*, vol. 67, no. 19, pp. 5046–5061, Oct. 2019.

- [73] A. Javaheri, H. Zayyani, M. A. Figueiredo, and F. Marvasti, "Robust sparse recovery in impulsive noise via continuous mixed norm," *IEEE Signal Process. Lett.*, vol. 25, no. 8, pp. 1146–1150, Apr. 2018.
- [74] Y. Chen, C. Caramanis, and S. Mannor, "Robust sparse regression under adversarial corruption," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 774–782.
- [75] F. Wen, L. Adhikari, L. Pei, R. F. Marcia, P. Liu, and R. C. Qiu, "Nonconvex regularization-based sparse recovery and demixing with application to color image inpainting," *IEEE Access*, vol. 5, pp. 11 513–11 527, 2017.
- [76] H. Zayyani, M. Korki, and F. Marvasti, "A distributed 1-bit compressed sensing algorithm robust to impulsive noise," *IEEE Commun. Lett.*, vol. 20, no. 6, pp. 1132–1135, 2016.
- [77] H. Liu, R. Zhang, Y. Zhou, X. Jing, and T.-K. Truong, "Speech Denoising Using Transform Domains in the Presence of Impulsive and Gaussian Noises," *IEEE Access*, vol. 5, pp. 21 193–21 203, 2017.
- [78] R. Hu, Y. Fu, Z. Chen, J. Xu, and J. Tang, "Robust DOA estimation via sparse signal reconstruction with impulsive noise," *IEEE Commun. Lett.*, vol. 21, no. 6, pp. 1333–1336, 2017.
- [79] H. Wang, G. Li, and G. Jiang, "Robust regression shrinkage and consistent variable selection through the lad-lasso," *J. Bus. Econ. Statist.*, vol. 25, no. 3, pp. 347–355, 2007.
- [80] J. Yang and Y. Zhang, "Alternating direction algorithms for ℓ_1 -problems in compressive sensing," *SIAM J. Sci. Comput.*, vol. 33, no. 1, pp. 250–278, 2011.
- [81] D. S. Pham and S. Venkatesh, "Efficient algorithms for robust recovery of images from compressed data," *IEEE Trans. Image Process.*, vol. 22, no. 12, pp. 4724–4737, 2013.
- [82] N. H. Nguyen and T. D. Tran, "Robust lasso with missing and grossly corrupted observations," *IEEE Trans. Inf. Theory*, vol. 59, no. 4, pp. 2036–2058, 2012.
- [83] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. Roy. Stat. Soc. B*, vol. 67, no. 2, pp. 301–320, 2005.
- [84] M. Dettling and P. Bühlmann, "Finding predictive gene groups from microarray data," *J. Multivariate Anal.*, vol. 90, no. 1, pp. 106–131, 2004.

- [85] U. Oswal, C. Cox, M. Lambon-Ralph, T. Rogers, and R. Nowak, “Representational similarity learning with application to brain networks,” in *Proc. Int. Conf. Mach. Learn.* PMLR, 2016, pp. 1041–1049.
- [86] X. Shen and H.-C. Huang, “Grouping pursuit through a regularization solution surface,” *J. Amer. Stat. Assoc.*, vol. 105, no. 490, pp. 727–739, 2010.
- [87] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight, “Sparsity and smoothness via the fused lasso,” *J. Roy. Stat. Soc. Ser. B*, vol. 67, no. 1, pp. 91–108, 2005.
- [88] Y. She, “Sparse regression with exact clustering,” *Electron. J. Stat.*, vol. 4, pp. 1055–1096, 2010.
- [89] H. D. Bondell and B. J. Reich, “Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR,” *Biometrics*, vol. 64, no. 1, pp. 115–123, 2008.
- [90] L. W. Zhong and J. T. Kwok, “Efficient sparse modeling with automatic feature grouping,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 9, pp. 1436–1447, 2012.
- [91] B. Vinzamuri, K. K. Padthe, and C. K. Reddy, “Feature grouping using weighted ℓ_1 norm for high-dimensional data,” in *Proc. Int. Conf. Data Mining.* IEEE, 2016, pp. 1233–1238.
- [92] X. Zeng and M. A. Figueiredo, “The ordered weighted ℓ_1 norm: Atomic formulation, projections, and algorithms,” *arXiv:1409.4271*, 2014.
- [93] M. Figueiredo and R. Nowak, “Ordered weighted ℓ_1 regularized regression with strongly correlated covariates: Theoretical aspects,” in *Proc. Artif. Intell. Statist.* PMLR, 2016, pp. 930–938.
- [94] M. Bogdan, E. Van Den Berg, C. Sabatti, W. Su, and E. J. Candès, “SLOPE—adaptive variable selection via convex optimization,” *Ann. Appl. Statist.*, vol. 9, no. 3, p. 1103, 2015.
- [95] H. Hu and Y. M. Lu, “SLOPE for sparse linear regression: Asymptotics and optimal regularization,” *IEEE Trans. Inf. Theory*, vol. 68, no. 11, pp. 7627–7664, 2022.
- [96] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, 2nd ed. New York: Springer, 2017.
- [97] R. A. Horn and C. R. Johnson, *Matrix Analysis*, 2nd ed. New York: Cambridge Univ. Press, 2013.

- [98] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge Univ. Press, 2004.
- [99] J.-B. Baillon and G. Haddad, “Quelques propriétés des opérateurs angle-bornés et n -cycliquement monotones,” *Israel J. Math.*, vol. 26, pp. 137–150, 1977.
- [100] M. Yukawa and I. Yamada, “Cocoercive gradient operator of convex function and its associated weakly convex function: Generalized proximity operator for case of unique minimizer,” in *Proc. Signal Processing Symposium*, 2023, Kyoto, Japan.
- [101] I. Yamada, M. Yukawa, and M. Yamagishi, “Minimizing the Moreau envelope of nonsmooth convex functions over the fixed point set of certain quasi-nonexpansive mappings,” *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, pp. 345–390, 2011.
- [102] T. M. Apostol, *Mathematical Analysis*, 2nd ed. Addison-Wesley, 1974.
- [103] D. L. Donoho, “De-noising by soft-thresholding,” *IEEE Trans. Inf. Theory*, vol. 41, no. 3, pp. 613–627, 1995.
- [104] F. Xu, J. Duan, and W. Liu, “Comparative study of non-convex penalties and related algorithms in compressed sensing,” *Digital Signal Process.*, vol. 135, p. 103937, 2023.
- [105] A. Böhm and S. J. Wright, “Variable smoothing for weakly convex composite functions,” *J. Optim. Theory Appl.*, vol. 188, pp. 628–649, 2021.
- [106] L. Chen and Y. Gu, “The convergence guarantees of a non-convex approach for sparse recovery,” *IEEE Trans. Signal Process.*, vol. 62, no. 15, pp. 3754–3767, 2014.
- [107] D. Andrews and F. Hampel, *Robust Estimates of Location: Survey and Advances*, ser. Princeton Legacy Library. Princeton Univ. Press, 2015.
- [108] R. E. Barlow, *Statistical Inference Under Order Restrictions: The Theory and Application of Isotonic Regression*. New York: Wiley, 1972.
- [109] J. d. Leeuw, K. Hornik, and P. Mair, “Isotone optimization in r: pool-adjacent-violators algorithm (pava) and active set methods,” *J. Stat. Software*, 2009.
- [110] M. J. Best and N. Chakravarti, “Active set algorithms for isotonic regression; a unifying framework,” *Mathematical Programming*, vol. 47, no. 1, pp. 425–439, 1990.

- [111] H.-Y. Gao and A. G. Bruce, “Waveshrink with firm shrinkage,” *Statistica Sinica*, pp. 855–874, 1997.
- [112] I. Bayram, “On the convergence of the iterative shrinkage/thresholding algorithm with a weakly convex penalty,” *IEEE Trans. Signal Process.*, vol. 64, pp. 1597–1608, 2016.
- [113] D. Gabay, “Chapter IX applications of the method of multipliers to variational inequalities,” in *Stud. Math. Appl.*, vol. 15. Elsevier, 1983, pp. 299–331.
- [114] A. Chambolle and T. Pock, “A first-order primal-dual algorithm for convex problems with applications to imaging,” *J. Math. Imag. Vis.*, vol. 40, no. 1, pp. 120–145, 2011.
- [115] P. L. Combettes and J.-C. Pesquet, “Primal-dual splitting algorithm for solving inclusions with mixtures of composite, Lipschitzian, and parallel-sum type monotone operators,” *Set-Valued Var. Anal.*, vol. 20, no. 2, pp. 307–330, 2012.
- [116] L. Condat, “A primal–dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms,” *J. Optim. Theory Appl.*, vol. 158, no. 2, pp. 460–479, 2013.
- [117] I. Loris and C. Verhoeven, “On a generalization of the iterative soft-thresholding algorithm for the case of non-separable penalty,” *Inverse Problems*, vol. 27, no. 12, pp. 1–15, 2011.
- [118] P. Chen, J. Huang, and X. Zhang, “A primal–dual fixed point algorithm for convex separable minimization with applications to image restoration,” *Inverse Problems*, vol. 29, no. 2, pp. 1–33, 2013.
- [119] J. Eckstein and D. P. Bertsekas, “On the douglas-rachford splitting method and the proximal point algorithm for maximal monotone operators,” *Mathematical Programming*, vol. 55, no. 1-3, pp. 293–318, 1992.
- [120] P. Giselsson, M. Fält, and S. Boyd, “Line search for averaged operator iteration,” in *Proc. Conf. Decis. Control.* IEEE, 2016, pp. 1015–1022.
- [121] D. M. Bradley and J. A. Bagnell, “Convex coding,” in *Proc. Conf. Uncertainty Artif. Intell. (UAI)*. AUAI Press, 2009, pp. 83–90.
- [122] N. Pustelnik and L. Condat, “Proximity operator of a sum of functions; application to depth map estimation,” *IEEE Signal Process. Lett.*, vol. 24, no. 12, pp. 1827–1831, 2017.
- [123] G. Pierra, “Eclatement de contraintes en parallèle pour la minimisation d’une forme quadratique,” in *IFIP Technical Conference on Optimization Techniques*. Springer, 1975, pp. 200–218.

- [124] ———, “Decomposition through formalization in a product space,” *Mathematical Programming*, vol. 28, no. 1, pp. 96–115, 1984.
- [125] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, “Physiobank, physiotookit, and physionet: components of a new research resource for complex physiologic signals,” *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.
- [126] R. Bousseljot, D. Kreiseler, and A. Schnabel, “Nutzung der EKG-Signaldatenbank CARDIODAT der PTB über das Internet,” *Biomedizinische Technik/Biomed. Eng.*, vol. 40, no. s1, pp. 317–318, 1995.
- [127] A. Singh and S. Dandapat, “Weighted mixed-norm minimization based joint compressed sensing recovery of multi-channel electrocardiogram signals,” *Comput. Elect. Eng.*, vol. 53, pp. 203–218, 2016.
- [128] R. R. Tamboli, M. A. Savkoor, S. Jana, and R. Manthalkar, “On the sparsest representation of electrocardiograms,” in *Comput. Cardiology*, 2013, pp. 479–482.
- [129] A. S. Al-Fahoum, “Quality assessment of ECG compression techniques using a wavelet-based diagnostic measure,” *IEEE Trans. Inf. Technol. Biomed.*, vol. 10, no. 1, pp. 182–191, 2006.
- [130] K. Suzuki and M. Yukawa, “Sparse stable outlier-robust regression with minimax concave function,” in *Proc. Int. Workshop Mach. Learn. Signal Process. (MLSP)*. IEEE, 2022, pp. 1–6.
- [131] N. Komodakis and J.-C. Pesquet, “Playing with duality: An overview of recent primal-dual approaches for solving large-scale optimization problems,” *IEEE Signal Process. Mag.*, vol. 32, no. 6, pp. 31–54, 2015.
- [132] Y. Xiao, H. Zhu, and S.-Y. Wu, “Primal and dual alternating direction algorithms for ℓ_1 - ℓ_1 -norm minimization problems in compressive sensing,” *Computational Optimization and Applications*, vol. 54, no. 2, pp. 441–459, 2013.
- [133] L. Wang, J. Zhu, and H. Zou, “Hybrid Huberized support vector machines for microarray classification,” in *Proc. Int. Conf. Mach. Learn.*, 2007, pp. 983–990.
- [134] C. K. Reddy, H. Dubey, K. Koishida, A. Nair, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner, and S. Srinivasan, “Interspeech 2021 deep noise suppression challenge,” in *INTERSPEECH*, 2021.

- [135] A. Beck and M. Teboulle, “A fast iterative shrinkage-thresholding algorithm for linear inverse problems,” *SIAM J. Imag. Sci.*, vol. 2, no. 1, pp. 183–202, 2009.
- [136] H. Kaneko and M. Yukawa, “Normalized least-mean-square algorithms with minimax concave penalty,” in *Proc. Int. Conf. Acoust., Speech, Signal Process.* IEEE, 2020, pp. 5445–5449.
- [137] P. K. Pokala, A. G. Mahurkar, and C. S. Seelamantula, “Firmnet: A sparsity amplified deep network for solving linear inverse problems,” in *Proc. Int. Conf. Acoust., Speech, Signal Process.* IEEE, 2019, pp. 2982–2986.
- [138] E. Z. Tan, C. Chaux, E. Soubies, and V. Y. Tan, “Deep unrolling for nonconvex robust principal component analysis,” in *Proc. IEEE Int. Workshop Mach. Learn. Signal Process.*, 2023, pp. 1–6.
- [139] M. Yukawa and I. Yamada, “Monotone Lipschitz-gradient denoiser: Explainability of operator regularization approaches and convergence to optimal point,” *arXiv:2406.04676*, 2024.
- [140] R. Gribonval and M. Nikolova, “A characterization of proximity operators,” *J. Math. Imag. Vis.*, vol. 62, no. 6-7, pp. 773–789, 2020.
- [141] H. H. Bauschke, E. Matoušková, and S. Reich, “Projection and proximal point methods: convergence results and counterexamples,” *Nonlinear Analysis: Theory, Methods and Applications*, vol. 56, no. 5, pp. 715–738, 2004.
- [142] H. H. Bauschke, R. Goebel, Y. Lucet, and X. Wang, “The proximal average: Basic theory,” *SIAM J. Optim.*, vol. 19, no. 2, pp. 766–785, 2008.
- [143] P. L. Combettes, “Resolvent and proximal compositions,” *Set-Valued and Variational Analysis*, vol. 31, no. 3, p. 22, 2023.
- [144] ———, “Monotone operator theory in convex optimization,” *Mathematical Programming*, vol. 170, pp. 177–206, 2018.
- [145] Y.-L. Yu, “On decomposing the proximal map,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 26, 2013.
- [146] P. L. Combettes and V. R. Wajs, “Signal recovery by proximal forward-backward splitting,” *Multiscale Model. Simul.*, vol. 4, no. 4, pp. 1168–1200, 2005.
- [147] V. Monga, Y. Li, and Y. C. Eldar, “Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing,” *IEEE Signal Process. Mag.*, vol. 38, no. 2, pp. 18–44, 2021.

- [148] K. Mitra, A. Veeraraghavan, and R. Chellappa, “Analysis of sparse regularization based robust regression approaches,” *IEEE Trans. Signal Process.*, vol. 61, no. 5, pp. 1249–1257, Mar. 2013.
- [149] K. Gregor and Y. LeCun, “Learning fast approximations of sparse coding,” in *Proc. Int. Conf. Mach. Learn.*, 2010, pp. 399–406.
- [150] I. Daubechies, M. Defrise, and C. De Mol, “An iterative thresholding algorithm for linear inverse problems with a sparsity constraint,” *Comm. Pure Appl. Math.*, vol. 57, no. 11, pp. 1413–1457, 2004.
- [151] L. Condat, D. Kitahara, A. Contreras, and A. Hirabayashi, “Proximal splitting algorithms for convex optimization: A tour of recent advances, with new twists,” *SIAM Review*, vol. 65, no. 2, pp. 375–435, 2023.
- [152] J. Von Zur Gathen and J. Gerhard, *Modern Computer Algebra*. Cambridge Univ. Press, 2003.
- [153] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 4th ed. JHU press, 2013.
- [154] Z. Zhou, X. Li, J. Wright, E. Candes, and Y. Ma, “Stable principal component pursuit,” in *Proc. Int. Symp. Inf. Theory*. IEEE, 2010, pp. 1518–1522.
- [155] M. Kowalski, “Sparse regression using mixed norms,” *Appl. Comput. Harmon. Anal.*, vol. 27, no. 3, pp. 303–324, 2009.
- [156] D. P. Bertsekas, A. Nedić, and A. E. Ozdaglar, *Convex Analysis and Optimization*. Cambridge, MA: Athena Scientific, 2003.
- [157] F. H. Clarke, “Generalized gradients and applications,” *Trans. Amer. Math. Soc.*, vol. 205, pp. 247–262, 1975.

Publications Related to the Dissertation

Articles in Journals

1. K. Suzuki and M. Yukawa, “Robust recovery of jointly-sparse signals using minimax concave loss function,” *IEEE Transactions on Signal Processing*, vol. 69, pp. 669–661, 2020.
2. K. Suzuki and M. Yukawa, “Sparse stable outlier-robust signal recovery under Gaussian noise,” *IEEE Transactions on Signal Processing*, vol. 71, pp. 372–387, 2023.

International Conference Papers

3. K. Suzuki and M. Yukawa, “External division of two proximity operators: An application to signal recovery with structured sparsity,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Seoul, Korea, pp. 9471–9475, April 2024.
4. K. Suzuki and M. Yukawa, “Sparse stable outlier-robust regression with minimax concave function,” in *Proc. IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, online, 6 pages, August 2022.
5. K. Suzuki and M. Yukawa, “On grouping effect of sparse stable outlier-robust regression,” in *Proc. IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, online, 6 pages, August 2022.
6. K. Suzuki and M. Yukawa, “Robust jointly-sparse signal recovery based on minimax concave loss function,” in *Proc. European Signal Processing Conference (EUSIPCO)*, online, pp. 2070–2074, January 2021.

Other Publications

Articles in Journals

1. M. Yukawa, H. Kaneko, K. Suzuki, and I. Yamada, “Linearly-involved Moreau-enhanced-over-subspace model: Debiased sparse modeling and stable outlier-robust regression,” *IEEE Trans. Signal Processing*, vol. 71, pp. 1232–1247, 2023.

International Conference Papers

2. M. Yukawa, K. Suzuki, I. Yamada, “Stable robust regression under sparse outlier and Gaussian noise”, in *Proc. European Signal Processing Conference (EUSIPCO)*, Belgrade, Serbia, pp. 2236–2240, August–September 2022.

Domestic Conference Papers

3. K. Suzuki and M. Yukawa, “Debiased estimation of signals with structured sparsity based on external division of two proximity operators”, in *Proc. IEICE Signal Processing Symposium*, Kyoto, Japan, 6 pages, Nov. 2023.
4. K. Suzuki and M. Yukawa, “Multiscale manifold clustering and embedding with multiple kernels”, in *Proc. Technical Report of IEICE*, vol. 122, no. 388, SIP2022-167, Okinawa, Japan, pp. 276–281, Mar. 2023.
5. K. Suzuki and M. Yukawa, “Sparse stable outlier-robust regression using minimax concave function”, in *Proc. IEICE Signal Processing Symposium*, online, pp. 96–101, Nov. 2021.
6. K. Suzuki and M. Yukawa, “A robust approach to jointly-sparse signal recovery based on minimax concave loss function”, in *Proc. Technical Report of IEICE*, vol. 119, no. 440, SIP2019-124, Okinawa, Japan (conference cancelled), pp. 123–128, Mar. 2020.