Linearly involved generalized Moreau enhanced models and their proximal splitting algorithm under overall convexity condition

Jiro Abe, Masao Yamagishi and Isao Yamada

Department of Information and Communications Engineering, Tokyo Institute of Technology, S3-60, Meguro-ku, Tokyo, 152-8552, Japan

E-mail: abe@sp.ce.titech.ac.jp, myamagi@sp.ce.titech.ac.jp and isao@sp.ce.titech.ac.jp

Received 5 June 2019, revised 22 October 2019 Accepted for publication 6 November 2019 Published 11 February 2020

Abstract
The convex envelopes of the direct discrete measures, for the sparsity of vectors or for the low-rankness of matrices, have been utilized extensively as practical penalties in order to compute a globally optimal solution of the corresponding regularized least-squares models. Motivated mainly by the ideas in Zhang (2010 Ann. Stat. 38 894–942; Selesnick 2017 IEEE Trans. Signal Process. 65 4481–94; Yin et al 2019 IEEE Trans. Signal Process. 67 2595–607) to exploit nonconvex penalties in the regularized least-squares models without losing their overall convexities, this paper presents the linearly involved generalized Moreau enhanced (LiGME) model as a unified extension of such utilizations of nonconvex penalties. The proposed model can admit multiple nonconvex penalties without losing its overall convexity and thus is applicable to much broader scenarios in the sparsity-rank-aware signal processing. Under the general overall-convexity condition of the LiGME model, we also present a novel proximal splitting type algorithm of guaranteed convergence to a globally optimal solution. Numerical experiments in typical examples of the sparsity-rank-aware signal processing demonstrate the effectiveness of the LiGME models and the proposed proximal splitting algorithm.

Keywords: convex optimization, nonconvex penalty, generalized minimax concave penalty function, linearly involved convexity-preserving models, proximal splitting, nonexpansive operator, signal recovery

(Some figures may appear in colour only in the online journal)
1. Introduction

Many tasks in inverse problems for data sciences and engineerings (see, e.g. [7, 8, 11, 27, 32, 43, 44, 59, 61] and references therein), including signal processing and machine learning, have been studied as estimations of an unknown vector \( x^* \in \mathcal{X} \) from the observed data \( y \in \mathcal{Y} \) that follows the linear regression model:

\[
y = Ax^* + \varepsilon,
\]

where \( (\mathcal{X}, \langle \cdot, \cdot \rangle_{\mathcal{X}}, \| \cdot \|_{\mathcal{X}}) \) and \( (\mathcal{Y}, \langle \cdot, \cdot \rangle_{\mathcal{Y}}, \| \cdot \|_{\mathcal{Y}}) \) are finite dimensional real Hilbert spaces. \( A : \mathcal{X} \to \mathcal{Y} \) is a known bounded linear operator and \( \varepsilon \in \mathcal{Y} \) is an unknown noise vector. A common approach for such estimation problems is to solve the regularized least-squares minimization problem:

\[
\min_{x \in \mathcal{X}} J_{\Psi \circ \mathcal{L}}(x) := \frac{1}{2} \| y - Ax \|_{\mathcal{Y}}^2 + \mu \Psi \circ \mathcal{L}(x), \quad \mu > 0,
\]

where \( \frac{1}{2} \| y - Ax \|_{\mathcal{Y}}^2 \) is the least-squares term that measures the distance between \( y \) and \( Ax \), \( \Psi \circ \mathcal{L} \) is a regularizer (or a penalty) designed strategically, e.g. based on a prior knowledge on \( x^* \), to obtain its better estimate as a minimizer of \( J_{\Psi \circ \mathcal{L}} \) with a certain real Hilbert space \( (\mathcal{Z}, \langle \cdot, \cdot \rangle_{\mathcal{Z}}, \| \cdot \|_{\mathcal{Z}}) \), a certain bounded linear operator \( \mathcal{L} : \mathcal{X} \to \mathcal{Z} \), a certain function \( \Psi : \mathcal{Z} \to [0, \infty) \) and a regularization parameter \( \mu > 0 \) providing the trade-off between the least-squares term and the regularizer. To study optimization algorithms for (2) with general \( \Psi \) which is not necessarily differentiable at every \( x \in \mathcal{X} \), the decoupled expression of \( \Psi \) and \( \mathcal{L} \) in (2) is very crucial even if \( \Psi \) is convex because we usually need many nontrivial ideas to deal with \( \Psi \) and \( \mathcal{L} \) separately. Design of \( (\Psi, \mathcal{L}, \mu) \) depends on applications as well as mathematical tractability for the optimization task. Typical examples are found as follows.

Example 1.

(a) Ridge regression or Tikhonov type regularization By letting \( \Psi(\cdot) = \| \cdot \|_{\mathcal{Z}}^2 \) and \( \mathcal{L} = \text{Id} \), the problem (2) reproduces a classical regularization known as the ridge regression estimator [35, 36], essentially based on common idea of the so-called Tikhonov type regularization [63, 64] which has been extensively studied and extended [7, 8, 30, 33, 34].

(b) \( \ell_1 \) regularization By letting \( \mathcal{X} = \mathcal{Z} := \mathbb{R}^n \), \( \Psi(\cdot) = \| \cdot \|_1 \) (\( \ell_1 \)-norm) and \( \mathcal{L} = \text{Id} \), the problem (2) reproduces the \( \ell_1 \) regularization problem which has been a standard model in applications demanding sparse estimates \( x = (x_1, \ldots, x_n) \in \mathcal{X} \) of \( x^* \). For example, in a classification task based on \( n \) features corresponding to the components of \( x^* \), not all features are informative, hence we want to keep the most informative components and make the less informative ones equal to zero. Since the naive approach by choosing \( \Psi(x) = \| x \|_0 \) where \( \| x \|_0 \) stands for the number of nonzero components of \( x \), makes the problem (2) in general NP-hard, its convex envelope \( \Psi(x) = \| x \|_1 := \sum_{i=1}^n |x_i| \) has been utilized in many applications. Although this type of regularizations appeared in 70s at the latest in seismology, e.g. [15, 55, 60], it has attracted an intensive revived interest in statistics [62], which addressed the LASSO (least absolute shrinkage and selection operator) task, as well as in signal processing and machine learning, in particular in compressed sensing [12, 24] and related sparsity aware applications [27, 61].

(c) Linearly involved \( \ell_p \) regularization/Wavelet-based regularization/total-variation based regularization By letting \( \mathcal{X} = \mathbb{R}^n \), \( \mathcal{Z} = \mathbb{R}^l \), \( \Psi(z) = (\| z \|_p)^p := \left( \sqrt[l]{\sum_{i=1}^{l} |z_i|^p} \right)^p \) \((p \geq 1)\) for \( z := (z_1, \ldots, z_l) \in \mathbb{R}^l \), the problem (2) reproduces the linearly involved \( \ell_p \)
regularizations. For example, setting $\mathcal{L} = W$, where $W$ is a wavelet transform matrix, the problem (2) reproduces the so-called wavelet-based regularization, e.g. in [22, 59]. If we set $\Psi(\cdot) = \| \cdot \|_1$, and $\mathcal{L} = D$, where $D$ is the first order differential operator (see (38)), the problem (2) reproduces the so-called convex total variation (TV) regularization [54]. The choices of $\Psi(\cdot) = (\| \cdot \|_p)^p (1 \leq p < 2)$, in such applications, have been preferred to $p = 2$ because smaller $p$ is more effective than $p = 2$ in order to promote the sparsity of $(\mathcal{L} \alpha)(x)$ in (2) and also because the choice $0 \leq p < 1$ looses the convexity of the function $\Psi$, which usually makes it very hard to find a global minimizer of $J_{\Psi, \Omega}$. The great success of the model $J_{\Psi, \Omega}$ with $\Psi(\cdot) = (\| \cdot \|_p)^p (1 \leq p < 2)$ especially for large scale applications has been achieved by the modern computational techniques, e.g. proximal splitting [6, 17, 26, 52, 53].

(d) (Regularized least-squares with multiple penalties) Thanks to the remarkable expressive ability of the abstract Hilbert space, the simple form of the regularized least-squares minimization problem in (2) is very flexible. For example, by letting $\mathcal{X} = \mathbb{R}^{m \times n} \times \mathbb{R}^{m \times n} = \{ x = (z_1, z_2), z_i \in \mathbb{R}^{m \times n} | i = 1, 2 \}$ equipped with the addition $\mathcal{X} \times \mathcal{X} \to \mathcal{X} : (x, y) \mapsto (x_1 + y_1, x_2 + y_2)$, the scalar multiplication $\mathbb{R} \times \mathcal{X} \to \mathcal{X} : (\alpha, x) \mapsto (\alpha x_1, \alpha x_2)$, the inner product $\langle \cdot, \cdot \rangle_{\mathcal{X}} : (x, y) \mapsto \text{tr}(x_1 y_1) + \text{tr}(x_2 y_2)$, we can use (2) for estimation of a pair of matrices. Moreover, the form (2) covers seemingly much more general case:

$$\min_{x \in \mathcal{X}} J_{\Psi, \Omega}(x) := \frac{1}{2} \| y - Ax \|_Y^2 + \sum_{i=1}^{M} \mu_i \Psi^{(i)}(\alpha_i x),$$

where multiple penalties are employed in terms of real Hilbert spaces $(\mathcal{Z}_i, \langle \cdot, \cdot \rangle_{\mathcal{Z}_i}, \| \cdot \|_{\mathcal{Z}_i})$, functions $\Psi^{(i)} : \mathcal{Z}_i \to (-\infty, \infty]$, bounded linear operators $\mathcal{L}_i : \mathcal{X} \to \mathcal{Z}_i$, and weights $\mu_i > 0 (i = 1, \ldots, M)$. This fact can be understood through the following simple translation (see, e.g. [16, 29, 50, 51, 67, 68]) of (3) into the form (2) by redefining a new Hilbert space

$$(\mathcal{Z} := Z_1 \times \cdots \times Z_M = \{ z = (z_1, \ldots, z_M) | z_i \in \mathcal{Z}_i (i = 1, \ldots, M) \})$$

equipped with the addition $\mathcal{Z} \times \mathcal{Z} \to \mathcal{Z} : (x, y) \mapsto (x_1 + y_1, \ldots, x_M + y_M)$, the scalar multiplication $\mathbb{R} \times \mathcal{Z} \to \mathcal{Z} : (\alpha, x) \mapsto (\alpha z_1, \ldots, \alpha z_M)$, and the inner product $(x, y) \mapsto \langle x, y \rangle_{\mathcal{Z}} := \sum_{i=1}^{M} \langle x_i, y_i \rangle_{\mathcal{Z}_i}$, and by introducing a new function

$$\Psi := \sum_{i=1}^{M} \mu_i \Psi^{(i)} : \mathcal{Z} \to (-\infty, \infty] : z = (z_1, \ldots, z_M) \mapsto \sum_{i=1}^{M} \mu_i \Psi^{(i)}(z_i),$$

together with a new bounded linear operator

$$\mathcal{L} : \mathcal{X} \to \mathcal{Z} : x \mapsto (\mathcal{L}_1 x, \ldots, \mathcal{L}_M x).$$

For example, by letting $\mathcal{X} = \mathbb{R}^{m \times n}$ with $\langle \cdot, \cdot \rangle_{\mathcal{X}} : (X, Y) \mapsto \text{tr}(X^\top Y)$, $\mathcal{Z} = \mathbb{R}^{M \times N}$, with $\langle \cdot, \cdot \rangle_{\mathcal{Z}} : (X_i, Y_i) \mapsto \text{tr}(X_i^\top Y_i)$, we can promote multiple desired features of $X \in \mathbb{R}^{m \times n}$ flexibly by the model (3) with $(\Psi^{(i)}, \mathcal{L}_i, \mu_i) (i = 1, 2, \ldots, M)$.

(e) (Convexity-preserving nonconvex penalties) The convexity is certainly a key for global optimization. Indeed, the popularity of $\| \cdot \|_1$ in (b) and (c) has been supported strongly by the fact that it is a convex envelope of $\| \cdot \|_0$, i.e. $\| \cdot \|_1$ is the largest convex minorant of $\| \cdot \|_0$ in a vicinity of $0 \in \mathbb{R}^d$. However restricting the choice of function $\Psi$ within convex functions is not the only realistic compromise for ensuring the convexity of $J_{\Psi, \Omega}$ in the
problem (2). For example, by designing strategically a regularizer \( \Psi \circ \mathcal{L} \) combined with the least-squares term in (2), we could have alternative possibility to achieve the overall convexity of (2), i.e. the convexity of \( J_{\Psi \circ \mathcal{L}} \). The so-called convexity-preserving nonconvex penalties were introduced, in late 80s by Blake and Zisserman [9], and followed for example by Nikolova [45–47], as nonconvex regularizers that can maintain the overall convexity after combined with some convex data-fidelity terms. For recent developments of the convexity-preserving nonconvex penalties, see [5, 13, 23, 39–42, 57, 58] and references therein. Most of these works rely on certain strong convexity assumptions in the least squares term, which corresponds to the assumption for the nonsingularity of \( A^*A \) in the scenario of (2), where \( A^* \) stands for the adjoint operator of \( A \). An exceptional example, which is free from such an assumption, has been introduced by Selesnick [56] as the \textit{generalized minimax concave (GMC) penalty function}\(^2\)

\[
(\| \cdot \|_B(\cdot) := \| \cdot \|_1 - \min_{v \in \mathbb{R}^n} \left[ \| \cdot \|_1 + \frac{1}{2} \| B(\cdot - v) \|_2^2 \right] )
\]

with a parameter \( B \in \mathbb{R}^{q \times n} \). The GMC penalty function is a parameterized multidimensional extension of the \textit{minimax concave (MC) penalty function} [71] (see also [4, 28])\(^3\). It is known that (i) the GMC penalty function \( (\| \cdot \|_1)_B \) is nonconvex except for \( (\| \cdot \|_1)_{\mathbb{R}^n} = \| \cdot \|_1 \) (see remark 3(ii)); (ii) for any \( A \in \mathbb{R}^{m \times n} \), \( (\| \cdot \|_1)_B \) can maintain the overall convexity of \( J_{(\| \cdot \|_1)_B} \) in (2) if \( A^*A - \mu B^*B \succeq O_n \) is satisfied (see proposition 1(b), remark 3(iii), and [56, theorem 1]).

The GMC penalty \( (\| \cdot \|_1)_B \) has great potential for dealing with many nonconvex variations of \( \| \cdot \|_1 \) under single umbrella of the modern convex analysis. Indeed, as will be seen in example 2, the GMC function can serve as a parametric penalty which bridges the gap between the direct discrete measure of sparsity and its convex envelope function. Moreover, for computing a global minimizer of

\[
\text{minimize } J_{(\| \cdot \|_1)_B} \circ \mathcal{L}(x) := \frac{1}{2} \| y - Ax \|_2^2 + \mu (\| \cdot \|_1)_B \circ \text{Id}(x), \ \mu > 0,
\]

an iterative algorithm was presented by Selesnick [56] (see appendix A) but only for a special case satisfying \( B^*B = (\theta / \mu) A^*A \ (0 \leq \theta \leq 1) \). Despite its great potential of the GMC penalty, so far the applicability of the algorithm in [56] is very limited. For example, it is not applicable directly to most scenarios in example 1(c) and (d).

To maximize the applicability of the excellent ideas of the MC penalty function [71] followed by the GMC penalty function \( (\| \cdot \|_1)_B \) [56], we are interested in the following questions:

\(^2\)We use the notation \( (\| \cdot \|_1)_B \) in place of its original notation \( \Psi_B \) used in [56] for the GMC penalty because the GMC penalty in [56] was introduced as a nonconvex alternative to \( \| \cdot \|_1 \) with \( B \in \mathbb{R}^{q \times n} \). In definition 1 of the present paper, we will use \( \Psi_B \) in much wider sense to denote a nonconvex alternative to a general proximable convex function \( \Psi \) defined on finite dimensional real Hilbert space.

\(^3\)The MC penalty

\[
\beta_{\text{MC}} : \mathbb{R} \to \mathbb{R}_{++} : x \mapsto \begin{cases} x - \frac{1}{\beta} x^2, & \text{if } |x| \leq \beta, \\ \frac{\beta}{2}, & \text{otherwise,} \end{cases}
\]

where \( \beta \in \mathbb{R}_{++} \), was introduced in [71] for achieving a nearly unbiased estimate by minimizing \( J_{\text{MC}} : \mathbb{R}^n \to \mathbb{R} : x = (x_1, \ldots, x_n)^T \mapsto \frac{1}{2} \| y - Ax \|_2^2 + \mu \sum_{i=1}^n \beta |x_i|_{\text{MC}} \). In fact, by setting \( B^*B = \beta \text{Id} \), the GMC penalty function \( (\| \cdot \|_1)_B \) reproduces the MC penalty function as \( (\| \cdot \|_1)_B(x) = \sum_{i=1}^n \beta |x_i|_{\text{MC}} \) [56, proposition 12].
(Q1) Can we extend the model (8) proposed in [56], without losing its inherent computational benefit, to

\[
\text{minimize } J_{\Psi \circ \Sigma}(x) := \frac{1}{2}\|y - Ax\|_Z^2 + \mu \Psi_B \circ \Sigma(x), \quad \mu > 0, \quad (9)
\]

where \(\mathcal{X}, \mathcal{Y}, \mathcal{Z}\) and \(\mathcal{L}\) are finite dimensional real Hilbert spaces, \(y \in \mathcal{Y}, A \in \mathcal{B}(\mathcal{X}, \mathcal{Y}), \Sigma \in \mathcal{B}(\mathcal{X}, \mathcal{Z})\) and

\[
\Psi_B(\cdot) := \min_{v \in \mathcal{Z}} \left[ \Psi(v) + \frac{1}{2} \|B(\cdot - v)\|_Z^2 \right]
\]

(10)

with \(\Psi \in \Gamma_0(\mathcal{Z})\) and \(B \in \mathcal{B}(\mathcal{Z}, \mathcal{L})\)?

(Q2) For given \(A \in \mathcal{B}(\mathcal{X}, \mathcal{Y})\) and \(\Sigma \in \mathcal{B}(\mathcal{X}, \mathcal{Z})\), what is the general condition for \(B \in \mathcal{B}(\mathcal{Z}, \mathcal{L})\) and \(\mu > 0\) to ensure the overall convexity of \(J_{\Psi \circ \Sigma}\) in (9)?

(Q3) Can we establish any iterative algorithm of guaranteed convergence to globally optimal solution of (9) under general overall-convexity condition?

(Q4) For given \(A \in \mathcal{B}(\mathcal{X}, \mathcal{Y})\) and \(\Sigma \in \mathcal{B}(\mathcal{X}, \mathcal{Z})\), can we choose \(B \in \mathcal{B}(\mathcal{Z}, \mathcal{L})\) and \(\mu > 0\) flexibly to ensure the overall-convexity \(J_{\Psi \circ \Sigma}\) in (9)?

**Remark 1.**

(a) (On Q1) The function \(\Psi_B\) in (10) is defined in a way similar to the GMC penalty function \((\|\cdot\|_1)_{B}\) in (7) and can be seen as a nonconvex enhanced penalty for a given much more general convex penalty \(\Psi \in \Gamma_0(\mathcal{Z})\) than \((\|\cdot\|_1)_{\text{Id}} \in \Gamma_0(\mathbb{R}^n)\).

(b) (On Q2) In [56] specially for \((\mathcal{X}, \mathcal{Z}, \Psi, \Sigma) = (\mathbb{R}^n, \mathbb{R}^n, \|\cdot\|_1, \text{Id})\), a sufficient condition is found for \(B\) and \(\mu\) to ensure the convexity of \(J_{(\|\cdot\|_1)_{\text{Id}}}\). We will see in remark 3 that this sufficient condition is indeed a necessary and sufficient condition to ensure the convexity of \(J_{(\|\cdot\|_1)_{\text{Id}}}\).

(c) (On Q3) Any iterative algorithm applicable, under fully general overall-convexity conditions, does not seem to have been reported yet even for \((\mathcal{X}, \mathcal{Z}, \Psi, \Sigma) = (\mathbb{R}^n, \mathbb{R}^n, \|\cdot\|_1, \text{Id})\).

As imaginable by the significant effort in the art of proximal splitting [3, 14, 17, 19–21, 65, 67, 69] for minimizing sum of nonsmooth convex functions, it is not trivial to establish algorithm for (9) due to the nonconvexity of \(\Psi_B\) for general \((\mathcal{X}, \mathcal{Z}, \Psi, \Sigma)\) even under the overall convexity condition.

(d) (On Q4) For practical applications, it is important to establish a flexible way to design \(B\) and \(\mu\) under the convexity of \(J_{\Psi \circ \Sigma}\).

The GMC penalties in the form of (9) with \((\mathcal{X}, \mathcal{Z}, \Psi)\) have already been reported (see, e.g. [25, 72]). However these reports do not present any mathematical analysis related to the above key questions (Q1)–(Q4).

This paper considers the questions (Q1)–(Q4) and presents a proximal splitting algorithm for problem (9) with (10) under as much general overall-convexity condition for \((A, B, \Sigma, \mu)\) as possible. After the preliminary section including short reviews on (i) the elements of convex analysis and optimization and (ii) fixed point theory of nonexpansive operators, we will present in proposition 1 useful conditions for the overall convexity of \(J_{\Psi \circ \Sigma}\) in (9). Under the overall convexity condition, we next propose a proximal splitting algorithm (algorithm 1) for problem (9). The proposed algorithm has theoretical guarantee of convergence to a global minimizer of (9) (see theorem 1 in section 3.1) and is designed in a way similar to an idea behind the primal-dual splitting method [21, 49, 65] which was established specially for minimization of sum of linearly involved convex terms. Furthermore, we also present a
Table 1. Singular values $\sigma_1 \geq \cdots \geq \sigma_{16} \geq 0$ of the original and estimated matrices in figure 10 and the numerical rank (num-rank) which is defined as the number of singular values greater than $10^{-4}$.

| Singular values | $\sigma_1$ | $\sigma_2$ | $\sigma_3$ | $\sigma_4$ | $\sigma_{16}$ | Num-rank |
|-----------------|-----------|-----------|-----------|-----------|-----------|---------|
| Original        | $6.48 \times 10^3$ | $9.01 \times 10^{-1}$ | $3.85 \times 10^{-1}$ | $0$ | $\ldots$ | $0$ | $3$ |
| Nuclear norm    | $6.42 \times 10^3$ | $8.55 \times 10^{-1}$ | $3.38 \times 10^{-1}$ | $6.66 \times 10^{-2}$ | $\ldots$ | $6.52 \times 10^{-11}$ | $8$ |
| LiGME           | $6.48 \times 10^3$ | $9.11 \times 10^{-1}$ | $3.89 \times 10^{-1}$ | $1.10 \times 10^{-14}$ | $\ldots$ | $8.26 \times 10^{-17}$ | $3$ |

flexible way to design $B$ and $\mu$ in proposition 2 for the convexity of $J_{\psi, \mu}$ and $\sigma$. To demonstrate the effectiveness of the proposed algorithm, we present numerical experiments in four different sparsity-rank-aware signal processing scenarios.

Preliminary short versions of this paper were presented at conferences [1, 66].

2. Preliminaries

Let $\mathbb{N}$, $\mathbb{R}$, $\mathbb{R}^+$, and $\mathbb{R}^{++}$ be the sets of natural numbers, real numbers, nonnegative real numbers, and positive real numbers, respectively. The superscript $\left(\cdot\right)^T$ denotes transpose. For a vector $x := (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$, we use $\|x\|_p := \left(\sum_{i=1}^n |x_i|^p\right)^{1/p} (0 < p < \infty)$, $\|x\|_{\infty} := \max\{|x_1|, \ldots, |x_n|\}$, and $\|x\|_0 := \#\{i \in \mathbb{N} \cap [1,n] \mid x_i \neq 0\}$. $O_n \in \mathbb{R}^{n \times n}$ stands for the zero vector. In sections 2.1 and 2.2, we use finite dimensional real Hilbert spaces $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}}, \| \cdot \|_{\mathcal{H}})$ and $(\mathcal{K}, \langle \cdot, \cdot \rangle_{\mathcal{K}}, \| \cdot \|_{\mathcal{K}})$. For $S \subseteq \mathcal{H}$, cone$(S)$ denotes the conical hull (see, e.g. [3, definition 6.1]) of $S$ and span$(S)$ the span of $S$. $B(\mathcal{H}, \mathcal{K})$ denotes the set of all bounded linear operators from $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}}, \| \cdot \|_{\mathcal{H}})$ to $(\mathcal{K}, \langle \cdot, \cdot \rangle_{\mathcal{K}}, \| \cdot \|_{\mathcal{K}})$. For $L \in B(\mathcal{H}, \mathcal{K})$, we use $\|L\|_{op} := \sup_{x \in \mathcal{H}, \|x\|_{\mathcal{H}} < 1} \|Lx\|_{\mathcal{K}}$. For $L \in B(\mathcal{H}, \mathcal{K})$, $L^* \in B(\mathcal{K}, \mathcal{H})$ denotes the adjoint of $L$, i.e. $\langle Lx, y \rangle_{\mathcal{K}} = \langle x, L^*y \rangle_{\mathcal{H}} \quad (\forall x, y \in \mathcal{H} \times \mathcal{K})$. We also use $I_d$ to denote the identity operator for general Hilbert spaces. $O_B(\mathcal{H}, \mathcal{K}) \subseteq B(\mathcal{H}, \mathcal{K})$ and $O_{\mathcal{H}} \subseteq B(\mathcal{H}, \mathcal{H})$ stand for the zero operators. For $L \in B(\mathcal{H}, \mathcal{K})$, $L^1 \in B(\mathcal{K}, \mathcal{H})$ stands for the Moore–Penrose pseudo inverse of $L$, $\text{ran}(L) := \{Lx \in \mathcal{K} \mid x \in \mathcal{H}\}$ and $\text{null}(L) := \{x \in \mathcal{H} \mid Lx = 0\}$ denote respectively the range and the null spaces of $L$. The positive definiteness and positive semidefiniteness of a self-adjoint operator $L \in B(\mathcal{H}, \mathcal{H})$ are expressed respectively as $L \succ O_{\mathcal{H}}$ and $L \succeq O_{\mathcal{H}}$. For $L \succeq O_{\mathcal{H}}$, $\mu(L)$ denotes the maximum eigenvalue of $L$. For any $L \succ O_{\mathcal{H}}$, by defining an inner product $\langle \cdot, \cdot \rangle_L: \mathcal{H} \times \mathcal{H} \rightarrow \mathbb{R}$, $(x, y) \mapsto \langle x, Ly \rangle_{\mathcal{H}}$ and its induced norm $\|x\|_L := \sqrt{\langle x, x \rangle_L}$, $(\mathcal{H}, \langle \cdot, \cdot \rangle_L, \| \cdot \|_L)$ becomes a real Hilbert space.

Note that, in any real finite dimensional space, a linear operator can be expressed with matrix multiplication and identified with a matrix. We use $I_d \in \mathbb{R}^{n \times n}$ to denote the identity matrix for $\mathbb{R}^n$. $O_{m,n} \in \mathbb{R}^{m \times n}$ and $O_n \in \mathbb{R}^{n \times n}$ stand for the zero matrices.

2.1. Selected elements of convex analysis and optimization

The class of proper lower semicontinuous convex functions $f: \mathcal{H} \rightarrow (-\infty, \infty]$, i.e. $f$ is convex function whose lower level set $\{x \in \mathcal{H} \mid f(x) \leq \alpha\}$ is closed for every $\alpha \in \mathbb{R}$ and $\text{dom}(f) := \{x \in \mathcal{H} \mid f(x) < \infty\} \neq \emptyset$, is denoted by $\Gamma_0(\mathcal{H})$. For convex $C \subseteq \mathcal{H}$, the relative interior of $C$ is $\text{ri}C := \{x \in \mathcal{H} \mid \text{cone}(C - x) = \text{span}(C - x)\}$ (see, e.g. [3, definition 6.9]).

*In real finite dimensional Hilbert space, $B(\mathcal{H}, \mathcal{K})$ is identical to the set of all linear operators.*
(Subdifferential) For a function \( f \in \Gamma_0(H) \), the \textit{subdifferential} of \( f \) is defined as the set valued operator
\[
\partial f: H \to 2^H: x \mapsto \{ u \in H \mid \langle y - x, u \rangle_H + f(x) \leq f(y), \ \forall y \in H \}.
\]

Subdifferential has the following properties:

(a) (Fermat’s rule [3, theorem 16.3]) Let \( f \in \Gamma_0(H) \) and \( \bar{x} \in H \). Then
\[
\bar{x} \in \arg \min_{x \in H} f(x) \iff 0 \in \partial f(\bar{x}). \tag{11}
\]

(b) (Sum rule [3, corollary 16.48]) Let \( f, g \in \Gamma_0(H) \) with \( \text{dom}(g) = H \). Then
\[
\partial(f + g) = \partial f + \partial g. \tag{12}
\]

(c) (Chain rule [3, corollary 16.53, fact 6.14i, section 6.2]) Let \( g \in \Gamma_0(H) \) and \( L \in B(H, K) \) satisfy \( 0_H \in \text{ri} (\text{dom}(g) - \text{ran} L) \). Then
\[
\partial(g \circ L) = L^* \circ (\partial g) \circ L. \tag{13}
\]

(d) [3, proposition 17.31] Let \( f \in \Gamma_0(H) \), let \( x \in \text{dom}(f) \), and suppose that \( f \) is (Gâteaux) differentiable at \( x \). Then \( \partial f(x) = \{ \nabla f(x) \} \).

(Legendre–Fenchel conjugate) For any \( f \in \Gamma_0(H) \), the function defined by
\[
\bar{f}^*: H \to (-\infty, \infty]: y \mapsto \sup_{x \in H} \{ \langle x, y \rangle_H - f(x) \}
\]
satisfies \( \bar{f}^* \in \Gamma_0(H) \). This function is called the \textit{conjugate} (also named \textit{Legendre–Fenchel conjugate}) of \( f \). Let \( f \in \Gamma_0(H) \). Then, for any \((x, u) \in H \times H\),
\[
u \in \partial f^*(u) \iff x \in \partial f^*(u). \tag{14}
\]

2.2. Selected elements of fixed point theory of nonexpansive operators

(Nonexpansive operator) An operator \( T: H \to H \) is said to be \( \kappa \)-Lipschitzian with constant \( \kappa > 0 \) if
\[
(\forall x, y \in H) \quad \|T(x) - T(y)\|_H \leq \kappa \|x - y\|_H.
\]

In particular, an operator \( T: H \to H \) is said to be \textit{nonexpansive} if it is 1-Lipschitzian, i.e.
\[
(\forall x, y \in H) \quad \|T(x) - T(y)\|_H \leq \|x - y\|_H.
\]

For \( \alpha \in (0, 1) \), a nonexpansive operator \( T \) is called \( \alpha \)-\textit{averaged} if there exists a nonexpansive operator \( \tilde{T}: H \to H \) such that
\[
T = (1 - \alpha)\text{Id} + \alpha \tilde{T},
\]
i.e. \( T \) is a convex combination of the identity operator \( \text{Id} \) and some nonexpansive operator \( \tilde{T} \).

Fact 1 (Compositions of averaged nonexpansive operators [48][20, proposition 2.4]). Suppose that each \( T_i: H \to H \) \((i = 1, 2)\) is \( \alpha_i \)-averaged nonexpansive for some \( \alpha_i \in (0, 1) \). Then \( T_1 \circ T_2 \) is \( \alpha \)-averaged nonexpansive for \( \alpha := \frac{\alpha_1 + \alpha_2 - 2 \alpha_1 \alpha_2}{1 - \alpha_1 \alpha_2} \in (0, 1) \).
Fact 2 (Krasnosel’skii–Mann iteration for finding a fixed point of averaged nonexpansive operator [3, section 5.2] [31]). For a nonexpansive operator $T : \mathcal{H} \to \mathcal{H}$ with $\text{Fix}(T) := \{ x \in \mathcal{H} \mid T(x) = x \} \neq \emptyset$ and any initial point $x_0 \in \mathcal{H}$, the sequence $(x_k)_{k \in \mathbb{N}} \subset \mathcal{H}$ generated by

$$x_{k+1} = [(1 - \alpha_k)\text{Id} + \alpha_kT](x_k)$$

(15)

converges weakly to a point in $\text{Fix}(T)$ if $(\alpha_k)_{k \in \mathbb{N}} \subset [0, 1]$ satisfies $\sum_{k \in \mathbb{N}} \alpha_k(1 - \alpha_k) = \infty$. In particular, if $T$ is $\alpha$-averaged for some $\alpha \in (0, 1)$, a simple iteration

$$x_{k+1} = T(x_k)$$

converges weakly to a point in $\text{Fix}(T)$.

(Monotone operator) A set-valued operator $T : \mathcal{H} \to 2^\mathcal{H}$ is said to be monotone if

$$(\forall (x,u) \in \text{gra}(T))(\forall (x',u') \in \text{gra}(T)) \quad (x - x', u - u')_\mathcal{H} \geq 0,$$

where $\text{gra}(T) := \{(x,u) \in \mathcal{H} \times \mathcal{H} \mid u \in T(x)\}$ is the graph of $T$. In particular, $T$ is called maximally monotone if, for every $(x,u) \in \mathcal{H} \times \mathcal{H},$

$$(x,u) \in \text{gra}(T) \iff (\forall (x',u') \in \text{gra}(T)), \quad (x - x', u - u')_\mathcal{H} \geq 0.$$

For a given $f \in \Gamma_0(\mathcal{H})$, $\partial f : \mathcal{H} \to 2^\mathcal{H}$ is maximally monotone if and only if the resolvent $R_T := (\text{Id} + T)^{-1} : \mathcal{H} \to 2^\mathcal{H} : u \mapsto \{ x \in \mathcal{H} \mid u \in x + T(x) \}$ is single-valued $(1/2)$-averaged nonexpansive operator.

(Proximity operator) The proximity operator of $f \in \Gamma_0(\mathcal{H})$ is defined by

$$\text{Prox}_f : \mathcal{H} \to \mathcal{H} : x \mapsto \arg \min_{y \in \mathcal{H}} \left[ f(y) + \frac{1}{2}\| x - y \|_\mathcal{H}^2 \right].$$

Note that $\text{Prox}_f(x) \in \mathcal{H}$ is well-defined for all $x \in \mathcal{H}$ due to the coercivity and the strict convexity of $f(\cdot) + \frac{1}{2}\| \cdot \|_\mathcal{H}^2 \in \Gamma_0(\mathcal{H})$. It is also well known that $\text{Prox}_f$ is nothing but the resolvent of $\partial f$, i.e. $\text{Prox}_f = (\text{Id} + \partial f)^{-1} = R_{\partial f}$, which implies that

$$\bar{x} \in \text{Fix}(\text{Prox}_f) \iff \text{Prox}_f(\bar{x}) = \bar{x} \iff (\text{Id} + \partial f)^{-1}(\bar{x}) = \bar{x}$$

(16)

$$\iff \bar{x} \in (\text{Id} + \partial f)(\bar{x}) \iff 0 \in \partial f(\bar{x}) \iff \bar{x} \in \arg \min_{x \in \mathcal{H}} f(x).$$

(17)

The proximity operator of $\Psi^*$ can be expressed as $\text{Prox}_{\Psi^*} = \text{Id} - \text{Prox}_\Psi$ (see e.g. [3, theorem 14.3(ii)]).

(Moreau envelope) For $f \in \Gamma_0(\mathcal{H}),$

$$\gamma f : \mathcal{H} \to \mathbb{R} : x \mapsto \min_{y \in \mathcal{H}} \left[ f(y) + \frac{1}{2\gamma}\| x - y \|_\mathcal{H}^2 \right],$$

(18)

is called the Moreau envelope of $f$ of index $\gamma > 0$. The Moreau envelope of $f \in \Gamma_0(\mathcal{H})$ converges pointwise to $f$ on $\text{dom}(f)$ as $\gamma \downarrow 0$, i.e. $\lim_{\gamma \downarrow 0} \gamma f(x) = f(x)$ for every $x \in \text{dom}(f)$. The function $\gamma f$ is Fréchet differentiable convex function with $(1/\gamma)$-Lipschitzian gradient.
\[
\n\nabla^* f : H \to H : x \mapsto \frac{x - \text{Prox}_f(x)}{\gamma}.
\]

3. Linearity involved generalized-Moreau-enhanced (LiGME) model and proximal splitting algorithm

In this section, after introducing LiGME model (see definition 1), we then presents a proximal splitting type algorithm of guaranteed convergence to a globally optimal solution of the model under an overall convexity condition (see theorem 1).

3.1. Linearly involved LiGME model

We impose the relatively strong assumption \(\text{dom} \, \Psi = Z\) for \(\Psi\) in (10), to reduce technical complexity in the later discussion, although there would be many ways to relax.

Definition 1 (Linearly involved generalized-Moreau-enhanced (LiGME) model). Let \((X, \langle \cdot, \cdot \rangle_X, \| \cdot \|_X), (Y, \langle \cdot, \cdot \rangle_Y, \| \cdot \|_Y), (Z, \langle \cdot, \cdot \rangle_Z, \| \cdot \|_Z), \) and \((\tilde{Z}, \langle \cdot, \cdot \rangle_{\tilde{Z}}, \| \cdot \|_{\tilde{Z}})\) be finite dimensional real Hilbert spaces, \(\Psi \in \Gamma_0(Z)\) coercive with \(\text{dom} \, \Psi = Z\), \(B \in B(Z, \tilde{Z})\), \(L \in B(X, Z)\), and \((A, L, \mu) \in B(X, Y) \times B(X, Z) \times \mathbb{R}^+\). Then:

(a) GME penalty function \(\Psi_B \in \Gamma_0(Z)\) is defined as
\[
\Psi_B(\cdot) := \Psi(\cdot) - \min_{v \in Z} \left[ \Psi(v) + \frac{1}{2} \|B(\cdot - v)\|_{\tilde{Z}}^2 \right].
\]

(b) Linearly involved generalized-Moreau-enhanced (LiGME) penalty is defined as \(\mu \Psi_B \circ \mathcal{L} \in \mathcal{L} \to (-\infty, \infty]\).

(c) LiGME model is defined as the minimization of
\[
J_{\Psi_B \circ \mathcal{L}} : \mathcal{L} \to \mathbb{R} : x \mapsto \frac{1}{2} \|y - Ax\|_Y^2 + \mu \Psi_B \circ \mathcal{L}(x).
\]

Example 2 (LiGME penalty bridges the gap between the direct discrete measures and their convex envelopes).

(a) (Normalized MC penalty) By letting \(X = Z = \mathbb{R}, \Psi = \| \cdot \|, \mathcal{L} = 1, B = \frac{1}{\sqrt{\gamma}}\) for \(\gamma \in \mathbb{R}^+\) and \(\mu = \frac{2}{\gamma}\), the function \(\mu \Psi_B \circ \mathcal{L}\) in (21) reproduces
\[
\frac{2}{\gamma} (\| \cdot \|_{MC}) : \mathbb{R} \to \mathbb{R} : x \mapsto \begin{cases} \frac{2}{\gamma} |x| - \frac{1}{\gamma} x^2, & \text{if } |x| \leq \gamma; \\ 1, & \text{otherwise}, \end{cases}
\]

which satisfies
\[
\lim_{\gamma \downarrow 0} \frac{2}{\gamma} (\| \cdot \|_{MC}) = \begin{cases} 0, & \text{if } x = 0; \\ 1, & \text{otherwise}. \end{cases}
\]

(b) (LiGME penalty bridges the gap between \(\| \cdot \|_0 \) and \(\| \cdot \|_1\)) Let \(X = Z = \mathbb{R}^n, \Psi = \| \cdot \|_l, \mathcal{L} = \text{Id}, B = \frac{1}{\sqrt{\gamma}} \text{Id}\) for \(\gamma \in \mathbb{R}^+\) and \(\mu = \frac{2}{\gamma}\). Then the function \(\mu \Psi_B \circ \mathcal{L}\) in (21) reproduces
\[
\frac{2}{\gamma} (\| \cdot \|_1) \frac{1}{\gamma} \text{Id} : \mathbb{R}^n \to \mathbb{R} : (x_1, \ldots, x_n) \mapsto \sum_{i=1}^{n} \frac{2}{\gamma} (\gamma |x_i|_{\text{nuc}}) \tag{24}
\]

which satisfies for \((x_1, \ldots, x_n) \in \mathbb{R}^n\)

\[
\lim_{\gamma \downarrow 0} \frac{2}{\gamma} (\| \cdot \|_1) \frac{1}{\gamma} \text{Id}(x_1, \ldots, x_n) = \| (x_1, \ldots, x_n) \|_0.
\tag{25}
\]

This fact together with \((\| \cdot \|_1)_{\text{O}_n}(x_1, \ldots, x_n) = \| (x_1, \ldots, x_n) \|_1\) validates that the LiGME penalty can serve as a parametrized bridge between \(\| \cdot \|_0\) and \(\| \cdot \|_1\).

(c) (LiGME penalty bridges the gap between rank(\(\cdot\)) and \(\| \cdot \|_{\text{nuc}}\)) Let \(\mathcal{X} = \mathbb{Z} = \mathbb{R}^{m \times n}, \quad \Psi = \| \cdot \|_{\text{nuc}}, \quad \Sigma = \text{Id}, \quad B = \frac{1}{\sqrt{\gamma}} \text{Id}\) for \(\gamma \in \mathbb{R}_{++}\) and \(\mu = \frac{\gamma}{\sqrt{\gamma}}\), where \(\| \cdot \|_{\text{nuc}} : \mathbb{R}^{m \times n} \to \mathbb{R}\) : \(X \mapsto \sum_{i=1}^{r} \sigma_i(X)\) with \(r = \text{rank}(X)\) and \(i\)th largest singular value \(\sigma_i(X)\) \((i = 1, 2, \ldots, r)\) of \(X\). It is well-known that \(\| \cdot \|_{\text{nuc}}\) is a convex envelope of rank(\(\cdot\)), i.e. the largest convex minorant of rank(\(\cdot\)), in a vicinity of \(\text{O}_{m,n}\). By [3, proposition 24.68], the function \(\mu \Psi_B \circ \Sigma\) in (21) reproduces

\[
\frac{2}{\gamma} (\| \cdot \|_{\text{nuc}}) \frac{1}{\gamma} \text{Id} : \mathbb{R}^{m \times n} \to \mathbb{R} : X \mapsto \sum_{i=1}^{r} \frac{2}{\gamma} (\gamma |\sigma_i(X)|_{\text{nuc}}) \tag{26}
\]

which satisfies for \(X \in \mathbb{R}^{m \times n}\)

\[
\lim_{\gamma \downarrow 0} \frac{2}{\gamma} (\| \cdot \|_{\text{nuc}}) \frac{1}{\gamma} \text{Id}(X) = \| (\sigma_1(X), \ldots, \sigma_r(X)) \|_0 = \text{rank}(X).
\tag{27}
\]

This fact together with \((\| \cdot \|_{\text{nuc}})_{\text{O}_n}(X) = \|X\|_{\text{nuc}}\) validates that the LiGME penalty can serve as a parametrized bridge between rank(\(\cdot\)) and \(\| \cdot \|_{\text{nuc}}\).

Example 3 (The sum of multiple LiGME penalties can be expressed as a single LiGME penalty on product space). Let \(Z_1, \tilde{Z}_i, (i = 1, 2, \ldots, \mathcal{M}), \quad \mathbb{Z} = Z_1 \times Z_2 \times \cdots \times Z_{\mathcal{M}}\) and \(\tilde{Z} = \tilde{Z}_1 \times \tilde{Z}_2 \times \cdots \times \tilde{Z}_{\mathcal{M}}\) be real Hilbert spaces. For coercive \(\Psi^{(i)} \in \Gamma_0(Z_i)\) with \(\text{dom} \Psi^{(i)} = Z_i, \quad B^{(i)} \in B(Z_i, \tilde{Z}_i)\) and \(\Sigma_i \in B(\mathcal{X}, Z_i)\) \((i = 1, 2, \ldots, \mathcal{M})\), let \(\Psi := \mu_1 \Psi^{(1)} \odot \mu_2 \Psi^{(2)} \odot \cdots \odot \mu_{\mathcal{M}} \Psi^{(\mathcal{M})}, \quad B : \mathcal{X} \to \tilde{Z} : (z_1, \ldots, z_\mathcal{M}) \mapsto (\sqrt{\mu_1} B^{(1)} z_1, \ldots, \sqrt{\mu_\mathcal{M}} B^{(\mathcal{M})} z_\mathcal{M}), \quad \Sigma : \mathcal{X} \to \mathcal{Z} : x \mapsto (\Sigma_i x)_{i=1}^{\mathcal{M}}\). Then we have

\[
\Psi_B \circ \Sigma = \sum_{i=1}^{\mathcal{M}} \mu_i (\Psi^{(i)})(B^{(i)} \circ \Sigma_i), \tag{28}
\]

where \((\Psi^{(i)})(B^{(i)} \circ \Sigma_i) := \Psi^{(i)}(\cdot) - \min_{v \in Z_i} \left[ \Psi^{(i)}(v) + \frac{1}{2} \| B^{(i)}(\cdot - v) \|^2_{Z_i} \right].\]

Remark 2. The LS-CNC penalty function in [70, definition 2] is reproduced as an LiGME penalty by setting \(\mathcal{X} = \mathcal{Z} = \mathbb{R}^{m \times n} \oplus \mathbb{R}^{m \times n}, \quad \Sigma = \text{Id}, \quad \Psi = \Psi_1 \oplus \Psi_2\) with \(\Psi_1 := \alpha \| \cdot \|_{\text{nuc}}\) and \(\Psi_2 := \beta \| \cdot \|_1\) in (5), where \(\alpha, \beta \geq 0\). Moreover, the LiGME penalty in example 3 can also be utilized to enhance the so-called morphological component analysis in [59].

Proposition 1 (Overall convexity condition for the LiGME model). The GME penalty function \(\Psi_B\) in definition 1 has the following properties:

\[(a) \quad \Psi_B \circ \Sigma(x) = \Psi(\Sigma_x) - \left[ \Psi(0_x) + \frac{1}{2} \| B \Sigma_x \|^2_{\Sigma_x} \right] \quad \text{if and only if} \quad B^* B \Sigma x \in \text{argmin}(\Psi^*). \]
(b) Let \( (A, \Sigma, \mu) \in \mathcal{B}(X, Y) \times \mathcal{B}(X, Z) \times \mathbb{R}^+ \). Then, for the three conditions
\[(C_1) A^*A - \mu L^*B^*L \succeq 0_X, \quad (C_2) J_{\Psi \circ \Sigma} \in \Gamma(0(X)) \text{ for any } y \in Y, \text{ and } (C_3) J_{\Psi \circ \Sigma}^{(0)} \triangleq \frac{1}{2} \|A \cdot \|_Y^2 + \mu \Psi_B \circ \Sigma \in \Gamma(0(X)) \text{, the relation } (C_1) \Rightarrow (C_2) \Rightarrow (C_3) \text{ holds.}
\]

In particular, if \( \Psi \) is a certain norm, say \( \| \cdot \| \), over the vector space \( Z \), these properties are enhanced as:
\[\text{(a') } \|\| \cdot \|\|_A \circ \Sigma(x) = \|\|\Sigma x\|\| - \frac{1}{2} \|B \Sigma x\|_Z^2 \text{ if and only if } \|\|B^*B \Sigma x\|\| \leq 1, \text{ where } \|\| \cdot \|\|_A \circ \Sigma \to \mathbb{R} : v \mapsto \sup_{w \in Z : \|\|w\|\|_Z \leq 1} \|\langle w, v \rangle \| \text{ is the dual norm}^5 \text{ of } \|\| \cdot \|\| \).
\[\text{(b')} \text{ The equivalence } (C_1) \Leftrightarrow (C_2) \Leftrightarrow (C_3) \text{ holds.}
\]

**Proof.** See appendix B. \(\square\)

**Remark 3.**

(i) Proposition 1(a’) for special case \( (X, Z, \|\| \cdot \|\|_A, \Sigma) = (\mathbb{R}^n, \mathbb{R}^n, \|\|_1, \text{Id}) \) reproduces \([56, \text{ corollary 2}] \) (i.e. \( \|\|x\|\|_A = \|\|x\|\|_1 - \|\|Bx\|\|_Z^2 \) if and only if \( \|\|B^*Bx\|\|_\infty \leq 1 \) because the dual norm of \( \|\| \cdot \|\|_A \) is \( \|\| \cdot \|\|_\infty \). For the whole shape of the graph of \( (\|\| \cdot \|\|_A) \), see the graphs in \([56, \text{ figures 3, 8, and 9}] \) of the GMC penalty.

(ii) Proposition 1(b’) specialized for \( \mu > 0 \) and \( A = O_B(X, Y) \) yields
\[B = O_B(Z, Z) \Rightarrow \langle B^*B \succeq O_Z \rangle \Leftrightarrow \|\| \cdot \|\|_A = \|\| \cdot \|\| \in \Gamma(0(X)). \quad (29)\]

(iii) \( (C_1) \Rightarrow (C_2) \) is found in \([56, \text{ theorem 1}] \) but only for special case \( \Psi = \|\| \cdot \|\|_1 \) (compare this with propositions 1(b) and (b')).

### 3.2. A proximal splitting algorithm for the LiGME model and its global convergence property

Our target is the following convex optimization problem:

**Problem 1** (LiGME model in definition 1 under an overall convexity condition). Assume that \( \Psi \in \Gamma(0(Z)) \) satisfies the even symmetry\(^6\) \( \Psi \circ (-\text{Id}) = \Psi \) and is proximable, i.e. \( \text{prox}_\Psi \) is available as a computable operator for every \( \gamma \in \mathbb{R}^+ \). Then for \( (A, \Sigma, B, y, \mu) \in \mathcal{B}(X, Y) \times \mathcal{B}(X, Z) \times \mathcal{B}(Z, Z) \times \mathcal{Y} \times \mathbb{R}^+ \) satisfying \( A^*A - \mu L^*B^*L \succeq 0_X \),

\[
\text{find } x^* \in S := \arg \min_{x \in X} J_{\Psi \circ \Sigma}(x). \quad (30)
\]

We will use a technical lemma below.

**Lemma 1.** In definition 1, if \( \Psi \) satisfies \( \Psi \circ (-\text{Id}) = \Psi \), we have
\[
0 \in \text{ri} \left( \text{dom} \left( \left( \Psi + \frac{1}{2} \|B \cdot \|_Z^2 \right)^* \right) - \text{ran}(B^*) \right). \quad (31)
\]

**Proof.** See appendix C. \(\square\)

In the next theorem, (a) and (b) show that the set \( S \) of all globally optimal solutions of problem 1 can be expressed in terms of the fixed-point set of a computable averaged nonexpansive

---

5 See, e.g. \([37, \text{ definition 5.4.12}], [38, \text{ definition 2.10.3}], \) and \([10, \text{ example 3.26}] \).

6 In this case, for \( B = O_Z \), we have \( \Psi_B(\cdot) = \Psi(\cdot) - \Psi(0_Z) \) (See also (29)).
operator in a certain real Hilbert space, and (c) presents an iterative algorithm, for problem 1, based on the Kransnosel’skii–Mann iteration in fact 2.

**Theorem 1 (Nonexpansive operator $T_{\text{LiGME}}$ and iterative algorithm for problem 1).** In problem 1, let $(H := X \times Z \times Z, \langle \cdot, \cdot \rangle_H, \|\cdot\|_H)$ be a real Hilbert space whose inner product $\langle \cdot, \cdot \rangle_H$ is defined as the one for the product space in example 1(d), and define $T_{\text{LiGME}} : H \to H : (x, v, w) \mapsto (\xi, \zeta, \eta)$, with $(\sigma, \tau) \in \mathbb{R}_+ \times \mathbb{R}_+$, by

\[
\xi := \text{Id} - \frac{1}{\sigma}(A^*A - \mu L^*B^*B) x - \frac{\mu}{\sigma} L^*B^*B v - \frac{\mu}{\sigma} L^*w + \frac{1}{\sigma} A^*y, \\
\zeta := \text{Prox}_{\psi} \left[ \frac{2\mu}{\tau} B^*B \xi - \frac{\mu}{\tau} B^*B \xi + \left( \text{Id} - \frac{\mu}{\tau} B^*B \right) v \right], \\
\eta := \text{Prox}_{\psi^*} \left( 2L^* - \xi + w \right).
\]

Then

(a) the solution set $\mathcal{S}$ of problem 1 can be expressed as

\[
\mathcal{S} = \Xi(\text{Fix}(T_{\text{LiGME}})) := \{ \Xi(x^*, v^*, w^*) \in X \mid (x^*, v^*, w^*) \in \text{Fix}(T_{\text{LiGME}}) \}
\]

with $\Xi : H \to X : (x, v, w) \mapsto x$.

(b) Choose $(\sigma, \tau, \kappa) \in \mathbb{R}_+ \times \mathbb{R}_+ \times (1, \infty)$ satisfying\(^7\)

\[
\sigma \text{Id} - \frac{\kappa}{2} A^*A - \mu L^*L > O_X, \\
\tau \geq \left( \frac{\kappa}{2} + \frac{\kappa}{2} \right) \mu \|B\|_{op}^2.
\]

Then

\[
\begin{bmatrix}
\sigma \text{Id} & -\mu L^*B^*B & -\mu L^*x \\
-\mu B^*L & \tau \text{Id} & O_Z \\
-\mu L & O_Z & \mu \text{Id}
\end{bmatrix} > O_H
\]

and $T_{\text{LiGME}}$ is $\frac{\kappa}{2\sigma}$-averaged nonexpansive in the Hilbert space $(H, \langle \cdot, \cdot \rangle_H, \|\cdot\|_H)$.

(c) Assume $(\sigma, \tau, \kappa) \in \mathbb{R}_+ \times \mathbb{R}_+ \times (1, \infty)$ satisfies (33). Then, for any initial point $(x_0, v_0, w_0) \in H$, the sequence $(x_k, v_k, w_k) \subseteq H$ generated by

\[
(x_{k+1}, v_{k+1}, w_{k+1}) = T_{\text{LiGME}}(x_k, v_k, w_k)
\]

converges weakly to a point $(x^*, v^*, w^*) \in \text{Fix}(T_{\text{LiGME}})$ and

\[
\lim_{k \to \infty} x_k = x^* \in \mathcal{S}.
\]

**Proof.** See appendix D.

Detailed description of the algorithm proposed in theorem 1 is shown in algorithm 1.

**Remark 4 (Algorithm 1 versus existing algorithms).**

(a) The derivation of algorithm 1 is inspired by Condat’s primal-dual algorithm [21] and is essentially based on the so-called forward-backward splitting method (see also (D.13)

\(^7\)For example, (33) is satisfied by any $\kappa > 1$ and

\[
\sigma := \| A^*A + \mu L^*L \|_{op} + (\kappa - 1), \\
\tau := \left( \frac{\kappa}{2} + \frac{\kappa}{2} \right) \mu \|B\|_{op}^2 + (\kappa - 1).
\]
demonstrating that $T_{\text{LiGME}}$ is a forward-backward operator). Since Condat’s primal-dual algorithm was proposed for minimization of sum of linearly involved convex terms, it is not directly applicable to the LiGME model involving nonconvex functions.

(b) The proposed algorithm in (35) differs clearly from Combettes–Pesquet primal-dual algorithm [18] which is for monotone inclusion problems and based on the so-called forward–backward–forward splitting method (or Tseng’s method), i.e. requires an extra forward step compared with the so-called forward-backward splitting method.

(c) Vu’s primal-dual algorithm [65] for monotone inclusion is also based on the so-called forward–backward splitting method. However, to the best of the authors’ knowledge, the strongly monotone assumption (of $D_\tau$) in [65, problem 1.1] prevents from applying directly the Vu’s primal-dual algorithm to problem 1 if $\text{null}(B) \neq \{0_Z\}$. Algorithm 1 is applicable to general $B \in B(Z, \tilde{Z})$.

Algorithm 1. For problem 1.

Choose $(x_0, v_0, w_0) \in H(= X \times Z \times Z)$.
Let $(\sigma, \tau, \kappa) \in \mathbb{R}_+ \times \mathbb{R}_+ \times (1, \infty)$ satisfying (33).
Define $\Psi$ as (34).
Define $k \leftarrow 0$.
Do

$x_{k+1} \leftarrow \left[1 - \frac{1}{\sigma}(A^* A - \mu L^* B^* B L)\right] x_k - \frac{\mu}{\sigma} L^* B v_k - \frac{\sigma}{\kappa} \frac{\mu}{\sigma} L^* w_k + \frac{1}{\kappa} A^* y$

$v_{k+1} \leftarrow \text{Prox}_{\frac{1}{\sigma}} \left[\frac{1}{2} B^* B L v_{k+1} - \frac{\mu}{\sigma} L^* B L x_{k+1} + (1 - \frac{\mu}{\sigma} B^* B) v_{k+1}\right]

w_{k+1} \leftarrow \text{Prox}_{\frac{1}{\kappa}} \left[\frac{2}{\kappa} L^* \frac{\mu}{\sigma} L^* w_{k} - \frac{\mu}{\sigma} L^* x_{k+1} + w_{k}\right]

k \leftarrow k + 1

while $\|x_k, v_k, w_k\|_2$ is not sufficiently small

return $x_k$

3.3. How to choose $B$ to ensure overall-convexity of $J_{\Psi_B \circ L}$

Choices of $B$ to guarantee $J_{\Psi_B \circ L} \in \Gamma_0(\mathbb{R}^n)$ are given, e.g. as follows.

Proposition 2 (A design of $B$ to ensure the overall-convexity condition in proposition 1(b)). Indefinition 1, let $(X, Y, Z) = (\mathbb{R}^n, \mathbb{R}^m, \mathbb{R}^l), (A, \Sigma, \mu) \in \mathbb{R}^{m \times n} \times \mathbb{R}^{l \times n} \times \mathbb{R}_+$, and $\text{rank}(\Sigma) = 1$. Choose a nonsingular $\Sigma \in \mathbb{R}^{l \times n}$ satisfying $\Sigma_{n \times (n-l)} I_l \Sigma = \Sigma$. Then

$B_\theta := \sqrt{\frac{\theta}{\mu}} \mu^{1/2} U^T \in \mathbb{R}^{l \times l}$, \quad $\theta \in [0, 1], \quad (36)$

ensures $J_{\Psi_B \circ L} \in \Gamma_0(\mathbb{R}^n)$, where

$[\hat{A}_1 \hat{A}_2] := A(\Sigma)^{-1}$

and $\Sigma U^T := \hat{A}_2^T \hat{A}_2 - \hat{A}_2^T \hat{A}_1 (\hat{A}_1^T \hat{A}_1)^{1/2} \hat{A}_1^T \hat{A}_2 \in \mathbb{R}^{l \times l}$ is an eigendecomposition.

Proof. See appendix E

The next corollary presents a way of design $B^{(i)} \in B(Z, \tilde{Z}) \ (i = 1, 2, \ldots, M)$ in example 3 for $\Psi_B \circ L$ in (28) to ensure the overall-convexity condition in proposition 1(b).

Such a choice is always possible. See corollary 1 and numerical experiments in four different scenarios in section 4.
Corollary 1 (A design of $B^{(i)}$ in example 3 to ensure the overall-convexity condition in proposition 1(b)). In example 3, let $(X, \mathcal{Y}, \mathcal{Z}_i) = (\mathbb{R}^n, \mathbb{R}^m, \mathbb{R}^l)$, $(A, \mathcal{L}, \mu) \in \mathbb{R}^{m \times n} \times \mathbb{R}^{l \times n} \times \mathbb{R}_{++}$, and rank($\mathcal{L}_i$) = $l_i$ $(i = 1, 2, \ldots, M)$. Choose nonsingular $\tilde{\mathcal{L}}_i \in \mathbb{R}^{n \times n}$ satisfying $[O_{l_i \times (n-l_i)} \ I_{l_i}] \tilde{\mathcal{L}}_i = \mathcal{L}_i$ $(i = 1, 2, \ldots, M)$ and $\omega_i \in \mathbb{R}_{++}$ $(i = 1, 2, \ldots, M)$ satisfying $\sum_{i=1}^M \omega_i = 1$. For each $i = 1, 2, \ldots, M$, apply proposition 2 to $(\sqrt{\mu} A, \mathcal{L}_i, \mu)$. Choose nonsingular $\tilde{\mathcal{L}}_i \in \mathbb{R}^{n \times n}$ satisfying $[O_{l_i \times (n-l_i)} \ I_{l_i}] \tilde{\mathcal{L}}_i = \mathcal{L}_i$ $(i = 1, 2, \ldots, M)$ and $\omega_i \in \mathbb{R}_{++}$ $(i = 1, 2, \ldots, M)$ satisfying $\sum_{i=1}^M \omega_i = 1$. For each $i = 1, 2, \ldots, M$, apply proposition 2 to $(\sqrt{\mu} A, \mathcal{L}_i, \mu)$. Choose nonsingular $\tilde{\mathcal{L}}_i \in \mathbb{R}^{n \times n}$ satisfying $[O_{l_i \times (n-l_i)} \ I_{l_i}] \tilde{\mathcal{L}}_i = \mathcal{L}_i$ $(i = 1, 2, \ldots, M)$ and $\omega_i \in \mathbb{R}_{++}$ $(i = 1, 2, \ldots, M)$ satisfying $\sum_{i=1}^M \omega_i = 1$. For each $i = 1, 2, \ldots, M$, apply proposition 2 to $(\sqrt{\mu} A, \mathcal{L}_i, \mu)$ to obtain $B^{(i)}_{\theta_0} \in \mathbb{R}^{l \times l_i}$ satisfying $(\sqrt{\mu} A)^T (\sqrt{\mu} A) - \mu_i \mathcal{L}_i^T B^{(i)}_{\theta_0} B^{(i)}_{\theta_0} \mathcal{L}_i \succeq O_{n \times n}$. Then $B_{\theta_0} : \mathbb{R}^l \times \mathbb{R}^{l_i} \times \cdots \times \mathbb{R}^{l_M} \to \mathbb{R}^l \times \mathbb{R}^{l_i} \times \cdots \times \mathbb{R}^{l_M} : (z_1, \ldots, z_M) \to (\sqrt{\mu} B^{(1)}_{\theta_0} z_1, \ldots, \sqrt{\mu} B^{(M)}_{\theta_0} z_M)$ ensures $J_{\Psi_{\omega_0} \circ \mathcal{L}} \in \Gamma_0(\mathbb{R}^n)$.

Proof. Verified by

$$A^T A - \mu \mathcal{L}^T B_{\theta_0} B_{\theta_0} \mathcal{L} = A^T A - \mu \sum_{i=1}^M \mu_i \mathcal{L}_i^T B^{(i)}_{\theta_0} B^{(i)}_{\theta_0} \mathcal{L}_i$$

$$= \mu \sum_{i=1}^M \left( \frac{\omega_i}{\mu} A^T A - \mu_i \mathcal{L}_i^T B^{(i)}_{\theta_0} B^{(i)}_{\theta_0} \mathcal{L}_i \right) \succeq O_{n \times n}. \quad \square$$

4. Numerical experiments

To demonstrate the effectiveness of the proposed penalties (LiGME penalties) and the proposed algorithm for the LiGME model (see algorithm 1), we present numerical experiments in four sparsity-rank-aware signal processing scenarios: (i) recovering a piecewise constant 1-d signal, (ii) deburring a piecewise constant image, (iii) filling missing entries of a low-rank matrix, which is a task so-called the matrix completion, (iv) filling missing entries of low-rank as well as piecewise constant matrix by handling two different LiGME penalties.

4.1. Piecewise constant 1-d signal recovery

In this section, we present a numerical experiment in a scenario of edge-preserving signal recovery by considering problem 1 with $(X, \mathcal{Y}, Z) = (\mathbb{R}^N, \mathbb{R}^M, \mathbb{R}^{N-1})$, $(N,M) := (128, 100)$, $\Psi = \| \cdot \|_1$, and $\mathcal{L}$ being the first order difference operator, i.e.

$$\mathcal{L} = D := \begin{bmatrix} -1 & 1 & \cdots & \cdots & \cdots & 1 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ -1 & 1 \end{bmatrix} \in \mathbb{R}^{(N-1) \times N}. \quad (38)$$

In this experiment, entries of $A \in \mathbb{R}^{M \times N}$ are drawn from i.i.d. zero-mean white Gaussian noise with unit variance. The observation $y \in \mathbb{R}^M$ is generated by $y = Ax^* + \varepsilon$, where $x^* \in \mathbb{R}^N$ is a piecewise constant signal (figure 3: dotted) and $\varepsilon \in \mathbb{R}^M$ is additive white Gaussian noise. The signal-to-noise ratio (SNR) is $-5$ dB, which is defined as

$$\text{SNR} : 10 \log_{10} \frac{\|x^*\|_2^2}{\|\varepsilon\|_2^2} \text{ (dB)}. \quad (39)$$

We compared minimizers of problem 1, estimated by algorithm 1, with two penalties: one is the standard convex total variation (TV), i.e. $(\| \cdot \|_1)_{B_0} \circ D = (\| \cdot \|_1)_{0} \circ D = \| \cdot \|_1 \circ D$, the
other is a nonconvex LiGME penalty $(\| \cdot \|_1)_B \circ D$ whose $B_{\theta} \in \mathbb{R}^{(N-1) \times (N-1)}$ is obtained by proposition 2 with

$$L = \tilde{D} := [e_1 | D^T]^T \in \mathbb{R}^{N \times N},$$

(40)

where $\theta = 0.99$ and $e_1 = (1, 0, \ldots, 0)^T \in \mathbb{R}^N$. Algorithm 1 with $\kappa = 1.001$ and $(\sigma, \tau)$ given in the footnote for theorem 1(b) is applied to the minimization problems, where the common initial estimate is set as $(x_0, v_0, w_0) = (0_X, 0_Z, 0_Z)$ for all experiments.

In algorithm 1, $\text{Prox}_{\gamma \| \cdot \|_1}$ for $\gamma \in \mathbb{R}^{++}$ can be calculated by the soft-thresholding whose $i$th component is

$$\left[\text{Prox}_{\gamma \| \cdot \|_1}\right]_i : \mathbb{R}^{N-1} \rightarrow \mathbb{R} : z = (z_1, \ldots, z_{N-1})^T \mapsto \begin{cases} 0, & \text{if } |z_i| \leq \gamma, \\ (|z_i| - \gamma) \frac{z_i}{|z_i|}, & \text{otherwise}. \end{cases}$$

(41)

Figure 1 shows dependency of recovering performance on the parameter $\mu$ in problem 1. The performance is measured by mean squared error (MSE) defined as the average of squared error (SE) over 100 independent realizations of the additive noise. From figure 1, we can see that (i) the best weights of the penalties are respectively $\mu_{TV} := 60$ for $\| \cdot \|_1 \circ D$ and $\mu_{LGME} := 900$ for $(\| \cdot \|_1)_B \circ D$ and (ii) the estimation by LiGME penalty with $\mu_{LGME}$ outperforms the standard convex TV penalty with $\mu_{TV}$ in the context of MSE.

Figure 2 shows dependency of the SE on the number of iterations under weights $(\mu_{TV}, \mu_{LGME})$. The accuracy of the approximation by the LiGME penalty becomes higher than the TV penalty after 400 iterations and SE for LiGME reaches 18.8% of SE for TV in the end.

Figure 3 shows the original signal and recovered signals by the penalties at 15 000 iteration. The estimation by LiGME $(\| \cdot \|_1)_B \circ D$ restores much more successfully the sharp edges than the standard convex TV, which also results in efficient noise suppression at 15 000 iteration depicted in figure 4.
4.2. Piecewise constant image deblurring

We present a numerical experiment in a scenario of image deblurring for piecewise constant $N$-by-$N$ image by considering problem 1 and example 3 with 

$$(\mathcal{M}, \mathcal{X}, \mathcal{Y}, Z_1, Z_2) = (2, \mathbb{R}^{N^2}, \mathbb{R}^{N^2}, \mathbb{R}^{N(N-1)}, \mathbb{R}^{N(N-1)}) \quad N = 16, \quad \Psi^{(1)} = \Psi^{(2)} = \| \cdot \|_1, \mu_1 = \mu_2 = 1, \quad \mathcal{L} = \mathcal{D} := [D_V, D_H]^T,$$

where the vertical difference operator $D_V \in \mathbb{R}^{N(N-1)\times N^2}$ and the horizontal difference operator $D_H \in \mathbb{R}^{N(N-1)\times N^2}$ are respectively defined as
with $D \in \mathbb{R}^{(N-1)\times N}$ in (38). The blur matrix\(^9\) $A \in \mathbb{R}^{N^2\times N^2}$ is designed by

$$A = \bar{A} \otimes \bar{A},$$

where $\otimes$ denotes the Kronecker product and the $(i,j)$-entry of the matrix $\bar{A} \in \mathbb{R}^{N\times N}$ is given by

$$\bar{A}_{ij} := \begin{cases} \frac{1}{\sqrt{1.62 \pi}} \exp \left(-\frac{|i-j|^2}{1.62}\right), & \text{if } |i-j| < 6, \\ 0, & \text{otherwise}. \end{cases}$$

The observation $y \in \mathbb{R}^{N^2}$ (figure 7(b)) is generated by $y = Ax^* + \varepsilon$, where $x^* \in \mathbb{R}^{N^2}$ is given by the vectorization\(^10\) of a piecewise constant image (figure 7(a)) and $\varepsilon \in \mathbb{R}^{N^2}$ is additive white Gaussian noise. The signal-to-noise ratio (SNR) defined in (39) is 20 dB. We compared minimizers of problem 1, estimated by algorithm 1, with two penalties: one is the anisotropic TV, i.e.

\[^9\] The blur matrix used in this experiment is more ill-conditioned than the random matrix used in section 4.1. The condition number, i.e. the ratio of the maximum singular value to the minimum singular value, of the blur matrix in (44) is about 593, and of the random matrix is about 12.4.

\[^10\] The vectorization of a matrix (or an image) is the mapping:

$$\text{vec} : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}^m : A \mapsto [a_1^\top, \ldots, a_n^\top]^\top,$$

where, for $i \in \{1, \ldots, n\}$, $a_i \in \mathbb{R}^m$ is the $i$th column vector of $A$. The inverse mapping of the vectorization $\text{vec}$ is denoted by $\text{vec}^{-1} : \mathbb{R}^m \rightarrow \mathbb{R}^{m \times n}$. 

---

**Figure 4.** Entries in $y - Ax^*$ (dotted black), $Ax_{\text{TV}} - Ax^*$ (dashed blue), and $Ax_{\text{LiGME}} - Ax^*$ (solid red), for $x^*$, $x_{\text{TV}}$, and $x_{\text{LiGME}}$ in figure 3.
\[(\| \cdot \|_1)_{\alpha_x} \circ D_N + (\| \cdot \|_1)_{\alpha_x} \circ D_H = \| \cdot \|_1 \circ D_N + \| \cdot \|_1 \circ D_H,\]

the other is a LiGME penalty \((\| \cdot \|_1)_{B_\theta} \circ \tilde{D}\) whose \(B_\theta = \begin{bmatrix} B_{\theta_1} & O_{N(N - 1)} \\ O_{N(N - 1)} & B_{\theta_2} \end{bmatrix} \in \mathbb{R}^{2N(N - 1) \times 2N(N - 1)}\)
is obtained by corollary 1 with \(\theta_1 = \theta_2 = 0.99, \omega_1 = \omega_2 = 1/2\), and \((\tilde{L}_1, \tilde{L}_2)\) given as

\[
\tilde{L}_1 = \tilde{D}_N := \begin{bmatrix} E \\ D_N \end{bmatrix} \in \mathbb{R}^{N^2 \times N^2}, \quad \tilde{L}_2 = \tilde{D}_H := \begin{bmatrix} I_N \\ O_{N(N - 1)} \\ D_H \end{bmatrix} \in \mathbb{R}^{N^2 \times N^2},
\]

where the \((i,j)\)-entry of \(E \in \mathbb{R}^{N^2 \times N^2}\) is defined as

\[
E_{ij} := \begin{cases} 1, & \text{if } (i - 1)N + 1 = j \\ 0, & \text{otherwise}. \end{cases}
\]

Algorithm 1 with \(\kappa = 1.001\) and \((\sigma, \tau)\) given in the footnote for theorem 1(b) is applied to the minimization problems, where the common initial estimate is set as \((x_0, v_0, w_0) = (0_X, 0_Z, 0_Z)\) for all experiments. The operator \(\text{Prox}_{\gamma \| \cdot \|_1}\) for \(\gamma \in \mathbb{R}^+\) in algorithm 1 can be calculated by (41).

Figure 5 shows dependency of recovering performance on the parameter \(\mu\) in problem 1. The performance is measured by mean squared error (MSE) defined as the average of SE in (42) over 100 independent realizations of the additive noise. From figure 5, we can see that (i) the best weights of the penalties are respectively \(\mu_{TV} := 0.013\) for \(\| \cdot \|_1 \circ \tilde{D}\) and \(\mu_{LiGME} := 0.03\) for \((\| \cdot \|_1)_{B_\theta} \circ \tilde{D}\) and (ii) the estimation by LiGME penalty with \(\mu_{LiGME}\) outperforms the anisotropic TV penalty with \(\mu_{TV}\) in the context of MSE.

Figure 6 shows dependency of the SE on the number of iterations under weights \((\mu_{TV}, \mu_{LiGME})\). The accuracy of the approximation by the LiGME penalty becomes higher than the anisotropic TV penalty from the beginning and SE for LiGME reaches 22.4% of SE for anisotropic TV in the end.

Figure 7 shows the original image, an observed image, and recovered images by the penalties at 5000 iteration. The deblurring by LiGME \((\| \cdot \|_1)_{B_\theta} \circ \tilde{D}\) restores much more successfully the sharp edges than the anisotropic TV.
4.3. Matrix completion by promoting low-rankness

We present a numerical experiment in a scenario of matrix completion by considering problem 1 with $(X, Y, Z) = (\mathbb{R}^{N_2}, \mathbb{R}^{N_2}, \mathbb{R}^{N_2})$, $N = 16$, $\Psi = \|\text{vec}^{-1}(\cdot)\|_{\text{nuc}}$ defined in example 2(c), and $\mathcal{L} = \text{Id}$. In this experiment, the $(i,j)$-entry of $A$ is given by

$$A_{ij} = \begin{cases} 1, & \text{if } i = j \in \Omega, \\ 0, & \text{otherwise}, \end{cases}$$

(48)

where $\Omega \subset \{1, \ldots, N_2\}$ satisfies $\#\Omega = N^2 - M$ with $M = 64$, i.e. 25% of entries are missing. The matrix $A^T A$ is singular because $\text{rank}(A) = N^2 - M$. The observation $y \in \mathbb{R}^{N_2}$ (figure 10(b)) is generated by $y = Ax^* + \varepsilon$, where $x^* \in \mathbb{R}^{N_2}$ is given by the vectorization of a low-rank matrix (figure 10(a)) and $\varepsilon \in \mathbb{R}^{N_2}$ is additive white Gaussian noise. The signal-to-noise ratio (SNR) defined in (39) is 30 dB. We compared minimizers of problem 1, estimated by algorithm 1, with two penalties: one is the nuclear norm, i.e. $(\|\text{vec}^{-1}(\cdot)\|_{\text{nuc}})_{B_0} = (\|\text{vec}^{-1}(\cdot)\|_{\text{nuc}})_{\Omega Z}$, the other is a LiGME penalty $(\|\text{vec}^{-1}(\cdot)\|_{\text{nuc}})_{B_\theta}$ whose $B_\theta \in \mathbb{R}^{N_2 \times N_2}$ is obtained by proposition 2 with $\theta = 0.99$ and $\mathcal{L} = \text{Id}$. Algorithm 1 with $\kappa = 1.001$ and $(\sigma, \tau)$ given in the footnote for theorem 1(b) is applied to the minimization problems, where the common initial estimate is set as $(x_0, v_0, w_0) = (0_X, 0_Z, 0_Z)$ for all experiments. In algorithm 1, the operator $\text{Prox}_{\gamma \|\text{vec}^{-1}(\cdot)\|_{\text{nuc}}}$ for $\gamma \in \mathbb{R}^{++}$ can be calculated by

$$\text{Prox}_{\gamma \|\text{vec}^{-1}(\cdot)\|_{\text{nuc}}}(z) = \text{vec} \left( U \text{diag}(\text{Prox}_{\gamma \|\cdot\|_1}([\sigma_1, \ldots, \sigma_N]^T)) V^T \right),$$

(49)

where $U \text{diag}([\sigma_1, \ldots, \sigma_N]) V^T$ ($\sigma_1 \geq \cdots \geq \sigma_N \geq 0$) is a singular value decomposition of $\text{vec}^{-1}(z) \in \mathbb{R}^{N^2 \times N}$.

Figure 8 shows dependency of recovering performance on the parameter $\mu$ in problem 1. The performance is measured by mean squared error (MSE) defined as the average of SE in (42) over 100 independent realizations of the additive noise. From figure 8, we can see that (i) the best weights of the penalties are respectively $\mu_{\text{nuc}} := 0.034$ for $\|\cdot\|_{\text{nuc}}$ and $\mu_{\text{LiGME}} := 0.1$.
Figure 7. (a) Original piecewise constant image whose pixels belong to \{0.25, 0.50, 0.75\}, (b) a noisy blurred image, (c) estimated image by using anisotropic TV penalty at \(k = 5000\) iteration, (d) estimated image by using LiGME penalty at \(k = 5000\) iteration. Each pixel is assigned a real value and displayed with under \(-0.2\) in black and over \(1.2\) in white.

Figure 8. MSE versus \(\mu\) in problem 1 at \(k = 500\) iteration for (a) the nuclear norm penalty \(\| \cdot \|_{\text{nuc}}\) and (b) LiGME penalty \((\| \cdot \|_{\text{nuc}})_{\theta_c}\).
Figure 9. SE versus iterations for the nuclear norm (dotted blue) and LiGME (solid red).

Figure 10. (a) Original low-rank matrix whose rank is 3, (b) observed matrix whose missing entries are displayed in white, (c) estimated matrix by using the nuclear norm penalty at $k = 500$ iteration, (d) estimated matrix by using LiGME penalty at $k = 500$ iteration. Each entry is displayed with under $-0.2$ in black and over $1.2$ in white.
for \( \| \cdot \|_{\text{nuc}} \) and (ii) the estimation by LiGME penalty with \( \mu_{\text{LiGME}} \) outperforms the nuclear norm penalty with \( \mu_{\text{nuc}} \) in the context of MSE.

Figure 9 shows dependency of the SE on the number of iterations under weights \( (\mu_{\text{nuc}}, \mu_{\text{LiGME}}) \). The accuracy of the approximation by the LiGME penalty becomes higher than the nuclear norm penalty from the beginning and SE for LiGME reaches 49.5% of SE for nuclear norm in the end.

Figure 10 shows the original matrix, an observed matrix, and recovered matrices by the penalties at 500 iteration and table 1 shows the singular values of the original matrix and the recovered matrices in figure 10. In the context of singular values in table 1, the recovered matrix by the LiGME penalty more accurately approximates the original than by the nuclear norm penalty. Especially, the number of singular values greater than \( 10^{-8} \) (num-rank) of the recovered matrix by the LiGME is equal to of the original.

### 4.4. Matrix completion by promoting low-rankness and smoothness

We present a numerical experiment in a scenario of matrix completion by considering problem 1 and example 3 with \(( \mathcal{M}, \mathcal{X}, \mathcal{Y}, \mathcal{Z}_1, \mathcal{Z}_2, \mathcal{Z}_3 ) = (3, \mathbb{R}^N, \mathbb{R}^N, \mathbb{R}^{(N-1)}, \mathbb{R}^{(N-1)}, \mathbb{R}^N), N = 16, (\Psi^{(1)}, \Psi^{(2)}, \Psi^{(3)}) = (\| \cdot \|_1, \| \cdot \|_1, \| \text{vec}^{-1}(\cdot) \|_{\text{nuc}}), \mu = 1, (\mathcal{L}_1, \mathcal{L}_2, \mathcal{L}_3) = (D_N, D_H, \text{Id})\).
The signal-to-noise ratio (SNR) defined in (39) is 20 dB, which is lower than the SNR set in (46). Algorithm 1 with

$\Psi_3 \circ \mathcal{L}$ (figure 13(b)) is generated by

In each penalty, $\omega$ is given by (48), which satisfies

- $a = 1$ for $B \in \{1, 2, 3\}$.

We compared minimizers of problem 1, estimated by algorithm 1, with four penalties:

(i) $\Psi_1 \circ \mathcal{L} := \mu_1 \left( \| \cdot \|_{B^{(1)}_0} \circ D_V + \| \cdot \|_{B^{(2)}_0} \circ D_H \right) + \mu_b \left( \| \text{vec}^{-1}(\cdot) \|_{\text{nuc}} \right)_{B^{(3)}_0}$

(ii) $\Psi_2 \circ \mathcal{L} := \mu_1 \left( \| \cdot \|_{B^{(1)}_{\omega_1}} \circ D_V + \| \cdot \|_{B^{(2)}_{\omega_2}} \circ D_H \right) + \mu_b \left( \| \text{vec}^{-1}(\cdot) \|_{\text{nuc}} \right)_{B^{(3)}_0}$

(iii) $\Psi_3 \circ \mathcal{L} := \mu_1 \left( \| \cdot \|_{B^{(1)}_{\omega_3}} \circ D_V + \| \cdot \|_{B^{(2)}_{\omega_3}} \circ D_H \right) + \mu_b \left( \| \text{vec}^{-1}(\cdot) \|_{\text{nuc}} \right)_{B^{(3)}_0}$

(iv) $\Psi_4 \circ \mathcal{L} := \mu_1 \left( \| \cdot \|_{B^{(1)}_\theta} \circ D_V + \| \cdot \|_{B^{(2)}_\theta} \circ D_H \right) + \mu_b \left( \| \text{vec}^{-1}(\cdot) \|_{\text{nuc}} \right)_{B^{(3)}_0}$.

In each penalty, $B^{(i)}_j$ ($i = 1, 2, 3$) are obtained by corollary 1 with $\mu_1 = \mu_2 = \mu_3 = \mu_b$, $a = 1$, and $B^{(1)}_j = \| \cdot \|_{B^{(1)}_j}$.

Figure 12. SE versus iterations for penalties $\Psi_1 \circ \mathcal{L}$ (dotted blue), $\Psi_3 \circ \mathcal{L}$ (dash-dotted green), $\Psi_4 \circ \mathcal{L}$ (dashed black), and $\Psi_{\Pi} \circ \mathcal{L}$ (solid red).

The observation $y \in \mathbb{R}^N$ (figure 13(b)) is generated by $y = Ax^* + \varepsilon$, where $x^* \in \mathbb{R}^N$ is given by the vectorization of a piecewise constant image (figure 13(a)) and $\varepsilon \in \mathbb{R}^N$ is additive white Gaussian noise. The signal-to-noise ratio (SNR) defined in (39) is 20 dB, which is lower than the SNR set in section 4.3.

We compared minimizers of problem 1, estimated by algorithm 1, with four penalties:

In each penalty, $B^{(i)}_j$ ($i = 1, 2, 3$) are obtained by corollary 1 with $\mu_1 = \mu_2 = \mu_3 = \mu_b$, $a = 1$, and $B^{(1)}_j = \| \cdot \|_{B^{(1)}_j}$.
Figure 13. (a) Original low-rank and piecewise-constant matrix which is the same as figures 10(a) and (b) observed matrix whose missing entries are displayed in white, (c) estimated matrix by using $\Psi_1 \circ \mathcal{J}$ at $k = 1000$ iteration, (d) estimated matrix by using $\Psi_2 \circ \mathcal{J}$ at $k = 1000$ iteration, (e) estimated matrix by using $\Psi_3 \circ \mathcal{J}$ at $k = 1000$ iteration, (f) estimated matrix by using $\Psi_4 \circ \mathcal{J}$ at $k = 1000$ iteration. Each entry is displayed with under $-0.2$ in black and over $1.2$ in white.

$(x_0, v_0, w_0) = (0, 0, 0)$ for all experiments. The operator $\text{Prox}_{\gamma}$ for $\gamma \in \mathbb{R}_{++}$ can be calculated by

$$\text{Prox}_{\gamma}(z) : Z_1 \times Z_2 \times Z_3 \to Z_1 \times Z_2 \times Z_3$$

$$(z_1, z_2, z_3) \mapsto (\text{Prox}_{\gamma}(z_1), \text{Prox}_{\gamma}(z_2), \text{Prox}_{\gamma}(z_3)).$$

(50)
where \text{Prox}_{\gamma \| \cdot \|_1} and \text{Prox}_{\gamma \| \text{vec}^{-1} \cdot \|_{\text{nuc}}} are given by (41) and (49) respectively.

Figure 11 shows dependency of recovering performance on the parameter \((\mu_a, \mu_b)\) in problem 1 and example 3. The performance is measured by mean squared error (MSE) defined as the average of SE in (42) over 100 independent realizations of the additive noise. From figure 11, we can see that (i) the best weights of the penalties are respectively \((\mu_{a, I}, \mu_{b, I}) = (0.015, 0.1)\) for \(\Psi_I \circ \mathcal{L}\), \((\mu_{a, II}, \mu_{b, II}) = (0.03, 0.15)\) for \(\Psi_{II} \circ \mathcal{L}\), and \((\mu_{a, III}, \mu_{b, III}) = (0.015, 0.15)\) for \(\Psi_{III} \circ \mathcal{L}\), and (ii) the estimations by \(\Psi_i \circ \mathcal{L}\) with \((\mu_{a, i}, \mu_{b, i})\) \((i = I, II, III, IV)\) outperform the convex penalty \(\Psi_I \circ \mathcal{L}\) in the context of MSE.

Figure 12 shows dependency of the SE on the number of iterations under weights \((\mu_{a, i}, \mu_{b, i})\) \((i = I, II, III, IV)\). The accuracy of the approximations by \(\Psi_i \circ \mathcal{L}\) \((i = I, II, III, IV)\) penalties become higher than the convex penalty \(\Psi_I \circ \mathcal{L}\) after 110 iterations and SE for \(\Psi_{II} \circ \mathcal{L}\), \(\Psi_{III} \circ \mathcal{L}\), and \(\Psi_{IV} \circ \mathcal{L}\) reaches respectively 78.1\%, 43.1\%, and 12.7\% of SE for \(\Psi_I \circ \mathcal{L}\) in the end.

Figure 13 shows the original matrix, an observed matrix, and recovered matrices by the penalties at 1000 iteration and figure 14 shows the difference between the original matrix and recovered matrices. From figure 14, the recovered matrix by \(\Psi_{IV} \circ \mathcal{L}\) approximates most accurately the original matrix.
5. Conclusion

In this paper, we have proposed the Linearly involved generalized Moreau enhanced (LiGME) model as a unified extension of the ideas in [56, 70, 71] for exploiting nonconvex penalties in the regularized least-squares models without losing their overall convexities. The proposed model can admit multiple nonconvex penalties without losing its overall convexity and thus is applicable to much broader scenarios including sparsity-rank-aware signal processing and machine learning. We have also proposed a proximal splitting type algorithm for the LiGME model under an overall-convexity condition. The proposed algorithm is guaranteed to converge to a globally optimal solution. Numerical experiments in four different sparsity-rank-aware signal processing scenarios demonstrate the effectiveness of the LiGME models and the proposed proximal splitting algorithm.

Acknowledgments

This work was supported in part by the Japan Society for the Promotion of Science under Grants-in-Aid 19H04134, 26730128 and 18K18122.

Appendix A. Selesnick’s algorithm for minimizing \( J(\|\cdot\|_1)_{\text{Prox}} \)

For problem (8) with \( \mathcal{X} = \mathbb{R}^n \), \( \mathcal{Y} = \mathbb{R}^m \), and a special \( B \), Selesnick presented an algorithm shown in fact A.1.

**Fact A.1 ([56, proposition 15]).** Let \( (A, B, y, \mu, \theta) \in \mathbb{R}^{m \times n} \times \mathbb{R}^{q \times n} \times \mathbb{R}^n \times \mathbb{R}_+ \times [0, 1] \). Suppose that \( B^* B = (\theta / \mu) A^* A \). Define \( T_{\text{Sel}} : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}^n \times \mathbb{R}^n : (x, v) \mapsto (\xi, \zeta) \) by

\[
\xi := \text{Prox}_{\tau \mu \| \cdot \|_1} \left[ x - \tau A^* \left( A (x + \theta (v - x)) - y \right) \right],
\zeta := \text{Prox}_{\tau \mu \| \cdot \|_1} \left[ v - \tau \theta A^* A (v - x) \right],
\]

where

\[
\tau \in \left( 0, \frac{2}{\max \left\{ 1, \frac{\theta}{1 - \theta} \right\} \sqrt{\rho(A^* A)}} \right).
\]

Then, for any initial point \( (x_0, v_0) \in \mathbb{R}^n \times \mathbb{R}^n \), the sequence \( (x_k)_{k \in \mathbb{N}} \subset \mathbb{R}^n \) generated by

\[
(x_{k+1}, v_{k+1}) = T_{\text{Sel}}(x_k, v_k)
\]

converges to a point in \( \arg \min_{x \in \mathbb{R}^n} J(\|\cdot\|_1)_{\text{Prox}}(x) \).

Appendix B. Proof of proposition 1

(a) Fermat’s rule (11) and the property (14) of conjugate functions yield

\[
B^* B \xi \in \arg \min_{\xi \in \mathbb{Z}} \Psi(\xi) + \frac{1}{2} \| B(\xi - \mathbb{L} \xi) \|_2^2 \ (0_Z) \quad \text{with dom} \left( \frac{1}{2} \| B(\cdot - \mathbb{L} \xi) \|_2^2 \right) = \mathbb{Z}
\]

also yields \( \partial \Psi(0_Z) + B^* B (0_Z - \mathbb{L} \xi) = \partial \left( \Psi(\cdot) + \frac{1}{2} \| B(\cdot - \mathbb{L} \xi) \|_2^2 \right) (0_Z), \) we have
\[ B^* B \mathbf{L} x \in \arg \min_{\hat{\mathbf{v}} \in Z} \Psi^*(\hat{\mathbf{v}}) \iff 0 \in \partial \left( \Psi(\cdot) + \frac{1}{2} \| B(\cdot - \mathbf{L} x) \|_Z^2 \right) (0) \iff 0 \in \arg \min_{\hat{\mathbf{v}} \in Z} \left( \Psi(\hat{\mathbf{v}}) + \frac{1}{2} \| B(\hat{\mathbf{v}} - \mathbf{L} x) \|_Z^2 \right), \]

where the 2nd last equivalence is due to (11) and the last equivalence is by definition of \( \Psi_B \).

(b) We shall show \((C_1) \Rightarrow (C_2)\). Fix \( y \in Y \) arbitrarily. Then we have, for every \( x \in X \),

\[
J_{\Psi_B}(x) = \frac{1}{2} \| y - Ax \|_Y^2 + \mu \Psi_B(\mathbf{L} x) = \frac{1}{2} \| y - Ax \|_Y^2 + \mu \Phi(\mathbf{L} x)
= \left( \frac{1}{2} \| y \|_Y^2 - \langle y, Ax \rangle \right) + \frac{1}{2} \| \mathbf{L} x \|_Z^2 + \mu \Phi(\mathbf{L} x)
- \mu \min_{\mathbf{v} \in Z} \left[ \Psi(\mathbf{v}) + \frac{1}{2} \| B(\mathbf{L} x - \mathbf{v}) \|_Z^2 \right]
= \frac{1}{2} \left( \| Ax \|_Y^2 - \mu \| B \xi \|_Z^2 \right) + \frac{1}{2} \| y \|_Y^2 - \langle y, Ax \rangle + \mu \Phi(\mathbf{L} x) + \mu \max_{\mathbf{v} \in Z} \psi_\mathbf{v}(\mathbf{v}),
\]

where

\[
\psi_\mathbf{v} : X \to \mathbb{R} : x \mapsto - \Psi(\mathbf{v}) + \frac{1}{2} \| Bv \|_Z^2 - \langle Bv, B(Lx) \rangle.
\]

Since \( \psi_\mathbf{v} \) is affine for every \( \mathbf{v} \in Z \) and \( \max_{\mathbf{v} \in Z} \psi_\mathbf{v}(0_X) \in \mathbb{R} \) due to \( \text{dom} \Psi = Z \) and coercivity of \( \Psi \), [3, proposition 9.3] yields \( \max_{\mathbf{v} \in Z} \psi_\mathbf{v} \in \Gamma_0(X) \).

Moreover, the assumption \( A^* A - \mu \xi_0 B^* B \Sigma \geq 0_X \) ensures that the function \( X \ni x \mapsto \frac{1}{2} \langle x, (A^* A - \mu \xi_0 B^* B \Sigma)x \rangle + \frac{1}{2} \| y \|_Y^2 - \langle y, Ax \rangle + \mu \Phi(\mathbf{L} x) \) also belongs to \( \Gamma_0(X) \).

Thus \( J_{\Psi_B}(x) \in \Gamma_0(X) \) holds.

Finally, since the affine function \( X \ni x \mapsto \frac{1}{2} \| Ax \|_Y^2 - \langle y, Ax \rangle \) in (B.2) does not affect the convexity of \( J_{\Psi_B} \), we have

\[
(C_2) \iff \frac{1}{2} \| Ax \|_Y^2 - \mu \| B \xi \|_Z^2 + \mu \Phi(\mathbf{L} x) + \mu \max_{\mathbf{v} \in Z} \psi_\mathbf{v}(\mathbf{v}) \in \Gamma_0(X) \iff (C_3),
\]

where the first equivalence holds by the expressions (B.2) and (B.1).

(a')

\[
\| B^* B \mathbf{L} x \|_\ast \leq 1 \iff B^* B \mathbf{L} x \in \arg \min_{\hat{\mathbf{v}} \in Z} \| \cdot \|_\ast^\dagger(\hat{\mathbf{v}})
\]

is verified by \( \| \cdot \|_\ast^\dagger = \begin{cases} 0 & \text{if } \| \cdot \|_\ast \leq 1, \\ +\infty & \text{otherwise}, \end{cases} \) (see e.g. [10, example 3.26]).

(b') We shall show \((C_3) \Rightarrow (C_1)\) by contraposition. Suppose \( A^* A - \mu \xi_0 B^* B \Sigma \not\geq 0_X \), i.e. there exists \( \hat{x} \in X \setminus \{ 0_X \} \) such that

\[
\langle \hat{x}, (A^* A - \mu \xi_0 B^* B \Sigma)\hat{x} \rangle < 0,
\]

which completes the proof.
and we shall prove $J_{\|\cdot\|_{\|\cdot\|_{x_0} \L}}^{(0)} \not\in \Gamma_0(\mathcal{A})$. By $\mu \in \mathbb{R}^+$, we have

$$ (B.5) \iff \|B\mathcal{L}\bar{x}\|^2_{\mathcal{B}} > \frac{1}{\mu} \|A\bar{x}\|^2_{\mathcal{B}} \geq 0 $$

implying thus $B^{*}B\mathcal{L}\bar{x} \neq 0_{\mathcal{B}}$ and $\|B^{*}B\mathcal{L}\bar{x}\|_{\mathcal{B}} = 1$ for $\bar{x} := (\|B^{*}B\mathcal{L}\bar{x}\|_{\mathcal{B}})^{-1} \mathcal{L} \bar{x} \in \mathcal{A}$. The statement (a') yields

$$ (\forall \lambda \in (0, 1)) J_{\|\cdot\|_{\|\cdot\|_{x_0} \L}}^{(0)}(\lambda \bar{x}) = \frac{1}{2} \|A(\lambda \bar{x})\|_2^2 + \mu \|\mathcal{L}(\lambda \bar{x})\|_{\mathcal{B}}^2 - \frac{\mu}{2} \|B\mathcal{L}(\lambda \bar{x})\|^2_{\mathcal{B}} = \frac{1}{2} \left( \|A\bar{x}\|^2_{\mathcal{B}} - \|B\mathcal{L}\bar{x}\|^2_{\mathcal{B}} \right) \lambda^2 + \mu \|\mathcal{L}\bar{x}\|_{\mathcal{B}}^2 $$

and $J_{\|\cdot\|_{\|\cdot\|_{x_0} \L}}^{(0)}(0_{\mathcal{A}}) = \frac{1}{2} \left( \|A0_{\mathcal{A}}\|^2_{\mathcal{B}} - \mu \|B0_{\mathcal{A}}\|_{\mathcal{B}}^2 \right) + \mu \|\mathcal{L}0_{\mathcal{A}}\|_{\mathcal{B}}^2 = 0$, from which we have

$$ \frac{1}{2} J_{\|\cdot\|_{\|\cdot\|_{x_0} \L}}^{(0)}(0_{\mathcal{A}}) + \frac{1}{2} J_{\|\cdot\|_{\|\cdot\|_{x_0} \L}}^{(0)}(\bar{x}) - J_{\|\cdot\|_{\|\cdot\|_{x_0} \L}}^{(0)} \left( \frac{0\bar{x} + \bar{x}}{2} \right) = \frac{1}{8} \left( \|A\bar{x}\|^2_{\mathcal{B}} - \|B\mathcal{L}\bar{x}\|^2_{\mathcal{B}} \right) < 0. $$

\[\square\]

**Appendix C. Proof of lemma 1**

We will show

$$ \text{span} \left( \text{dom} \left( \left( \Psi + \frac{1}{2} \|B \cdot \|_2^2 \right)^* \right) \right) - \text{ran}(B^{*}) \subset \text{cone} \left( \text{dom} \left( \left( \Psi + \frac{1}{2} \|B \cdot \|_2^2 \right)^* \right) \right) - \text{ran}(B^{*}), $$

which is equivalent to (31). By the even symmetry of $\Psi$, we have for $v \in \mathcal{Z}$

$$ \left( \Psi + \frac{1}{2} \|B \cdot \|_2^2 \right)^* v = \sup_{w \in \mathcal{Z}} \left( w, v \right)_\mathcal{Z} - \left( \Psi(w) - \frac{1}{2} \|Bw\|_2^2 \right) = \left( \Psi + \frac{1}{2} \|B \cdot \|_2^2 \right)^* v $$

and

$$ D := \text{dom} \left( \left( \Psi + \frac{1}{2} \|B \cdot \|_2^2 \right)^* \right) = - \text{dom} \left( \left( \Psi + \frac{1}{2} \|B \cdot \|_2^2 \right)^* \right). $$

Let $v = \sum_{i \in I} \alpha_i (v_i - w_i) \in \text{span}(D - \text{ran}(B^{*}))$ for some $(\alpha_i, v_i, w_i)_{i \in I} \subset (\mathbb{R} \setminus \{0\}) \times \mathcal{D} \times \text{ran}(B^{*})$ with finite $I \subset \mathbb{N}$. Then we have

$$ v = \sum_{i \in I} |\alpha_i| (\text{sgn}(\alpha_i) v_i) - \sum_{i \in I} \alpha_i w_i = \sum_{i \in I} |\alpha_i| \left( \sum_{i \in I} \frac{|\alpha_i|}{\sum_{i \in I} |\alpha_i|} (\text{sgn}(\alpha_i) v_i) - \frac{\sum_{i \in I} \alpha_i w_i}{\sum_{i \in I} |\alpha_i|} \right), $$

where $\sum_{i \in I} \alpha_i w_i \in \text{ran}(B^{*})$. Moreover, by $\text{sgn}(\alpha_i) v_i \in D$ ($i \in I$) due to (C.1) and by $\sum_{i \in I} (\text{sgn}(\alpha_i) v_i) \in D$ due to the convexity of $D$, (C.2) implies $v \in \text{cone}(D - \text{ran}(B^{*}))$.

**Appendix D. Proof of theorem 1**

**Proof of (a)**: Recall that, under the assumption in problem 1, (B.2) gives an expression of $J_{\Psi_0 \mathcal{L}}$ as a sum of convex functions. The proof of (a) is decomposed into two steps.
(Step 1) By applying properties of the subdifferential in section 2.1, we will derive an alternative characterization of $S = \{ x^* \in X \mid 0 \in \partial J_{\Psi}(x^*) \}$ in terms of zeros of the sum of an affine operator $F$ and a set-valued operator $G$ involving $\partial \Psi$ (see claim D.1).

Claim D.1. In problem 1, for any $x^* \in X$, we have $x^* \in S$ if and only if there exists $(v^*, w^*) \in Z \times Z$ s.t.

$$
(0_X, 0_Z, 0_Z) \in F(x^*, v^*, w^*) + G(x^*, v^*, w^*),
$$

(12)

where $F: H \to H$ and $G: H \to 2^H$ are defined as

$F(x, v, w) := ((A^*A - \mu \Sigma^* B^* B \Sigma)x - A^*y, \mu B^* B v, 0_Z)$,

$G(x, v, w) := \{ \mu \Sigma^* B^* B v + \mu \Sigma^* w \} \times (-\mu B^* B \Sigma x + \mu \partial \Psi(v)) \times (-\mu \Sigma x + \mu (\partial \Psi)^*(w))$.

(Step 2) By using (34), we will confirm for any $x^* \in X$ that

$$(x^*, v^*, w^*) \in \text{Fix}(T_{\text{LIGME}}) \iff T_{\text{LIGME}}(x^*, v^*, w^*) = (x^*, v^*, w^*)$$

(13)

$$\iff (\mathcal{Q} - F)(x^*, v^*, w^*) \in (\mathcal{Q} + G)(x^*, v^*, w^*) \quad (\iff (D.1))$$

implying thus, with claim D.1, $x^* \in S$ if and only if there exists $(v^*, w^*) \in Z \times Z$ such that $(x^*, v^*, w^*) \in \text{Fix}(T_{\text{LIGME}})$.

For proof of claim D.1, we will use, in (B.2) and (B.3),

$$
(x \in X) \quad \max_{v \in Z} \psi_v(x) = \max_{v \in Z} \left( (v, B^* B \Sigma x - \Psi(v) - \frac{1}{2} \| Bv \|_Z^2 \right)
$$

$$= \left( \Psi + \frac{1}{2} \| B \cdot \|_Z \right)^* \circ B^* \circ B \Sigma(x),
$$

(14)

and

$$\text{dom} \left( \left( \Psi + \frac{1}{2} \| B \cdot \|_Z \right)^* \circ B^* \right) = \tilde{Z},
$$

(15)

where (15) is verified, with the coercivity of $\Psi$, by

$$
(z \in \tilde{Z}) \quad \left( \Psi + \frac{1}{2} \| B \cdot \|_Z \right)^* \circ B^* \circ B \Sigma(x) = \sup_{v \in Z} \left( (v, B^* z - \Psi(v) - \frac{1}{2} \| Bv \|_Z^2 \right)
$$

$$\leq \sup_{v \in Z} \left( -\Psi(v) \right) + \sup_{v \in Z} \left( (Bv, z - \frac{1}{2} \| Bv \|_Z^2 \right)
$$

$$= \max_{v \in Z} \left( -\Psi(v) \right) + \max_{v \in Z} \left( (Bv, z - \frac{1}{2} \| Bv \|_Z^2 \right) < \infty.
$$

Now, we shall prove steps 1 and 2.

Step 1: Proof of claim D.1. Since the first three terms of the RHS of (B.2) are differentiable over $X$, the sum rule (12) implies

$$
\partial J_{\Psi}(x) = \nabla \left( \frac{1}{2} \langle x, (A^*A - \mu \Sigma^* B^* B \Sigma)x \rangle_X + \frac{1}{2} \| v \|_Z^2 - \langle y, Ax \rangle_Y \right) + \partial \left( \mu \Psi \circ \Sigma \circ \mu \max_{v \in Z} \psi_v \right)(x)
$$

$$= (A^*A - \mu \Sigma^* B^* B \Sigma)x - A^*y + \mu \partial \left( \Psi \circ \Sigma \circ \mu \max_{v \in Z} \psi_v \right)(x).
$$

(16)
Moreover, by $\text{dom}(\max_{\xi \in Z} \psi_\eta) = \mathcal{X}$ due to (D.3) and (D.4), the sum rule (12) decomposes (D.5) as
\[ \partial J_{\Phi^{\psi_\eta}}(x) = (A^*A - \mu \Sigma^\ast B^* B \Sigma)x - A^*y + \mu \partial (\Psi \circ \Sigma)(x) + \mu \partial \left( \max_{\nu \in Z} \nu_\eta \right)(x). \]
(D.6)

Apply the chain rule (13) to $\partial(\Psi \circ \Sigma)$ with $\text{dom}(\Psi) = Z$ for simplification
\[ \partial J_{\Phi^{\psi_\eta}}(x) = (A^*A - \mu \Sigma^\ast B^* B \Sigma)x - A^*y + \mu \Sigma^\ast \partial \Psi(\Sigma x) + \mu \partial \left( \max_{\nu \in Z} \nu_\eta \right)(x). \]
(D.7)

Apply again the chain rule (13) to (D.3) with (D.4) for
\[ \partial \left( \max_{\nu \in Z} \nu_\eta \right) = (B \Sigma)^\ast \partial \left[ \left( \Psi + \frac{1}{2} \|B \cdot \|_Z^2 \right)^\ast \circ B^\ast \right] \circ B \Sigma, \]
and to $\partial \left( \left( \Psi + \frac{1}{2} \|B \cdot \|_Z^2 \right)^\ast \circ B^\ast \right)$ in (D.8) with (31) to deduce further simplification
\[ \partial J_{\Phi^{\psi_\eta}}(x) = (A^*A - \mu \Sigma^\ast B^* B \Sigma)x - A^*y + \mu \Sigma^\ast \partial \Psi(\Sigma x) + \mu (B^* B \Sigma)^\ast \partial \left( \Psi + \frac{1}{2} \|B \cdot \|_Z^2 \right)^\ast (B^* B \Sigma x). \]
(D.9)

Furthermore, by $w^* \in \partial \Psi(\Sigma x)$ $\Leftrightarrow \Sigma x^* \in \partial \Psi^\ast(w^*)$ and $v^* \in \partial \left( \Psi + \frac{1}{2} \|B \cdot \|_Z^2 \right)^\ast (B^* B \Sigma x^*) \Leftrightarrow B^* B \Sigma x^* \in \partial \left( \Psi + \frac{1}{2} \|B \cdot \|_Z^2 \right)^\ast (B^* B \Sigma x^*)$ due to the property (14) and the sum rule (12) with $\text{dom}(\|B \cdot \|_Z^2) = Z$, we deduce from (D.9)
\[
\begin{align*}
x^* & \in S \Leftrightarrow 0 \chi \in \partial J_{\Phi^{\psi_\eta}}(x^*) \Leftrightarrow \begin{cases} 0 \chi = (A^*A - \mu \Sigma^\ast B^* B \Sigma)x^* - A^*y + \mu \Sigma^\ast w^* + \mu (B^* B \Sigma)^\ast v^*, \\
\Sigma x^* \in \partial \Psi^\ast(w^*) \end{cases} \Leftrightarrow \begin{cases} 0 \chi = (A^*A - \mu \Sigma^\ast B^* B \Sigma)x^* - A^*y + \mu \Sigma^\ast w^* + \mu \Sigma^\ast w^*, \\
0 \Sigma \in \partial \Psi^\ast(w^*) \end{cases} \Leftrightarrow 0 \chi, 0 \Sigma, 0 \Sigma \in F(x^*, v^*, w^*) + G(x^*, v^*, w^*) \end{align*}
\]
which completes the proof of claim D.1.

**Step 2:** (D.2) is verified by the definitions of $T_{\text{LIGME}}$ and $\Psi$ in theorem 1 as
\[
T_{\text{LIGME}}(x, v, w) = (\xi, \zeta, \eta)
\]
\[ \begin{align*}
\begin{cases} 0 \chi & = (A^*A - \mu \Sigma^\ast B^* B \Sigma)x - \mu \Sigma^\ast B^* B v - \mu \Sigma^\ast w + A^*y = \sigma \xi, \\
0 \Sigma & \in \partial \mu \Sigma x^* + \mu \partial \Psi^\ast(w^*) \\
0 \Sigma & \in - \mu B^* B \Sigma x^* + \mu B^* B v^* + \mu \partial \Psi^\ast(v^*) \\
0 \Sigma & \in - \mu \Sigma x^* + \mu \partial \Psi^\ast(w^*) \end{cases} \Leftrightarrow (0 \chi, 0 \Sigma, 0 \Sigma) \in F(x^*, v^*, w^*) + G(x^*, v^*, w^*) 
\end{align*}
\]
which we used the expression of the proximity operator as the resolvent of a subdifferential.

**Proof of (b):** We first prove $\Psi \succeq O_X$ under the condition (33). The Schur complement (see e.g. [37, theorem 7.7.6]) yields
\[ \mathcal{P} \gg \mathcal{O}_H \Leftrightarrow \sigma \mathrm{Id} - \left[ \frac{\mu A^*B^*B}{\tau} - \frac{\mu L^*L}{\tau} \right] \begin{bmatrix} \tau \mathrm{Id} & \mu B^* \mathcal{O} \\ \mu \mathrm{Id} & -\mu L^*L \end{bmatrix}^{-1} \begin{bmatrix} \mu B^* \mathcal{O} \\ -\mu L^*L \end{bmatrix} \gg \mathcal{O}_X. \]

\[ \Leftrightarrow \sigma \mathrm{Id} - \frac{\mu^2}{\tau} \begin{bmatrix} L^*(B^*B)^2L \\ -\mu \mathcal{O} \end{bmatrix} \gg \mathcal{O}_X. \]

\[ \Leftrightarrow \left( \sigma \mathrm{Id} - \frac{\kappa}{2} A^*A - \mu L^*L \right) + \left( \frac{\kappa}{2} A^*A - \frac{\mu^2}{\tau} L^*(B^*B)^2L \right) \gg \mathcal{O}_X. \]

From the condition (33), it is sufficient to show that \( \frac{\kappa}{2} A^*A - \frac{\mu^2}{\tau} L^*(B^*B)^2L \geq \mathcal{O}_X \). Recalling \( \|B^*B\|_{\text{op}} = \|B^*\|_{\text{op}}^2 \) for \( B \in \mathcal{B}(\mathcal{Z}, \mathcal{Z}) \) and using the condition (33), we have

\[ (\forall x \in \mathcal{X}) \quad \left\langle x, \left( \frac{\mu^2}{\tau} L^*(B^*B)^2L \right) x \right\rangle_{\mathcal{X}} = \frac{\mu^2}{\tau} \|B^*B\mathcal{O}x\|_2^2 \leq \frac{\mu^2}{\tau} \|B\|_{\text{op}} \|B\mathcal{O}x\|_{\mathcal{Z}}^2 \]

\[ \leq \mu^2 \left( \frac{\kappa}{2} + \frac{2}{\tau} \right) \|B\|_{\text{op}}^2 \|B\mathcal{O}x\|_{\mathcal{Z}}^2 \leq \mu \frac{\kappa}{2} \|B\mathcal{O}x\|_{\mathcal{Z}}^2. \]

which yields

\[ (\forall x \in \mathcal{X}) \quad \left\langle x, \left( \frac{\kappa}{2} A^*A - \frac{\mu^2}{\tau} L^*(B^*B)^2L \right) x \right\rangle_{\mathcal{X}} \geq \frac{\kappa}{2} \left\langle x, (A^*A - \mu L^*B^*B) x \right\rangle_{\mathcal{X}} \geq 0, \]

where the last inequality is due to the assumption \( A^*A - \mu L^*B^*B \geq \mathcal{O}_X \) in problem 1.

Next, we prove that \( T_{\text{LGME}} \) is \( \frac{\kappa}{2\tau} \)-averaged nonexpansive over \((\mathcal{H}, \langle \cdot, \cdot \rangle, \| \cdot \|_{\mathcal{H}})\). By applying \( \mathcal{P} \gg \mathcal{O}_H \) to (D.10), we have for \((x, v, w) \in \mathcal{H} \times \mathcal{X} \times \mathcal{X}\),

\[ T_{\text{LGME}}(x, v, w) = (\xi, \zeta, \eta) \Leftrightarrow (\mathrm{Id} - \mathcal{P}^{-1} \circ F)(x, v, w) \in (\mathrm{Id} + \mathcal{P}^{-1} \circ G)(\xi, \zeta, \eta). \]

Moreover, as will be shown in the end of this proof, \( \mathcal{P}^{-1} \circ G \) is maximally monotone over \((\mathcal{H}, \langle \cdot, \cdot \rangle, \| \cdot \|_{\mathcal{H}})\), by which the resolvent \((\mathrm{Id} + \mathcal{P}^{-1} \circ G)^{-1}\) is guaranteed to be single-valued and therefore

\[ T_{\text{LGME}} = (\mathrm{Id} + \mathcal{P}^{-1} \circ G)^{-1} \circ (\mathrm{Id} - \mathcal{P}^{-1} \circ F), \]

which is \( \frac{1}{2\tau} \)-averaged nonexpansive in \((\mathcal{H}, \langle \cdot, \cdot \rangle, \| \cdot \|_{\mathcal{H}})\).

To show that \( T_{\text{LGME}} \) is \( \frac{\kappa}{2\tau} \)-averaged nonexpansive in \((\mathcal{H}, \langle \cdot, \cdot \rangle, \| \cdot \|_{\mathcal{H}})\), fact 1 tells us it is sufficient to show the nonexpansiveness of \( \mathrm{Id} - \kappa \mathcal{P}^{-1} \circ F \). Define first

\[ \mathcal{P} \gg \mathcal{O}_H \Leftrightarrow \sigma \mathrm{Id} \quad \text{satisfies} \quad F(x, v, w) = M \begin{bmatrix} x \\ v \\ w \end{bmatrix} + \begin{bmatrix} -A^*y \\ 0_{\mathcal{Z}} \\ 0_{\mathcal{Z}} \end{bmatrix} \text{ for every } (x, v, w) \in \mathcal{H}, \quad M^* = M, \quad \text{and} \quad \mathcal{M} \gg \mathcal{O}_H \text{ (due to the assumption } A^*A - \mu L^*B^*B \geq \mathcal{O}_X \text{ in problem 1). Then we have for all } u_1, u_2 \in \mathcal{H}, \]

31
where the last equivalence is due to the Schur complement. Moreover, since for every \( \mathbf{u}_1, \mathbf{u}_2 \in \mathcal{H} \)

\[
\langle \mathbf{u}_1, \mathbf{Mu}_1 \rangle_{\mathcal{H}} + \langle \mathbf{u}_1, \mathbf{M} \mathbf{u}_2 \rangle_{\mathcal{H}} + \langle \mathbf{u}_2, \mathbf{M} \mathbf{u}_1 \rangle_{\mathcal{H}} + 2\kappa^{-1} \langle \mathbf{u}_2, \mathbf{P} \mathbf{u}_2 \rangle_{\mathcal{H}}
\]

we have

\[
\mathbf{u}_1 - \mathbf{u}_2 \|^2_\mathcal{H} = 2\kappa \left( \langle \mathbf{u}_1, \mathbf{u}_2 \rangle_{\mathcal{H}} - \frac{\kappa}{2} \mathbf{M} \mathbf{P}^{-1} \mathbf{M} \right) (\mathbf{u}_1 - \mathbf{u}_2)_{\mathcal{H}},
\]

which implies

\[
(\mathbf{Id} - \kappa \mathbf{P}^{-1} \circ \mathbf{F}) \text{ is nonexpansive} \iff M - \frac{\kappa}{2} \mathbf{M} \mathbf{P}^{-1} \mathbf{M} \succeq \mathcal{O}_{\mathcal{H}} \iff \begin{bmatrix} M & M \\ M & 2\kappa^{-1} \mathbf{P} \end{bmatrix} \succeq \mathcal{O}_{\mathcal{H} \times \mathcal{H}},
\]

Indeed, by the Schur complement, we have

\[
\begin{bmatrix}
\sigma \mathbf{Id} - (\kappa/2) \mathbf{A}^* \mathbf{A} & \mathbf{O}_{B(\mathcal{Z}, \mathcal{X})} \\
\mathbf{O}_{B(\mathcal{X}, \mathcal{Z})} & \mathbf{O}_{\mathcal{Z} \times \mathcal{Z}}
\end{bmatrix} \succ \mathcal{O}_{\mathcal{H}} \iff \sigma \mathbf{Id} - \frac{\kappa}{2} \mathbf{A}^* \mathbf{A} - \mu \mathbf{\Sigma} \succ \mathcal{O}_{\mathcal{X}} \quad \text{(\ref{eq:D.14})}
\]

and

\[
\begin{bmatrix}
(\kappa/2) \mathbf{A}^* \mathbf{B} \mathbf{L} & \mu \mathbf{\Sigma} \mathbf{B}^* \\
-\mu \mathbf{B} \mathbf{L} & \mathbf{Id} - (\kappa/2) \mathbf{B}^* \mathbf{B} \mathbf{L}
\end{bmatrix}
\]

\[
\begin{bmatrix}
\mathbf{L} & \mathbf{O}_{\mathcal{Z}} \\
\mathbf{O}_{\mathcal{Z}} & \mathbf{Id}
\end{bmatrix}
\]

\[
\begin{bmatrix}
\mu \mathbf{B}^* \\
-\mu \mathbf{B}^* \mathbf{L} & \mathbf{Id} - (\kappa/2) \mathbf{B}^* \mathbf{B} \mathbf{L}
\end{bmatrix}
\]

\[
\begin{bmatrix}
\mathbf{L} & \mathbf{O}_{\mathcal{Z}} \\
\mathbf{O}_{\mathcal{Z}} & \mathbf{Id}
\end{bmatrix}
\]

\[
\begin{bmatrix}
\mu \mathbf{B}^* \\
-\mu \mathbf{B}^* \mathbf{L} & \mathbf{Id} - (\kappa/2) \mathbf{B}^* \mathbf{B} \mathbf{L}
\end{bmatrix}
\]

\[
\begin{bmatrix}
\mathbf{L} & \mathbf{O}_{\mathcal{Z}} \\
\mathbf{O}_{\mathcal{Z}} & \mathbf{Id}
\end{bmatrix}
\]

\[
\begin{bmatrix}
\mu \mathbf{B}^* \\
-\mu \mathbf{B}^* \mathbf{L} & \mathbf{Id} - (\kappa/2) \mathbf{B}^* \mathbf{B} \mathbf{L}
\end{bmatrix}
\]

\[
\begin{bmatrix}
\mathbf{L} & \mathbf{O}_{\mathcal{Z}} \\
\mathbf{O}_{\mathcal{Z}} & \mathbf{Id}
\end{bmatrix}
\]
implying thus (RHS of (D.14)) \( \mathcal{Q} - \frac{\tau}{2} M \succeq O_\mathcal{H} \). The RHS of (D.14) is ensured by (33) because for every \( v_1, v_2 \in \mathcal{Z} \)

\[
\begin{bmatrix}
\begin{bmatrix}
\frac{\partial^2 \mathcal{B}}{\partial \mathcal{Y} \partial \mathcal{X}} - \mu \mathcal{B} \mathcal{B}^* \mathcal{B} \\
-\mu \mathcal{B} \mathcal{B}^* \mathcal{B}
\end{bmatrix}
\begin{bmatrix}
\mathcal{V}_1 \\
\mathcal{V}_2
\end{bmatrix}
\end{bmatrix}
\]

\( Z \times Z \)

\[ = (v_1, \frac{\kappa}{2} \mathcal{B} \mathcal{B}^* v_2) Z - (v_1, \mu \mathcal{B} \mathcal{B}^* v_2) Z + (v_2, \tau \mathcal{I} \mathcal{D} - \frac{\kappa}{2} \mathcal{B}^* \mathcal{B}) v_1 Z
\]

\[ = 2 \mu \kappa \frac{\kappa}{2} \mathcal{B} \mathcal{B}^* v_2 Z + (\tau \mathcal{V}^2) Z - \frac{\mu}{\kappa} \mathcal{V}^2 Z - \mu \left( \frac{\kappa}{2} + \frac{\kappa}{2} \right) \| \mathcal{V} \|_Z^2
\]

\[ \geq \tau \| \mathcal{V} \|_Z^2 - \mu \left( \frac{\kappa}{2} + \frac{\kappa}{2} \right) \| \mathcal{V} \|_Z^2 \geq \left( \tau - \mu \left( \frac{\kappa}{2} + \frac{\kappa}{2} \right) \| \mathcal{V} \|_Z^2 \right) \| \mathcal{V} \|_Z^2 \geq 0.
\]

Therefore we have proved that \( \mathcal{I} \mathcal{D} - \mathcal{Q}^{-1} \circ \mathcal{G} \) is \( \frac{1}{\kappa} \)-averaged nonexpansive and that \( T_{\text{LGE}} = (\mathcal{I} \mathcal{D} - \mathcal{Q}^{-1} \circ \mathcal{G})^{-1} \circ (\mathcal{I} \mathcal{D} - \mathcal{Q}^{-1} \circ \mathcal{F}) \) is \( \frac{1}{\kappa} \)-averaged nonexpansive over \( (\mathcal{H}, \langle \cdot, \cdot \rangle_\mathcal{H}, \| \cdot \|_\mathcal{H}) \).

Finally, the maximal monotonicity of \( \mathcal{Q}^{-1} \circ \mathcal{G} \) over \( (\mathcal{H}, \langle \cdot, \cdot \rangle_\mathcal{H}, \| \cdot \|_\mathcal{H}) \) is shown as follows. Let \( G_1 : \mathcal{H} \times \mathcal{Z} \times Z \to 2^\mathcal{H} : (v, w, l) \mapsto \{ 0 \} \times (\mu \mathcal{D} \mathcal{W}(v)) \times (\mu \mathcal{D} \mathcal{W}(w)) \) and \( G_2 : \mathcal{H} \to \mathcal{H} : (v, w) \mapsto (\mu \mathcal{D} \mathcal{W} \mathcal{B}^* \mathcal{B} \mathcal{B}^* \mathcal{W} \mathcal{L} \mathcal{X} - \mu \mathcal{D} \mathcal{X} \mathcal{L}) \). Then \( G_1 \) is maximally monotone over \( (\mathcal{H}, \langle \cdot, \cdot \rangle_\mathcal{H}, \| \cdot \|_\mathcal{H}) \) by [3, theorem 20.48, proposition 16.9 and 20.23]. Also, \( G_2 \) is a bounded linear skew-symmetric operator, i.e. \( G_2^* = -G_2 \), and is thus maximally monotone over \( (\mathcal{H}, \langle \cdot, \cdot \rangle_\mathcal{H}, \| \cdot \|_\mathcal{H}) \) by [3, example 20.35]. Then, by \( \text{dom}(G_2) = \mathcal{H} \) and [3, corollary 25.51], \( G = G_1 + G_2 \) is maximally monotone over \( (\mathcal{H}, \langle \cdot, \cdot \rangle_\mathcal{H}, \| \cdot \|_\mathcal{H}) \), which implies the monotonicity of \( \mathcal{Q}^{-1} \circ \mathcal{G} \) over \( (\mathcal{H}, \langle \cdot, \cdot \rangle_\mathcal{H}, \| \cdot \|_\mathcal{H}) \). Finally, we confirm the maximal monotonicity of \( \mathcal{Q}^{-1} \circ \mathcal{G} \) over \( (\mathcal{H}, \langle \cdot, \cdot \rangle_\mathcal{H}, \| \cdot \|_\mathcal{H}) \) by contradiction. Assume that there exists \( (u, z) \not\in \text{gra}(\mathcal{Q}^{-1} \circ \mathcal{G}) \), which means \( (u, \mathcal{Q} z) \not\in \text{gra}(\mathcal{G}) \), such that for all \( (u', z') \in \text{gra}(\mathcal{Q}^{-1} \circ \mathcal{G}) \), \( \langle u - u', z - z' \rangle_\mathcal{H} \geq \langle u - u', \mathcal{Q} (z - z') \rangle_\mathcal{H} \geq 0 \). However, since \( (u', \mathcal{Q} z') \in \text{gra}(\mathcal{G}) \), it contradicts the maximal monotonicity of \( \mathcal{G} \) over \( (\mathcal{H}, \langle \cdot, \cdot \rangle_\mathcal{H}, \| \cdot \|_\mathcal{H}) \).

Proof of (C): Thanks to (b), the direct application of Krasnosel’skiǐ–Mann iteration in fact 2 to \( T_{\text{LGE}} \) yields (c).

\[ \square \]

Appendix E. Proof of proposition 2

If \( \theta = 0 \), we have \( \mathcal{F}_\theta = \mathcal{O} \) and \( J_{\mathcal{F}_\theta \circ \mathcal{D}} \in \Gamma_0(\mathbb{R}^n) \) for all \( y \in \mathbb{R}^m \) by proposition 1(b).

Let \( \theta \in (0, 1) \). Proposition 1(b) shows

\[
(\mathcal{F}_\theta \circ \mathcal{D} \circ \mathcal{L}) \mathcal{L} = \mathcal{L}
\]

and the definition (37) of \( \mathcal{A}_1 \) and \( \mathcal{A}_2 \) show

\[
O_n \preceq (\mathcal{A}_2^{-1})^{\top} \mathcal{A}_1^{-1} A = \mathcal{A}_1 A - \mu \mathcal{D}^\top \mathcal{B}_n \mathcal{B}_n \mathcal{L}
\]

\(
\Leftrightarrow O_n \preceq (\mathcal{A}_2^{-1})^{\top} (\mathcal{A}_1^{-1}) - \mu (O_n O_n) I_1 \mathcal{L} = \mathcal{A}_1 \mathcal{A}_1 \mathcal{A}_1 - \mu (O_n O_n) I_1 \mathcal{L}
\]

\[
= \left[ \begin{array}{c}
\mathcal{A}_1 \\
\mathcal{A}_2
\end{array} \right] \left[ \begin{array}{c}
\mathcal{A}_1 \\
\mathcal{A}_2
\end{array} \right] - \mu \mathcal{B}_n \mathcal{B}_n \mathcal{L}
\]

\[
= \left[ \begin{array}{c}
\mathcal{A}_1^\top \mathcal{A}_1 \\
\mathcal{A}_2^\top \mathcal{A}_2 - \mu \mathcal{B}_n \mathcal{B}_n \mathcal{L}
\end{array} \right].
\]
Note that $\tilde{A}_1^\top \tilde{A}_1 \succeq O$, holds obviously and $\tilde{A}_1^\top \tilde{A}_1 (\tilde{A}_1^\top \tilde{A}_1)^\dagger \tilde{A}_2 = \tilde{A}_1^\top \tilde{A}_2$ holds due to $\text{ran}(\tilde{A}_1^\top \tilde{A}_2) \subset \text{ran}(\tilde{A}_1) = \text{ran}(\tilde{A}_1^\top \tilde{A}_1) \subset \text{null}(\tilde{A}_1^\top \tilde{A}_1 (\tilde{A}_1^\top \tilde{A}_1)^\dagger - I_{n-1})$. Thus [2, theorem 1] implies

\[
[(\text{LHS of (E.1)}) \iff (\text{RHS of (E.2)})] \iff \tilde{A}_2^\top \tilde{A}_2 - \mu \tilde{B}_B^\top \tilde{B}_B - \tilde{A}_1^\top \tilde{A}_1 (\tilde{A}_1^\top \tilde{A}_1)^\dagger \tilde{A}_1^\top \tilde{A}_2 \succeq O. \quad (E.3)
\]

Since $B_B$ in (36) satisfies $\theta^{-1} \mu \tilde{B}_B^\top \tilde{B}_B = \tilde{A}_2^\top \tilde{A}_2 - \tilde{A}_1^\top \tilde{A}_1 (\tilde{A}_1^\top \tilde{A}_1)^\dagger \tilde{A}_1^\top \tilde{A}_2$, we have

\[
(\text{RHS of (E.3)}) \iff \theta^{-1} \mu \tilde{B}_B^\top \tilde{B}_B - \mu \tilde{B}_B^\top \tilde{B}_B \succeq O. \quad (E.4)
\]

Since RHS of (E.4) holds due to $\theta^{-1} \geq 1$ and $B_B^\top B_B \succeq O$, $J_{\Psi_{\phi} \circ \Omega} \in \Gamma_{\Omega}(\mathbb{R}^n)$ has been proven.

\[\square\]

ORCID iDs

Isao Yamada @ https://orcid.org/0000-0002-6563-7526

References

[1] Abe J, Yamagishi M and Yamada I 2019 Convexity-edge-preserving signal recovery with linearly involved generalized minimax concave penalty function Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing pp 4918–22
[2] Arthur A 1969 Conditions for positive and nonnegative definiteness in terms of pseudoinverses SIAM J. Appl. Math. 17 434–40
[3] Bauschke H H and Combettes P L 2017 Convex Analysis and Monotone Operator Theory in Hilbert Spaces 2nd edn (New York: Springer)
[4] Bayram I 2015 Penalty functions derived from monotone mappings IEEE Signal Process. Lett. 22 265–9
[5] Bayram I 2016 On the convergence of the iterative shrinkage/thresholding algorithm with a weakly convex penalty IEEE Trans. Signal Process. 64 1597–608
[6] Beck A and Teboulle M 2010 Gradient-based algorithms with applications to signal-recovery problems Convex Optimization in Signal Processing and Communications ed D P Palomar and Y C Eldar (Cambridge: Cambridge University Press) pp 42–88
[7] Ben-Israel A and Greville T N E 2003 Generalized Inverses: Theory and Applications 2nd edn (New York: Springer)
[8] Bertero M and Boccacci P 1989 Introduction to Inverse Problems in Imaging (Bristol: Institute of Physics)
[9] Blake A and Zisserman A 1987 Visual Reconstruction (Cambridge, MA: MIT Press)
[10] Boyd S and Vandenberghe L 2004 Convex Optimization (Cambridge: Cambridge University Press)
[11] Byrne C L 2007 Applied Iterative Methods (Boca Raton, FL: CRC Press)
[12] Candès E J, Romberg J K and Tao T 2006 Stable signal recovery from incomplete and inaccurate measurements Commun. Pure Appl. Math. 59 1207–23
[13] Carlsson M 2018 On convex envelope and regularization of non-convex functionals without moving global minima J. Optim. Theory Appl. 183 66–84
[14] Cambasolle A 2004 An algorithm for total variation minimization and applications J. Math. Imaging Vis. 20 89–97
[15] Claerbout J and Muir F 1973 Robust modelling of elastic data Geophysics 38 826–44

\[\text{11} \quad \text{[2, theorem 1]} \text{ shows that for a block matrix}
\]

\[
S := \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix}
\]

\text{with } S_{11}, S_{22} \text{ symmetric, } S \succeq O \text{ if and only if } S_{11} \succeq O, S_{11} S_{12} = S_{12}, \text{ and } S_{22} - S_{21}^\top S_{11} S_{12} \succeq O. \]
[48] Ogura N and Yamada I 2002 Non-strictly convex minimization over the fixed point set of an asymptotically shrinking nonexpansive mapping Numer. Funct. Anal. Optim. 23 113–37
[49] Ono S and Yamada I 2015 Hierarchical convex optimization with primal-dual splitting IEEE Trans. Signal Process. 63 373–88
[50] Pierra G 1976 Méthodes de décomposition et croisement d’algorithmes pour des problèmes d’optimisation PhD Thesis University of Grenoble
[51] Pierra G 1984 Decomposition through formalization in a product space Math. Program. 28 96–115
[52] Rockafellar R T 1970 Convex Analysis (Princeton, NJ: Princeton University Press)
[53] Rockafellar R T 2009 Variational Analysis (New York: Springer)
[54] Rudin L I, Osher S and Fatemi E 1992 Nonlinear total variation based noise removal algorithms Physica D 60 259–68
[55] Santosa F and Symes W W 1986 Linear inversion of band limited reflection seismograms SIAM J. Sci. Comput. 7 1307–30
[56] Selesnick I 2017 Sparse regularization via convex analysis IEEE Trans. Signal Process. 65 4481–94
[57] Selesnick I and Farshchian M 2017 Sparse signal approximation via nonseparable regularization IEEE Trans. Signal Process. 65 2561–75
[58] Soubies E, Blanc-Féraud L and Aubert G 2015 A continuous exact L0 penalty (CEL0) for least squares regularized problem SIAM J. Imaging Sci. 8 1607–39
[59] Starck J L, Murtagh F and Fadili J 2015 Sparse Image and Signal Processing: Wavelets and Related Geometric Multiscale Analysis (Cambridge: Cambridge University Press)
[60] Taylor H, Bank S and McCoy J 1979 Deconvolution with the l1-norm Geophysics 44 39–52
[61] Theodoridis S 2015 Machine Learning—a Bayesian and Optimization Perspective (New York: Academic)
[62] Tibshirani R 1996 Regression shrinkage and selection via the lasso J. R. Stat. Soc. B 58 267–88
[63] Tikhonov A N 1963 Solution of incorrectly formulated problems and the regularization method Soviet Math. Dokl. 4 1035–8
[64] Tikhonov A N and Arsenin Y V 1977 Solutions of Ill-Posed Problems (New York: Winston)
[65] Vă B C 2013 A splitting algorithm for dual monotone inclusions involving cocoercive operators Adv. Comput. Math. 38 667–81
[66] Yamada I and Yamagishi M 2019 Global optimization of sum of convex and nonconvex functions with proximal splitting techniques Int. Conf. on Continuous Optimization
[67] Yamada I and Yamagishi M 2019 Hierarchical convex optimization by the hybrid steepest descent method with proximal splitting operators—Enhancements of SVM, Lasso Splitting Algorithm, Modern Operator Theory and Applications ed H H Bauschke et al (New York: Springer) pp 413–89
[68] Yamada I, Yukawa M and Yamagishi M 2011 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 ed H H Bauschke et al (New York: Springer) pp 345–90
[69] Yamagishi M and Yamada I 2017 Nonexpansiveness of a linearized augmented Lagrangian operator for hierarchical convex optimization Inverse Problems 33 35
[70] Yin L, Parekh A and Selesnick I 2019 Stable principal component pursuit via convex analysis IEEE Trans. Signal Process. 67 2595–607
[71] Zhang C H 2010 Nearly unbiased variable selection under minimax concave penalty Ann. Stat. 38 894–942
[72] Zhong D, Yi C, Xiao H, Zhang H and Wu A 2018 A novel fault diagnosis method for rolling bearing based on improved sparse regularization via convex optimization Complexity 2018 10