A Discrete Measure for Debiased Feature Grouping: A Limit of Moreau-Enhanced OSCAR Regularizer and Its Proximity Operator

Kyohei Suzuki
Department of Electronics and Electrical Engineering
Keio University
Yokohama, Japan
suzuki.k.439f@m.isct.ac.jp

Masahiro Yukawa
Department of Electronics and Electrical Engineering
Keio University
Yokohama, Japan
yukawa@elec.keio.ac.jp

***Abstract*—Octagonal shrinkage and clustering algorithm for regression (OSCAR) is an effective method for feature grouping, which aims to select important highly-correlated groups of features relevant to the observations. Unfortunately, it is known that OSCAR may cause estimation bias, which is undesirable for many applications. Whereas the Moreau enhancement of convex regularizers promoting sparsity or low-rankness has been studied extensively to reduce the estimation bias, its use in the feature grouping task still remains unexplored. In this paper, we investigate the debiasing effect of the discrete measure defined by a limit of the Moreau-enhanced OSCAR regularizer, which is referred to as the LME-OSCAR regularizer. The proximity operator of the LME-OSCAR regularizer can be computed efficiently by using the dynamic programming. Numerical examples demonstrate the efficacy of the proposed discrete measure.**

***Index Terms*—proximity operator, Moreau enhancement, OSCAR, feature grouping.**

I. Introduction

Highly correlated features may be included in the observations in some applications of sparse regression. When the least absolute shrinkage and selection operator (lasso) [1], which is a standard method of sparse regression, is employed in this situation, only one feature from each group tends to be selected [2]. However, it is often desirable to select all the important groups of features relevant to the observations to achieve better interpretability of the regression results. Such a situation occurs in various fields such as gene expression analysis [3], brain imaging [4], and analysis of protein-protein interaction networks [5].

To extract important groups of the features, many feature grouping methods have been proposed such as the elastic net [2], the fused lasso [6], the clustered lasso [7], and the octagonal shrinkage and clustering algorithm for regression (OSCAR) [8], [9]. Unfortunately, a solution obtained by the elastic net does not have the identical coefficients for the highly correlated features in general, and this may lead to difficulty in interpretation of the group structure [8], [9]. 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 [9], [10]. In contrast to those methods, OSCAR is free from such limitations. Unfortunately, it is known that OSCAR may overpenalize the large pairwise coefficient differences [5], [9], which may cause estimation bias.

To reduce the estimation bias caused by convex regularizers which promote sparsity of a vector or low-rankness of a matrix, the generalized Moreau enhancement (GME) penalty has been proposed in [11][1], and related techniques have been studied intensively [12]–[14]. It is known that the Moreau enhancement parametrically bridges the gap between a direct discrete measure and its convex envelope for some functions. For example, the Moreau enhancement of the ℓ_1 norm (known as the minimax concave (MC) penalty [15], [16]) parametrically bridges the ℓ_0 psuedonorm (which counts the number of nonzero components) and the ℓ_1 norm.

A natural way to reduce the estimation bias of the OSCAR regularizer would be the use of the Moreau-enhanced OSCAR regularizer. Recently, the debiased OSCAR (DOSCAR) shrinkage is proposed as an extension of the single-valued proximity operator of the Moreau-enhanced OSCAR regularizer, and its debiasing effect is studied [17]. However, unlike the ℓ_1 norm, no direct discrete measure corresponding to the OSCAR regularizer is known, to the best of our knowledge. Hence, it is still unclear if the Moreau-enhanced OSCAR regularizer can be seen as an approximation of a discrete measure which has a desirable property.

In this paper, we investigate the discrete measure for debiased feature grouping defined by a limit of the Moreau-enhanced OSCAR regularizer, which we refer to as the LME-OSCAR regularizer. It turns out that the LME-OSCAR regularizer yields smaller values for vectors which are sparse or have group structures (in a sense that some coefficients have identical values). Hence, a sparse or grouped solution is likely to be obtained when the LME-OSCAR regularizer is adopted as a penalty function. This implies that the Moreau-enhanced OSCAR regularizer yields a parametric bridge between the OSCAR regularizer and a desirable discrete measure. We derive an efficient algorithm based on the dynamic pro-

K. Suzuki is now with the Department of Information and Communications Engineering, Tokyo Institute of Technology, Yokohama, 226-8502, Japan. This work was supported by the Grants-in-Aid for Scientific Research (KAKENHI) under Grant Numbers 22KJ2720 and 23K22762.

[1]In [11], a more general penalty function than the GME penalty is proposed, which is applicable to broader scenarios.

gramming to compute the proximity operator of the LME-OSCAR regularizer. In numerical examples, we demonstrate that the LME-OSCAR regularizer significantly outperforms the OSCAR regularizer in the feature grouping task.

II. Preliminaries

This section briefly introduces the notation, definitions, and selected elements of convex analysis.

A. Notation and Definitions

Throughout the paper, let $\mathbb{R}$ and $\mathbb{N}$ denote the set of real numbers and nonnegative integers, respectively. For any $n \in \mathbb{N}_* := \mathbb{N} \setminus \{0\}$, let $\overline{1,n} := \{1, 2, \ldots, n\}$. For any $m, n \in \mathbb{N}_*$, the ith column of $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ is denoted as $\boldsymbol{a}_i$. The matrix transpose is denoted as $(\cdot)^\top$. For any $\boldsymbol{x} \in \mathbb{R}^n$ and $p \geq 1$, we define the ℓ_p norm by $\|\boldsymbol{x}\|_p := (\sum_{i=1}^n |x_i|^p)^{1/p}$.

B. Selected Elements of Convex Analysis

A function $f : \mathbb{R}^n \to (-\infty, +\infty]$ is convex if $f(a\boldsymbol{x} + (1-a)\boldsymbol{\xi}) \leq af(\boldsymbol{x}) + (1-a)f(\boldsymbol{\xi})$ for any $\boldsymbol{x}, \boldsymbol{\xi} \in \mathbb{R}^n$ and any $a \in (0,1)$. A function $f : \mathbb{R}^n \to (-\infty, +\infty] := \mathbb{R} \cup \{+\infty\}$ is proper if $\operatorname{dom} f := \{\boldsymbol{x} \in \mathbb{R}^n \mid f(\boldsymbol{x}) < +\infty\} \neq \emptyset$. A function $f : \mathbb{R}^n \to (-\infty, +\infty]$ is lower-semicontinuous on $\mathbb{R}^n$ if the level set $\operatorname{lev}_{\leq a} f := \{\boldsymbol{x} \in \mathbb{R}^n \mid f(\boldsymbol{x}) \leq a\}$ is closed for any $a \in \mathbb{R}$. Given any proper function $f : \mathbb{R}^n \to (-\infty, +\infty]$, we define the proximity operator of f of index $\gamma > 0$ by

$$\operatorname{Prox}_{\gamma f} : \mathbb{R}^n \to 2^{\mathbb{R}^n} : \boldsymbol{x} \mapsto \operatorname*{argmin}_{\boldsymbol{\xi} \in \mathbb{R}^n} \left(f(\boldsymbol{\xi}) + \frac{1}{2\gamma}\|\boldsymbol{x} - \boldsymbol{\xi}\|_2^2 \right), \tag{1}$$

where $2^{\mathbb{R}^n}$ denotes the power set (the family of all subsets) of $\mathbb{R}^n$. Given a proper lower-semicontinuous convex function $f : \mathbb{R}^n \to (-\infty, +\infty]$, the Moreau envelope of f of index $\gamma > 0$ is defined by [18]–[20]

$$\begin{aligned} {}^{\gamma}f : \mathbb{R}^n \to \mathbb{R} : \boldsymbol{x} &\mapsto \min_{\boldsymbol{\xi} \in \mathbb{R}^n} \left(f(\boldsymbol{\xi}) + \frac{1}{2\gamma}\|\boldsymbol{x} - \boldsymbol{\xi}\|_2^2 \right) \\ &= f(\operatorname{Prox}_{\gamma f}(\boldsymbol{x})) + \frac{1}{2\gamma}\|\boldsymbol{x} - \operatorname{Prox}_{\gamma f}(\boldsymbol{x})\|_2^2. \end{aligned} \tag{2}$$

III. Proposed Discrete Measure and Its Proximity Operator

In this section, we derive a closed-form expression of the LME-OSCAR regularizer, and show that its proximity operator can be computed efficiently. The theoretical result in this paper is presented without proof. An extended version of this work including a full proof will be published elsewhere.

A. A Limit of Moreau-Enhanced OSCAR Regularizer

We consider the task of estimating the sparse coefficient vector $\boldsymbol{x}_\diamond \in \mathbb{R}^n$ from a given input matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ and an observation vector modeled as

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_\diamond + \boldsymbol{\varepsilon} \in \mathbb{R}^m, \tag{3}$$

where $\boldsymbol{\varepsilon} \in \mathbb{R}^m$ is the zero-mean additive white Gaussian noise. The matrix $\boldsymbol{A}$ is assumed to have groups of highly correlated column vectors. OSCAR aims to obtain a sparse coefficient vector in which the coefficients corresponding to the highly correlated vectors are identical. The formulation of OSCAR is defined by [8]

$$\min_{\boldsymbol{x} \in \mathbb{R}^n} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \Omega_{\lambda_1,\lambda_2}^{\mathrm{OSCAR}}(\boldsymbol{x}), \tag{4}$$

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

$$\begin{aligned} \Omega_{\lambda_1,\lambda_2}^{\mathrm{OSCAR}} : \mathbb{R}^n &\to [0, +\infty) \\ \boldsymbol{x} &\mapsto \lambda_1\|\boldsymbol{x}\|_1 + \lambda_2 \sum_{i<j} \max\{|x_i|, |x_j|\}. \end{aligned} \tag{5}$$

For the Moreau enhancement[2] of the OSCAR regularizer

$$(\Omega_{\lambda_1,\lambda_2}^{\mathrm{OSCAR}})_{\gamma^{-1/2}\boldsymbol{I}_n} := \Omega_{\lambda_1,\lambda_2}^{\mathrm{OSCAR}} - {}^{\gamma}\Omega_{\lambda_1,\lambda_2}^{\mathrm{OSCAR}}, \tag{6}$$

it holds by [21, Proposition 12.33] that

$$\lim_{\gamma \to +\infty} (\Omega_{\lambda_1,\lambda_2}^{\mathrm{OSCAR}})_{\gamma^{-1/2}\boldsymbol{I}_n}(\boldsymbol{x}) = \Omega_{\lambda_1,\lambda_2}^{\mathrm{OSCAR}}(\boldsymbol{x}), \ \forall \boldsymbol{x} \in \mathbb{R}^n. \tag{7}$$

We investigate the LME-OSCAR regularizer defined by

$$\Upsilon_{\lambda_1,\lambda_2} := \lim_{\gamma \downarrow 0} 2\gamma^{-1}(\Omega_{\lambda_1,\lambda_2}^{\mathrm{OSCAR}})_{\gamma^{-1/2}\boldsymbol{I}_n}. \tag{8}$$

By (7) and (8), one can see that the Moreau-enhanced OSCAR regularizer parametrically bridges the OSCAR and the LME-OSCAR regularizers. The following proposition shows a closed-form expression of the LME-OSCAR regularizer in the two-dimensional case. Let

$$\mathcal{K}_{\geq,+}^n := \{\boldsymbol{x} \in \mathbb{R}^n \mid x_1 \geq x_2 \geq \ldots \geq x_n \geq 0\}, \tag{9}$$

and $|\boldsymbol{x}|_\downarrow := \boldsymbol{P}(|\boldsymbol{x}|)|\boldsymbol{x}| \in \mathbb{R}^n$, where $\boldsymbol{P}(|\boldsymbol{x}|) \in \mathbb{R}^{n \times n}$ denotes a permutation matrix which sorts the components of $|\boldsymbol{x}| := [|x_1|, |x_2|, \ldots, |x_n|]^\top \in \mathbb{R}^n$ in non-increasing order. Then, it is sufficient to consider the case in which $\boldsymbol{x} \in \mathcal{K}_{\geq,+}^n$ since

$$(\Omega_{\lambda_1,\lambda_2}^{\mathrm{OSCAR}})_{\gamma^{-1/2}\boldsymbol{I}_n}(\boldsymbol{x}) = (\Omega_{\lambda_1,\lambda_2}^{\mathrm{OSCAR}})_{\gamma^{-1/2}\boldsymbol{I}_n}(|\boldsymbol{x}|_\downarrow). \tag{10}$$

Proposition 1 (LME-OSCAR regularizer: two-dimensional case)**.** *Let $n := 2$. Then, for any $\boldsymbol{x} \in \mathcal{K}_{\geq,+}^2$, it holds that*

$$\Upsilon_{\lambda_1,\lambda_2}(\boldsymbol{x}) = \begin{cases} (\lambda_1 + \lambda_2)^2 + \lambda_1^2, & \text{if } x_1 > x_2 > 0, \\ (2\lambda_1 + \lambda_2)^2/2, & \text{if } x_1 = x_2 > 0, \\ (\lambda_1 + \lambda_2)^2, & \text{if } x_1 > x_2 = 0, \\ 0, & \text{if } x_1 = x_2 = 0. \end{cases} \tag{11}$$

Proof. The proof is due to the general result presented in Proposition 2. □

Equation (11) indicates the LME-OSCAR regularizer is a certain discrete function shown in Figure 1. This function yields a smaller value outside the set $\{\boldsymbol{x} \in \mathcal{K}_{\geq,+}^2 \mid x_1 > x_2 > 0\}$. Hence, when (11) is used as a regularizer, the solution $\hat{\boldsymbol{x}}$ is likely to satisfy

$$\hat{x}_1 = \hat{x}_2 \geq 0 \text{ or } \hat{x}_1 \geq \hat{x}_2 = 0, \tag{12}$$

which indicates that a sparse or grouped solution is likely to be obtained. Figure 2 shows the contours of the Moreau-enhanced OSCAR regularizer $(\Omega_{\lambda_1,\lambda_2}^{\mathrm{OSCAR}})_{\gamma^{-1/2}\boldsymbol{I}_n}$ for $\gamma := 2$ in

[2]Let $(\mathcal{X}, \langle\cdot,\cdot\rangle_{\mathcal{X}})$ and $(\mathcal{Z}, \langle\cdot,\cdot\rangle_{\mathcal{Z}})$ be the finite-dimensional real Hilbert spaces. For any proper lower-semicontinuous convex function $\Psi : \mathcal{X} \to \mathcal{X}$ coercive with $\operatorname{dom}\Psi = \mathcal{X}$, and an arbitrary bounded linear operator $\boldsymbol{B} : \mathcal{X} \to \mathcal{Z}$, the generalized-Moreau-enhanced penalty is defined by [11] $\Psi_{\boldsymbol{B}} : \boldsymbol{x} \mapsto \Psi(\boldsymbol{x}) - \min_{\boldsymbol{z} \in \mathcal{X}} \left(\Psi(\boldsymbol{z}) + \|\boldsymbol{B}(\boldsymbol{x} - \boldsymbol{z})\|_{\mathcal{Z}}^2/2\right)$.

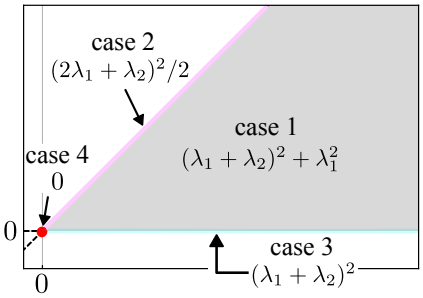

Fig. 1: Visualization of the four cases of (11).

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 (12). The following proposition extends the result of Proposition 1 to the n-dimensional case for $n \geq 2$.

Proposition 2 (LME-OSCAR regularizer: general case)**.** *Let* $\boldsymbol{w} \coloneqq [w_1, w_2, \ldots, w_n]^\top \in \mathbb{R}^n$ *such that*

$$w_i = \lambda_1 + \lambda_2(n-i), \ \forall i \in \overline{1,n}. \tag{13}$$

Let $\mathcal{K}^n_> \coloneqq \{\boldsymbol{x} \in \mathbb{R}^n \mid x_1 > x_2 > \ldots > x_n\}$, $\chi_{\mathbb{R}_{++}} : \mathbb{R}_{++} \to \{0,1\} : x \mapsto \begin{cases} 1, & \text{if } x \in \mathbb{R}_{++}, \\ 0, & \text{if } x \notin \mathbb{R}_{++}, \end{cases}$ *and the set* S_l *with* $\mathrm{card}(S_l) \geq 2$ *be the* l*th group of consecutive indices for* $l = 1, 2, \ldots, q$ *with* $q \in \overline{1,n}$*, such that*

$$x_j = x_k, \ \forall j,k \in S_l, \ \text{and}, \tag{14}$$

$$x_j \neq x_k, \ \forall j \in S_l, \ \forall k \in \overline{1,n} \setminus S_l. \tag{15}$$

For example, it holds that $S_1 = \{3,4,5\}$, $S_2 = \{7,8\}$, $S_3 = \{9,10\}$ *when*

$$x_1 > x_2 > \underbrace{x_3 = x_4 = x_5}_{\textit{first group}} > x_6 > \underbrace{x_7 = x_8}_{\textit{second group}} > \underbrace{x_9 = x_{10}}_{\textit{third group}} > x_{11}. \tag{16}$$

Then, it holds for any $n \in \mathbb{N}_*$ *and* $\boldsymbol{x} \in \mathcal{K}^n_{\geq,+}$ *that*

$$\Upsilon_{\lambda_1,\lambda_2}(\boldsymbol{x}) = \begin{cases} \|\boldsymbol{w}\|_2^2, & \text{if } \boldsymbol{x} \in \mathbb{R}^n_{++} \cap \mathcal{K}^n_> \quad \textit{(case 1)}, \\ \|\boldsymbol{w}\|_2^2 - \displaystyle\sum_{l=1}^{q} \sum_{j \in S_l} \left(w_j - \frac{\sum_{k \in S_l} w_k}{\mathrm{card}(S_l)} \right)^2, & \\ & \text{if } \boldsymbol{x} \in \mathbb{R}^n_{++} \cap (\mathcal{K}^n_>)^{\mathsf{c}} \quad \textit{(case 2)}, \\ \|\boldsymbol{w}\|_2^2 - w_n^2, & \text{if } \boldsymbol{x} \in (\mathbb{R}^n_{++})^{\mathsf{c}} \cap \mathcal{K}^n_> \quad \textit{(case 3a)}, \\ \|\boldsymbol{w}\|_2^2 - \displaystyle\sum_{l=1}^{q} \sum_{j \in S_l} \left(w_j - \frac{\sum_{k \in S_l} w_k}{\mathrm{card}(S_l)} \chi_{\mathbb{R}_{++}}(x_j) \right)^2 & \\ \quad - w_n^2 \chi_{\mathbb{R}_{++}}(x_{n-1}), & \\ & \text{if } \boldsymbol{x} \in \mathcal{K}^n_{\geq,+} \setminus (\mathbb{R}^n_{++} \cup \mathcal{K}^n_> \cup \{\boldsymbol{0}\}) \\ & \textit{(case 3b)}, \\ 0, & \text{if } \boldsymbol{x} = \boldsymbol{0} \quad \textit{(case 4)}, \end{cases} \tag{17}$$

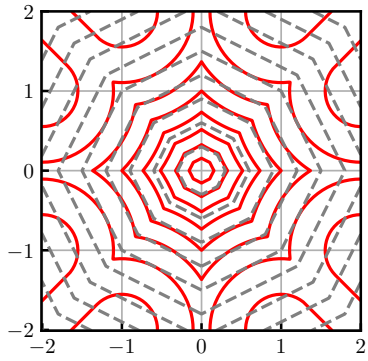

Fig. 2: Contours of the Moreau-enhanced OSCAR regularizer (red) for $\lambda_1 = \lambda_2 \coloneqq 0.5$ and $\gamma \coloneqq 2$ and the OSCAR regularizer (gray) in the two-dimensional case.

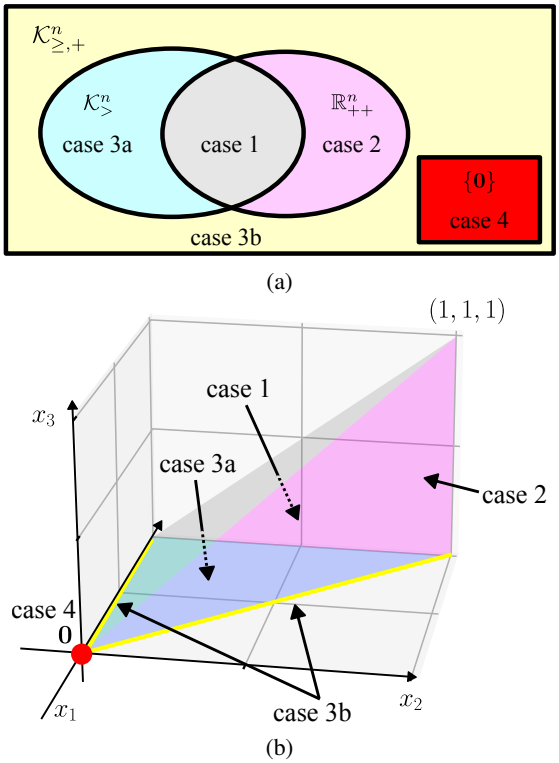

Fig. 3: (a) Inclusion relation among the five cases of (17). (b) Visualization of the five cases in the three-dimensional case.

where S^{c} *denotes the compliment set of any set* $S \subset \mathbb{R}^n$.

The inclusion relation of the cases of (17) and visualization of the cases in the three-dimensional case are shown in Figures 3(a) and 3(b), respectively. Unlike (11), (17) is divided into five cases (case 3b does not exist in the two-dimensional case).

Proposition 2 indicates that, the LME-OSCAR regularizer is bounded above by the constant $\|\boldsymbol{w}\|_2^2$, and it is strictly smaller than this bound on the set $(\mathcal{K}^n_>)^{\mathsf{c}} \cup (\mathbb{R}^n_{++})^{\mathsf{c}}$. Since $\boldsymbol{x} \in \mathcal{K}^n_{\geq,+}$ by assumption, $\boldsymbol{x} \in \mathcal{K}^n_>$ implies that $\boldsymbol{x}$ has no group of equal indices, and $\boldsymbol{x} \in \mathbb{R}^n_{++}$ implies that $\boldsymbol{x}$ is dense. Hence, when (17) is used as a penalty, a sparse or grouped solution is likely to be obtained.

B. Proximity Operator of the LME-OSCAR Regularizer $\Upsilon_{\lambda_1,\lambda_2}$

For any $\boldsymbol{x} \in \mathbb{R}^n$, the proximity operator of $\Upsilon_{\lambda_1,\lambda_2}$ defined in (17) can be computed efficiently as in the following two propositions.

Proposition 3 (Proximity operator of the LME-OSCAR regularizer)**.** *For any $\boldsymbol{x} \in \mathbb{R}^n$, it holds that*

$$\mathrm{Prox}_{\Upsilon_{\lambda_1,\lambda_2}}(\boldsymbol{x}) = \mathrm{Sign}(\boldsymbol{x}) \circ \boldsymbol{P}(|\boldsymbol{x}|)^\top \mathrm{Prox}_{\Upsilon_{\lambda_1,\lambda_2}}(|\boldsymbol{x}|_\downarrow), \quad (18)$$

where $\mathrm{Sign} : \mathbb{R}^n \to \{-1,1\}^n : \boldsymbol{x} \mapsto [\mathrm{sign}(x_1), \mathrm{sign}(x_2), \ldots, \mathrm{sign}(x_n)]^\top$ *with* $\mathrm{sign}(a) \coloneqq 1$ *if* $a \geq 0$*;* $\mathrm{sign}(a) \coloneqq -1$ *otherwise.*

The vector $\mathrm{Prox}_{\Upsilon_{\lambda_1,\lambda_2}}(|\boldsymbol{x}|_\downarrow)$, where $|\boldsymbol{x}|_\downarrow \in \mathcal{K}^n_{\geq,+}$, can be computed by using the following proposition.

Proposition 4 (Computation of $\mathrm{Prox}_{\Upsilon_{\lambda_1,\lambda_2}}(|\boldsymbol{x}|_\downarrow)$ in (18))**.** *Let $\boldsymbol{x} \in \mathcal{K}^n_{\geq,+}$. Define $\boldsymbol{w}$ as in* (13)*. Let $\mathscr{S}_{i,\mathfrak{q}} \coloneqq \{\overline{(k_j, l_j)}_{j=1}^{\mathfrak{q}} \subset \overline{1,i}^{\mathfrak{q}} \mid 1 \leq k_1 < l_1 < k_2 < l_2 < \cdots < k_{\mathfrak{q}} < l_{\mathfrak{q}} \leq i\}$ for any $i \in \overline{2,n}$ and $\mathfrak{q} \in \overline{1, \lceil i/2 \rceil}$, where $\lceil \cdot \rceil$ is the ceiling function. Let*

$$\eta_{\min} \coloneqq \min_{i \in \overline{1,6}} \eta_i, \quad (19)$$

where $\eta_1 \coloneqq \|\boldsymbol{w}\|_2^2$*,* $\eta_2 \coloneqq \|\boldsymbol{w}\|_2^2 - w_n^2 + (1/2)x_n^2$*,* $\eta_3 \coloneqq \|\boldsymbol{w}\|_2^2 - d_n$*,* $\eta_4 \coloneqq \|\boldsymbol{w}\|_2^2 - d_{n-1} - w_n^2 + (1/2)x_n^2$*,* $\eta_5 \coloneqq \|\boldsymbol{w}\|_2^2 - \max_{i \in \overline{2,n-1}}\{d_{i-1} + \sum_{j=i}^n (w_j^2 - (1/2)x_j^2)\}$*, and* $\eta_6 \coloneqq (1/2)\|\boldsymbol{x}\|_2^2$*. Here,*

$$d_i \coloneqq \begin{cases} 0, & \text{if } i = 1, \\ \max\limits_{\mathfrak{q} \in \overline{1, \lceil i/2 \rceil}} \max\limits_{(\mathfrak{S}_j)_{j=1}^{\mathfrak{q}} \in \mathscr{S}_{i,\mathfrak{q}}} \sum\limits_{l=1}^{\mathfrak{q}} v_{\mathfrak{S}_j}, & \text{if } i \in \overline{2,n}, \end{cases} \quad (20)$$

where

$$v_{\mathfrak{S}_j} \coloneqq \sum_{j \in \mathfrak{S}_j} \left[\left(w_j - \frac{\sum_{k \in \mathfrak{S}_j} w_k}{\mathrm{card}(\mathfrak{S}_j)} \right)^2 - \frac{1}{2} \left(x_j - \frac{\sum_{k \in \mathfrak{S}_j} x_k}{\mathrm{card}(\mathfrak{S}_j)} \right)^2 \right]. \quad (21)$$

Let $\hat{\mathfrak{q}} \in \mathrm{argmax}_{\mathfrak{q} \in \overline{1, \lceil i/2 \rceil}} \max_{(\mathfrak{S}_j)_{j=1}^{\mathfrak{q}} \in \mathscr{S}_{i,\mathfrak{q}}} \sum_{l=1}^{\mathfrak{q}} v_{\mathfrak{S}_j}$ *and* $(\hat{\mathfrak{S}}_j)_{j=1}^{\hat{\mathfrak{q}}} \in \mathrm{argmax}_{(\mathfrak{S}_j)_{j=1}^{\hat{\mathfrak{q}}} \in \mathscr{S}_{i,\hat{\mathfrak{q}}}} \sum_{l=1}^{\hat{\mathfrak{q}}} v_{\mathfrak{S}_j}$*. Then, it holds that*

$$\mathrm{Prox}_{\Upsilon_{\lambda_1,\lambda_2}}(\boldsymbol{x}) = \bigcup_{i=1}^{6} C_i, \quad (22)$$

where $C_i \coloneqq \begin{cases} \{\boldsymbol{p}^{(i)}\}, & \text{if } \eta_{\min} = \eta_i, \\ \emptyset, & \text{otherwise}, \end{cases}$ *for any* $i \in \overline{1,6}$*. Here,* $\boldsymbol{p}^{(1)} \coloneqq \boldsymbol{x}$*,* $p_j^{(2)} \coloneqq \begin{cases} 0, & \text{if } j = n, \\ x_j, & \text{if } j \neq n, \end{cases}$ $p_j^{(3)} \coloneqq \begin{cases} \frac{\sum_{k \in \hat{\mathfrak{S}}_l} x_k}{\mathrm{card}(\hat{\mathfrak{S}}_l)}, & \text{if } j \in \hat{\mathfrak{S}}_l, \\ x_j, & \text{if } j \notin \bigcup_{l=1}^{\hat{\mathfrak{q}}} \hat{\mathfrak{S}}_l, \end{cases}$ $p_j^{(4)} \coloneqq \begin{cases} 0, & \text{if } j = n, \\ p_j^{(3)}, & \text{if } j \neq n, \end{cases}$ $p_j^{(5)} \coloneqq \begin{cases} 0, & \text{if } j \in \hat{\mathfrak{S}}_{\hat{\mathfrak{q}}}, \\ p_j^{(3)}, & \text{if } j \notin \hat{\mathfrak{S}}_{\hat{\mathfrak{q}}}, \end{cases}$ *and* $\boldsymbol{p}^{(6)} \coloneqq \boldsymbol{0}$ *for any* $j \in \overline{1,n}$ *and* $l \in \overline{1, \hat{\mathfrak{q}}}$*.*

Algorithm 1 Dynamic programming for $(d_i)_{i=1}^n$ and $(\hat{\mathfrak{S}}_l)_{l=1}^{\hat{\mathfrak{q}}}$

Input: $\boldsymbol{x} \in \mathbb{R}^n$, $\lambda_1, \lambda_2 > 0$
Compute $\boldsymbol{w}$ by (13)
$d_1 \coloneqq 0$, $\mathcal{S}_1 \coloneqq \emptyset$
for $i = 2, 3, \ldots, n$ **do**
  $t \coloneqq d_{i-1}$
  for $j = 1, 2, \ldots, i-1$ **do**
    Compute $v_{\overline{j,i}}$ by (21)
    if $t < d_{\max\{j-1,1\}} + v_{\overline{j,i}}$ **then**
      $t \coloneqq d_{\max\{j-1,1\}} + v_{\overline{j,i}}$
      $\mathcal{S}_i \coloneqq \mathcal{S}_{\max\{j-1,1\}} \cup \{j, i\}$
    end if
  end for
  $d_i \coloneqq t$
end for
$\hat{\mathfrak{q}} \coloneqq \mathrm{card}(\mathcal{S}_n)/2$
for $l = 1, 2, \ldots, \hat{\mathfrak{q}}$ **do**
  $\hat{\mathfrak{S}}_l \coloneqq \overline{s_{2l-1}, s_{2l}}$, where s_l is the lth smallest element of $\mathcal{S}_n$.
end for
Output: $(d_i)_{i=1}^n$, $(\hat{\mathfrak{S}}_l)_{l=1}^{\hat{\mathfrak{q}}}$

For any $\boldsymbol{x} \in \mathcal{K}^n_{\geq,+}$, $\mathrm{Prox}_{\Upsilon_{\lambda_1,\lambda_2}}(\boldsymbol{x})$ given in (22) depends on $(d_i)_{i=1}^n$ and $(\hat{\mathfrak{S}}_l)_{l=1}^{\hat{\mathfrak{q}}}$, which can be efficiently computed by the dynamic programming [22] as shown in Algorithm 1. Here, Algorithm 1 is derived from the following recurrence:

$$d_i = \begin{cases} 0, & \text{if } i = 1, \\ \max\left\{ d_{i-1}, \max\limits_{j \in \overline{1,i-1}} \left(d_{\max\{j-1,1\}} + v_{\overline{j,i}} \right) \right\}, & \text{if } i \in \overline{2,n}. \end{cases} \quad (23)$$

One can obtain $\hat{\mathfrak{q}}$ and $(\hat{\mathfrak{S}}_l)_{l=1}^{\hat{\mathfrak{q}}}$ simultaneously in Algorithm 1. Note that $\mathrm{card}(\mathcal{S}_i)/2$ is an integer for all $i \in \overline{1,n}$. Let $R_{\lambda_1,\lambda_2} : \mathbb{R}^n \to \mathbb{R}^n$ be a selection of $\mathrm{Prox}_{\Upsilon_{\lambda_1,\lambda_2}}$, *i.e.*, $R_{\lambda_1,\lambda_2}(\boldsymbol{x}) \in \mathrm{Prox}_{\Upsilon_{\lambda_1,\lambda_2}}(\boldsymbol{x})$ for any $\boldsymbol{x} \in \mathbb{R}^n$. Then, the computational cost to obtain $R_{\lambda_1,\lambda_2}(\boldsymbol{x})$ scales in $\mathcal{O}(n^2)$.

IV. Numerical Examples

We consider the sparse estimation task, particularly when there are groups of highly correlated features. The standard linear model in (3) with $m \coloneqq 80$ and $n \coloneqq 30$ is used. The noise vector $\boldsymbol{\varepsilon} \in \mathbb{R}^m$ is generated i.i.d. from the zero-mean Gaussian distribution with signal-to-noise ratio (SNR) 20 dB. We consider the following two toy datasets, which are similar to those used in [2], [8], [9]:

A. The column vectors of $\boldsymbol{A} \in \mathbb{R}^{80 \times 30}$ are generated as $\boldsymbol{a}_i \coloneqq \begin{cases} \tilde{\boldsymbol{a}}_1 + \boldsymbol{\epsilon}_i, & \text{if } i \in G_1 \coloneqq \overline{1,5}, \\ \tilde{\boldsymbol{a}}_2 + \boldsymbol{\epsilon}_i, & \text{if } i \in G_2 \coloneqq \overline{6,10}, \\ \tilde{\boldsymbol{a}}_3 + \boldsymbol{\epsilon}_i, & \text{if } i \in G_3 \coloneqq \overline{11,15}, \\ \boldsymbol{\epsilon}_i, & \text{if } i \in G_4 \coloneqq \overline{16,30}, \end{cases}$ where the components of $\tilde{\boldsymbol{a}}_i \in \mathbb{R}^{80}$ $(i \in \overline{1,3})$ are generated i.i.d. from the standard Gaussian distribution, and those of $\boldsymbol{\epsilon}_i \in \mathbb{R}^{80}$ $(i \in \overline{1,30})$ are generated i.i.d. from $N(0, \sigma_\epsilon^2)$ with $\sigma_\epsilon > 0$. The coefficient vector $\boldsymbol{x}_\diamond \in \mathbb{R}^n$ is set as

$$x_{\diamond,i} \coloneqq \begin{cases} 3, & \text{if } i \in G_1, \\ 2, & \text{if } i \in G_2, \\ 1.5, & \text{if } i \in G_3, \\ 0, & \text{if } i \in G_4. \end{cases}$$

B. The difference from dataset A is the following: $G_1 \coloneqq \{1, 4, 7, 10, 13\}$, $G_2 \coloneqq \{2, 5, 8, 11, 14\}$, $G_3 \coloneqq \{3, 6, 9, 12, 15\}$.

For this task, we consider the proximal gradient method to solve the least square loss regularized by $\Upsilon_{\lambda_1,\lambda_2}$ given in (17) as follows:

$$\boldsymbol{x}_{k+1} \coloneqq R_{\lambda_1,\lambda_2}(\boldsymbol{x}_k - \mu \boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{x}_k - \boldsymbol{y})),\ k \in \mathbb{N}, \qquad (24)$$

where $\mu > 0$ is a step size. The algorithm is initialized to $\boldsymbol{x}_0 \coloneqq \boldsymbol{0}_n$. We compare the proposed method with the methods to solve lasso [1], the MC penalty [15], [16], the fused lasso [6], and OSCAR [8]. The hyperparameter for the MC penalty is chosen to guarantee the convexity of the cost function. All the hyperparameters are tuned to attain the best performance. The evaluation metric is the system mismatch defined by $\|\hat{\boldsymbol{x}} - \boldsymbol{x}_\diamond\|_2^2 / \|\boldsymbol{x}_\diamond\|_2^2$, where $\hat{\boldsymbol{x}}$ is an estimate of $\boldsymbol{x}_\diamond$. The results are averaged over 300 trials.

Figure 4 shows the system mismatch across σ_ϵ under dataset A and B. The parameter σ_ϵ controls the correlation of the column vectors of $\boldsymbol{A}$: a large σ_ϵ corresponds to small correlations among $\boldsymbol{a}_i$'s, and vice versa. It can be seen that the proposed method outperforms the other methods in a wide range. Especially when the correlation is high, the proposed method successfully reduces the estimation bias, unlike OSCAR. We finally mention that the proposed method achieves the best overall performance for both datasets, while the performance of the fussed lasso degrades for dataset B significantly as it assumes the active coefficients to be consecutive.

V. Conclusion

We investigated the LME-OSCAR regularizer, which is defined by a limit of the Moreau-enhanced OSCAR regularizer. It turned out that the LME-OSCAR regularizer induces sparse or grouped solutions effectively while reducing the estimation bias. As a result, it was clarified that the Moreau-enhanced OSCAR regularizer can be seen as an approximation of a desirable discrete measure. An efficient way of computing the proximity operator of the proposed discrete measure was derived by using the dynamic programming. The bias-reducing effect of the proposed method for feature grouping was demonstrated by simulations.

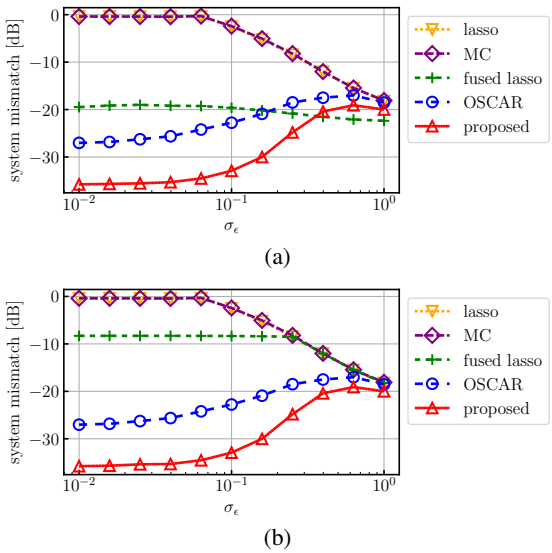

Fig. 4: System mismatch across σ_ϵ under (a) dataset A and (b) dataset B.

References

[1] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Stat. Soc. B*, vol. 58, no. 1, pp. 267–288, 1996.

[2] 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.

[3] M. Dettling and P. Bühlmann, "Finding predictive gene groups from microarray data," *J. Multivariate Anal.*, vol. 90, no. 1, pp. 106–131, 2004.

[4] 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.

[5] 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.

[6] 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.

[7] Y. She, "Sparse regression with exact clustering," *Electron. J. Stat.*, vol. 4, pp. 1055–1096, 2010.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] A. Parekh and I. W. Selesnick, "Enhanced low-rank matrix approximation," *IEEE Signal Process. Lett.*, vol. 23, no. 4, pp. 493–497, 2016.

[15] C. H. Zhang, "Nearly unbiased variable selection under minimax concave penalty," *Ann. Statist.*, vol. 38, no. 2, pp. 894–942, Apr. 2010.

[16] I. Selesnick, "Sparse regularization via convex analysis," *IEEE Trans. Signal Process.*, vol. 65, no. 17, pp. 4481–4494, Sep. 2017.

[17] K. Suzuki and M. Yukawa, "External division of two proximity operators: An application to signal recovery with structured sparsity," in *Proc. Int. Conf. Acoust., Speech, Signal Process.* IEEE, 2024, pp. 9471–9475.

[18] J. J. Moreau, "Décomposition orthogonale d'un espace hilbertien selon deux cônes mutuellement polaires," *Comptes rendus hebdomadaires des séances de l'Académie des sciences*, vol. 255, pp. 238–240, 1962.

[19] J. J. Moreau, "Fonctions convexes duales et points proximaux dans un espace hilbertien," *C. R. Acad. Sci. Paris Ser. A Math.*, vol. 255, pp. 2897–2899, 1962.

[20] 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.

[21] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, Springer, New York, 2nd edition, 2017.

[22] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*, MIT press, 2nd edition, 2022.