Hierarchical Convex Optimization by the Hybrid Steepest Descent Method with Proximal Splitting Operators — Enhancements of SVM and Lasso

Isao Yamada and Masao Yamagishi

Abstract The breakthrough ideas in the modern proximal splitting methodologies allow us to express the set of all minimizers of a superposition of multiple nonsmooth convex functions as the fixed point set of computable nonexpansive operators. In this paper, we present practical algorithmic strategies for the hierarchical convex optimization problems which require further strategic selection of a most desirable vector from the solution set of the standard convex optimization. The proposed algorithms are established by applying the hybrid steepest descent method to special nonexpansive operators designed through the art of proximal splitting. We also present applications of the proposed strategies to certain unexplored hierarchical enhancements of the support vector machine and the Lasso estimator.

Key words: convex optimization, proximal splitting algorithms, hybrid steepest descent method, Support Vector Machine (SVM), Lasso, TREX, signal processing, machine learning, statistical estimation
AMS 2010 Subject Classification: 49M20, 65K10, 90C30

1 Introduction

Convex optimization has been playing a central role in a broad range of mathematical sciences and engineering. Many optimization tasks in such applications can be interpreted as special instances of the following simple model:

$$\text{minimize } f(x) + g(Ax) \text{ subject to } x \in \mathscr{X}, \tag{1}$$

where $(\mathscr{X}, \langle\cdot,\cdot\rangle_{\mathscr{X}}, \|\cdot\|_{\mathscr{X}})$, $(\mathscr{K}, \langle\cdot,\cdot\rangle_{\mathscr{K}}, \|\cdot\|_{\mathscr{K}})$ are real Hilbert spaces, $f : \mathscr{X} \to (-\infty,\infty]$ and $g : \mathscr{K} \to (-\infty,\infty]$ are proper lower semicontinuous convex functions, i.e., $f \in \Gamma_0(\mathscr{X})$ and $g \in \Gamma_0(\mathscr{K})$, and $A : \mathscr{X} \to \mathscr{K}$ is a bounded linear operator, i.e., $A \in \mathscr{B}(\mathscr{X}, \mathscr{K})$. Such a unified simplification is indebted entirely to the remarkable expressive ability of the abstract Hilbert space. For example, a seemingly much more general model:

Isao Yamada
Department of Information and Communications Engineering, Tokyo Institute of Technology, S3-60, Ookayama, Meguro-ku, Tokyo 152-8552, Japan, e-mail: isao@ict.e.titech.ac.jp

Masao Yamagishi
Department of Information and Communications Engineering, Tokyo Institute of Technology, S3-60, Ookayama, Meguro-ku, Tokyo 152-8552, Japan, e-mail: myamagi@ict.e.titech.ac.jp

This manuscript was prepared before publication as Chapter 16 (pp. 413–489) in: H. H. Bauschke, R. S. Burachik, D. R. Luke (eds.) Splitting Algorithms, Modern Operator Theory, and Applications, Springer, 2019. This version is not completely same as the final version in the book. For precise information, we would like to encourage to consult the final version.

$$\text{find } x^\star \in \mathscr{S} := \operatorname*{argmin}_{x\in\mathscr{X}} \left[\Phi(x) := f(x) + \sum_{i=1}^{m} g_i(A_i x) \right] \neq \varnothing, \tag{2}$$

where $(\mathscr{X}, \langle\cdot,\cdot\rangle_{\mathscr{X}}, \|\cdot\|_{\mathscr{X}})$ and $(\mathscr{K}_i, \langle\cdot,\cdot\rangle_{\mathscr{K}_i}, \|\cdot\|_{\mathscr{K}_i})$ $(i = 1,2,\ldots,m)$ are real Hilbert spaces, $f \in \Gamma_0(\mathscr{X})$, $g_i \in \Gamma_0(\mathscr{K}_i)$ $(i = 1,2,\ldots,m)$, and $A_i \in \mathscr{B}(\mathscr{X},\mathscr{K}_i)$ $(i = 1,2,\ldots,m)$, can also be translated into the problem in (1) by redefining a new Hilbert space

$$\mathscr{K} := \mathscr{K}_1 \times \cdots \times \mathscr{K}_m = \{\mathbf{x} = (x_1,\ldots,x_m) \mid x_i \in \mathscr{K}_i \ (i = 1,\ldots,m)\} \tag{3}$$

equipped with the addition $(\mathbf{x},\mathbf{y}) \mapsto (x_1 + y_1,\ldots,x_m + y_m)$, the scalar multiplication $(\alpha,\mathbf{x}) \mapsto (\alpha x_1,\ldots,\alpha x_m)$, and the inner product $(\mathbf{x},\mathbf{y}) \mapsto \langle \mathbf{x},\mathbf{y}\rangle_{\mathscr{K}} := \sum_{i=1}^{m} \langle x_i, y_i\rangle_{\mathscr{K}_i}$, a new convex function

$$g := \bigoplus_{i=1}^{m} g_i : \mathscr{K} \to (-\infty,\infty] : (x_1,\ldots,x_m) \mapsto \sum_{i=1}^{m} g_i(x_i), \tag{4}$$

and a new bounded linear operator

$$A : \mathscr{X} \to \mathscr{K} : x \mapsto (A_1 x,\ldots,A_m x). \tag{5}$$

Indeed, for many years, the model (2) has been accepted widely as a standard, where all players, $f, g_i \circ A_i \in \Gamma_0(\mathscr{X})$ $(i = 1,\ldots,m)$ in (2) are designed strategically by users in order to achieve, after optimization, a valuable vector satisfying their requirements.

The so-called *proximal splitting methodology* has been built, on the rich mathematical foundations of convex analysis, monotone operator theory and fixed point theory of nonexpansive operators (see, e.g., [9, 45, 47, 139]), in order to broaden the applicability of the proximity operators of convex functions [101], e.g., to the model (2). It is well-known that the solution set $\mathscr{S}$ in (2) can be characterized completely as the zero of the set-valued operator $\partial\Phi : \mathscr{X} \to 2^{\mathscr{X}} : x \mapsto \{u \in \mathscr{X} \mid \Phi(y) \geq \Phi(x) + \langle y - x, u\rangle_{\mathscr{X}} \ (\forall y \in \mathscr{X})\}$, called the *subdifferential* of Φ. The maximal monotonicity of $\partial\Phi$ provides us with further equivalent fixed point characterization in terms of a single valued operator $\operatorname{prox}_{\Phi} := (\mathrm{I} + \partial\Phi)^{-1} : \mathscr{X} \to \mathscr{X} : x \mapsto \operatorname{argmin}_{y\in\mathscr{X}} \Phi(y) + \frac{1}{2}\|x - y\|_{\mathscr{X}}^2$ called the proximity operator of Φ (see Section 2.2) [Note: The identity operator is denoted by I: $\mathscr{X} \to \mathscr{X}$ but the common notation I is going to be used for the identity operator on any real Hilbert space in this paper]. This fact is simply stated as

$$(\forall z \in \mathscr{X}) \quad z \in \mathscr{S} \Leftrightarrow 0 \in \partial\Phi(z) \Leftrightarrow z = \operatorname{prox}_{\Phi}(z)$$

and some algorithms can generate, from any $x_0 \in \mathscr{X}$, a weakly convergent sequence $x_n \in \mathscr{X}$ $(n \in \mathbb{N})$ to a point in $\mathscr{S}$ in (2) if $\operatorname{prox}_{\Phi}$ is available as a computational tool (see, e.g., [98, 99, 121]). A simplest example of such algorithms generates a sequence $(x_n)_{n\in\mathbb{N}}$ by

$$x_{n+1} = \operatorname{prox}_{\Phi}(x_n) \quad (n = 0,1,2,\ldots). \tag{6}$$

The algorithm (6) can be interpreted as a straightforward application of *Krasnosel'skiĭ-Mann Iterative Process* (see Fact 6 in Section 2.2) because $\operatorname{prox}_{\Phi}$ is known to be firmly nonexpansive, i.e., $2\operatorname{prox}_{\Phi} - \mathrm{I} : \mathscr{X} \to \mathscr{X}$ is a nonexpansive operator (see (21)). Although the above strategy in (6) is conceptually simple and elegant, its applicability has been very limited because the computation of $\operatorname{prox}_{\Phi}(x)$ requires to solve a regularized convex optimization problem $\min \Phi(\cdot) + \frac{1}{2}\|x - \cdot\|_{\mathscr{X}}^2$ whose unique solution is still hard to be computed for most scenarios of type (2) in many application areas.

On the other hand, there are many scenarios that fall in the model (2) where the proximity operators of the all players, i.e., $\operatorname{prox}_f : \mathscr{X} \to \mathscr{X}$ and $\operatorname{prox}_{g_i} : \mathscr{K}_i \to \mathscr{K}_i$ $(i = 1,\ldots,m)$, are available as computational tools while $\operatorname{prox}_{\Phi}$ is not practically available (see, e.g., [42, 45]). A major goal of recent active studies (see, e.g., [40, 44, 47, 139, 150]) on the *proximal splitting methodology* has been the creation of more applicable iterative algorithms, for (2) and its variations, than (6) by utilizing computable tools prox_f and $\operatorname{prox}_{g_i}$ $(i = 1,\ldots,m)$

simultaneously. Such effort has culminated in many powerful algorithms which have been applied successfully to the broader classes of optimizations including the standard model (2).

Usually, the standard model (2) is formulated in the form of a weighted average of multiple convex functions and the weights are designed in accordance with the level of importance of each convex function. However quantification of the level of importance is often challenging as well as influential to the final results of optimizations (see Section 5.1 for a recent advanced strategy of such a parameter tuning for the Lasso estimator which is a standard sparsity aware statistical estimation method). By keeping in mind (i) the remarkable flexibility of the standard model (2) proven extensively in many successful applications of the modern proximal splitting methodology, as well as (ii) the inherent difficulty in the weight design of multiple convex functions in (2), a question arises: *Is there any alternative model of (2) which can also serve as a natural optimization strategy for multiple convex criteria* ? To see the light of the tunnel regarding this primitive question, let us start to imagine important elements for us to consider in *finding residence*. We may consider the house rent, the residential environment including living space and housing equipment, the neighborhood environment, the accessibility to public transportation systems, and the commuting time, etc. We would prioritize the elements, e.g., firstly by narrowing down the candidates to the set S_1 of all residents of which the rents and commuting times are in your acceptable range. Next we may try to narrow down the candidates to the set $S_2(\subset S_1)$ of all residents whose living spaces achieve maximum level among all in S_1. Further, we may probably like to select residents in S_2 as final choices by choosing the best ones, e.g., in the sense of the neighborhood environment or the housing equipment. This simple example suggests that we often optimize multiple criteria one by one hierarchically rather than optimize the sum of different criteria at once certainly because there exists no universal justification for adding different criteria. In fact, many mathematicians and scientists have been challenging to pave the way for the so-called hierarchical convex optimization problems (see, e.g., [3, 8, 18, 32, 33, 36, 46, 55, 95, 107, 114, 133, 142, 146–149]). Landmark theories toward $\mathscr{M}$-stage hierarchical convex optimization are found, e.g., in [3, 18] where, for given $\Phi_i \in \Gamma_0(\mathscr{X})$ $(i = 0,1,\ldots,\mathscr{M})$ satisfying $S_i := \underset{x\in S_{i-1}}{\operatorname{argmin}} \Phi_i(x) \neq \varnothing$ $(i = 0,1,\ldots,\mathscr{M})$ with $S_{-1} := \mathscr{X}$, their major goals are set to establish computational strategies for iterative approximation of a point in $S_{\mathscr{M}}$. (Note: Every point in S_i of the hierarchical convex optimization is called a *viscosity solution* of S_{i-1} $(i = 1,2,\ldots,\mathscr{M})$. To avoid confusion with "bilevel optimization" in the sense of [19, 30, 138], we do not use the designation *bilevel optimization* for S_1 in our hierarchical convex optimization). Under the assumptions that $\dim(\mathscr{X}) < \infty$ and that $\Phi_i \in \Gamma_0(\mathscr{X})$ $(i = 1,2,\ldots,\mathscr{M})$ are real valued, Cabot [18] showed that the sequence $(x_n)_{n\in\mathbb{N}}$ defined by

$$x_{n+1} := \operatorname{prox}_{\left(\Phi_0+\varepsilon_n^{(1)}\Phi_1+\varepsilon_n^{(2)}\Phi_2+\cdots+\varepsilon_n^{(\mathscr{M})}\Phi_{\mathscr{M}}\right)}(x_n) = \left(\mathrm{I}+\partial\left(\Phi_0+\varepsilon_n^{(1)}\Phi_1+\varepsilon_n^{(2)}\Phi_2+\cdots+\varepsilon_n^{(\mathscr{M})}\Phi_{\mathscr{M}}\right)\right)^{-1}(x_n) \quad (7)$$

satisfies (i) $\lim_{n\to\infty} d(x_n, S_{\mathscr{M}}) = 0$ and (ii) $(\forall i \in \{0,1,\ldots,\mathscr{M}\})$ $\lim_{n\to\infty}\Phi_i(x_n) = \min_{x\in S_{i-1}}\Phi_i(x)$ if positive number sequences $(\varepsilon_n^{(0)} := 1)_{n\in\mathbb{N}}$ and $(\varepsilon_n^{(i)})_{n\in\mathbb{N}}$ $(i \in \{1,\ldots,\mathscr{M}\})$ satisfy certain technical conditions including $\lim\limits_{n\to\infty}\varepsilon_n^{(i)} = 0$, $\lim\limits_{n\to\infty}\dfrac{\varepsilon_n^{(i)}}{\varepsilon_n^{(i-1)}} = 0$ $(i = 1,2,\ldots,\mathscr{M})$ and $\sum_{n=0}^{\infty}\varepsilon_n^{(\mathscr{M})} = \infty$ [Note: The scheme (7) is a simplified version of the original scheme in [18] by restricting to the case $\lambda_n = 1$ and $\eta_n = 0$ $(n \in \mathbb{N})$].

Clearly, the algorithms (7) and (6) have essentially a common limitation in their practical applicabilities because (7) requires $\operatorname{prox}_{\left(\Phi_0+\varepsilon_n^{(1)}\Phi_1+\varepsilon_n^{(2)}\Phi_2+\cdots+\varepsilon_n^{(\mathscr{M})}\Phi_{\mathscr{M}}\right)}$, or its very good approximation, for every update in generation of $(x_n)_{n\in\mathbb{N}}$. By recalling the breakthrough ideas developed in the recent *proximal splitting methodology* for resolution of the inherent limitation in (6), an ideal as well as possibly realistic assumption to be imposed upon each player $\Phi_i : \mathscr{X} \to (-\infty,\infty]$ $(i = 0,1,\ldots,\mathscr{M})$ in the above $\mathscr{M}$-stage hierarchical convex optimization seems to be certain differentiability assumptions or proximal decomposability assumptions, e.g.

$$\Phi_i(x) := f_i(x) + \sum_{\iota_{(i)}=1}^{M_i} g_{\iota_{(i)}}(A_{\iota_{(i)}} x),$$

with real Hilbert spaces $(\mathscr{K}_{\iota_{(i)}}, \langle \cdot,\cdot \rangle_{\mathscr{K}_{\iota_{(i)}}}, \|\cdot\|_{\mathscr{K}_{\iota_{(i)}}})$ $(\iota_{(i)} = 1,2,\ldots,M_i)$, $f_i \in \Gamma_0(\mathscr{X})$, $g_{\iota_{(i)}} \in \Gamma_0(\mathscr{K}_i)$ $(i = 0,1,\ldots,\mathscr{M})$, and bounded linear operators $A_{\iota_{(i)}} : \mathscr{X} \to \mathscr{K}_{\iota_{(i)}}$ $(\iota_{(i)} = 1,2,\ldots,M_i)$, where $\operatorname{prox}_{f_i} : \mathscr{X} \to \mathscr{X}$ and $\operatorname{prox}_{g_{\iota_{(i)}}} : \mathscr{K}_{\iota_{(i)}} \to \mathscr{K}_{\iota_{(i)}}$ $(\iota_{(i)} = 1,\ldots,M_i)$, are available as computational tools while $\operatorname{prox}_{\Phi_i}$ is not necessarily available.

In this paper, we choose to cast our primary target in the iterative approximation of a solution of

$$\text{minimize } \Psi(x^\star) \text{ subject to } x^\star \in \operatorname*{argmin}_{x\in\mathscr{X}} \left[\Phi(x) := f(x) + \sum_{i=1}^{m} g_i(A_i x) \right] \neq \varnothing, \tag{8}$$

i.e., a viscosity solution of the convex optimization problem (2), where we assume that $\Psi \in \Gamma_0(\mathscr{X})$ is Gâteaux differentiable with Lipschitzian gradient $\nabla\Psi : \mathscr{X} \to \mathscr{X}$, i.e.,

$$(\exists \kappa > 0,\ \forall x,y \in \mathscr{X}) \quad \|\nabla\Psi(x) - \nabla\Psi(y)\| \leq \kappa\|x-y\|,$$

and that $\operatorname{prox}_f : \mathscr{X} \to \mathscr{X}$ and $\operatorname{prox}_{g_i} : \mathscr{K}_i \to \mathscr{K}_i$ $(i = 1,\ldots,m)$ are available as computational tools.

Although the application of such iterative algorithms is certainly restrictive compared to the overwhelming potential of the general hierarchical convex optimization, our target is realistic and still allows us to cover many applications of interest to practitioners who are searching for a step ahead optimization strategy and yet to be able to exploit maximally the central ideas in the modern proximal splitting methodologies. Especially for practitioners, we remark that if the suppression of $\sum_{k=1}^{L} \psi_k \circ B_k \in \Gamma_0(\mathscr{X})$ over $\operatorname{argmin}_{x\in\mathscr{X}} \Phi(x)$ is required, where, for each $k \in \{1,2,\ldots,L\}$, $\mathscr{Y}_k$ is a real Hilbert space, $\psi_k \in \Gamma_0(\mathscr{Y}_k)$, $\operatorname{prox}_{\gamma\psi_k} : \mathscr{Y}_k \to \mathscr{Y}_k$ $(\gamma > 0)$ is available as computational tools, and $B_k \in \mathscr{B}(\mathscr{X},\mathscr{Y}_k)$, such a mission could be achieved satisfactorily by considering an alternative problem below of type (8):

$$\begin{aligned} &\text{minimize } \Psi(x^\star) := \sum_{k=1}^{L} {}^{\gamma}\psi_k(B_k x^\star) \\ &\text{subject to } x^\star \in \operatorname*{argmin}_{x\in\mathscr{X}} \left[\Phi(x) := f(x) + \sum_{i=1}^{m} g_i(A_i x) \right] \neq \varnothing, \end{aligned} \tag{9}$$

where (i) ${}^{\gamma}\psi_k : \mathscr{Y}_k \to \mathbb{R} : y_k \mapsto \min_{y\in\mathscr{Y}_k} \psi_k(y) + \frac{1}{2\gamma}\|y - y_k\|^2_{\mathscr{Y}_k}$ $(k = 1,2,\ldots,L)$ are *the Moreau envelopes (or the Moreau-Yosida regularizations)* of a sufficiently small index $\gamma > 0$ (see Fact 8 in Section 2.2 and [149]). This is because $\lim_{\gamma\downarrow 0} {}^{\gamma}\psi_k(y_k) = \psi_k(y_k)$ $(\forall y_k \in \operatorname{dom}\psi_k := \{y \in \mathscr{Y}_k \mid \psi_k(y) < \infty\})$ and ${}^{\gamma}\psi_k$ is Gâteaux differentiable with $\frac{1}{\gamma}$-Lipschitzian $\nabla^{\gamma}\psi_k : \mathscr{Y}_k \to \mathscr{Y}_k : y_k \mapsto \frac{y_k - \mathrm{prox}_{\gamma\psi_\mathrm{k}}(\mathrm{y_k})}{\gamma}$ and therefore

$$\begin{aligned} &(\forall x_1,x_2 \in \mathscr{X}) \quad \left\| \nabla \sum_{k=1}^{L} ({}^{\gamma}\psi_k \circ B_k)(x_1) - \nabla \sum_{k=1}^{L} ({}^{\gamma}\psi_k \circ B_k)(x_2) \right\|_{\mathscr{X}} \\ &= \left\| \sum_{k=1}^{L} B_k^* \nabla^{\gamma}\psi_k(B_k x_1) - \sum_{k=1}^{L} B_k^* \nabla^{\gamma}\psi_k(B_k x_2) \right\|_{\mathscr{X}} \leq \sum_{k=1}^{L} \frac{\|B_k\|^2_{\mathrm{op}}}{\gamma} \|x_1 - x_2\|_{\mathscr{X}}, \end{aligned}$$

where $B_k^* \in \mathscr{B}(\mathscr{Y}_k,\mathscr{X})$ is the conjugate of $B_k \in \mathscr{B}_k(\mathscr{X},\mathscr{Y}_k)$ and $\|\cdot\|_{\mathrm{op}}$ stands for the operator norm.

Fortunately, by introducing the exactly same translation used in the reformulation of Problem (2) as an instance of Problem (1), our problem (8) can also be simplified as

$$\text{minimize } \Psi(x^\star) \text{ subject to } x^\star \in \mathscr{S}_p := \operatorname*{argmin}_{x\in\mathscr{X}} [f(x) + g(Ax)] \neq \varnothing, \tag{10}$$

where $\mathscr{K}$, $g : \mathscr{K} \to (-\infty,\infty]$, and $A : \mathscr{X} \to \mathscr{K}$ are defined respectively[1] by (3), (4), and (5), and we can assume that (i) $\Psi \in \Gamma_0(\mathscr{X})$ is Gâteaux differentiable with Lipschitzian gradient $\nabla\Psi : \mathscr{X} \to \mathscr{X}$, and that (ii) $\mathrm{prox}_f : \mathscr{X} \to \mathscr{X}$ and $\mathrm{prox}_g : \mathscr{K} \to \mathscr{K}$ are available as computational tools because

$$\begin{aligned}\mathrm{prox}_g(\mathbf{x}) &:= \operatorname*{argmin}_{\mathbf{y}\in\mathscr{K}} \left[g(\mathbf{y}) + \frac{1}{2}\|\mathbf{y}-\mathbf{x}\|^2_{\mathscr{K}} \right] \\ &= \operatorname*{argmin}_{(y_1,\dots,y_m)\in\mathscr{K}_1\times\cdots\times\mathscr{K}_m} \sum_{i=1}^{m} \left[g_i(y_i) + \frac{1}{2}\|y_i - x_i\|^2_{\mathscr{K}_i} \right] \\ &= \left(\mathrm{prox}_{g_1}(x_1),\dots,\mathrm{prox}_{g_m}(x_m)\right). \qquad (11)\end{aligned}$$

The following two scenarios suggest the remarkable advantage achieved by algorithmic solutions to (8).

Scenario 1 *(Unification of Conditional Optimization Models)* Let $f_{\langle\mathscr{D}\rangle} \in \Gamma_0(\mathscr{X})$ and $g_{i\langle\mathscr{D}\rangle} \in \Gamma_0(\mathscr{K}_i)$ $(i = 1,2,\dots,m)$ be nonnegative valued functions which are defined with observed data $\mathscr{D}$. Suppose that there exists a well-established data analytic strategy which utilizes with $\Psi \in \Gamma_0(\mathscr{X})$ as

$$\begin{aligned}&\text{find } x^\star \in \operatorname*{argmin}_{x\in\mathscr{S}_0} \Psi(x), \\ &\text{where } \mathscr{S}_0 := \left\{ x \in \mathscr{X} \mid f_{\langle\mathscr{D}\rangle}(x) = g_{i\langle\mathscr{D}\rangle}(A_i x) = 0 \ (i = 1,2,\dots,m) \right\}, \qquad (12)\end{aligned}$$

provided that the data $\mathscr{D}$ is consistent, i.e., it satisfies $\mathscr{S}_0 \neq \varnothing$.

However, to deal with more general data $\mathscr{D}$, it is important to establish a mathematically sound extension of the above data analytic strategy to be applicable even to inconsistent data $\mathscr{D}$ s.t. $\mathscr{S}_0 = \varnothing$. One of the most natural extensions of (12) would be the following hierarchical formulation:

$$\begin{aligned}&\text{find } x^{\star\star} \in \operatorname*{argmin}_{x^\star\in\mathscr{S}_{\langle\mathscr{D}\rangle}} \Psi(x^\star), \\ &\text{where } \mathscr{S}_{\langle\mathscr{D}\rangle} := \operatorname*{argmin}_{x\in\mathscr{X}} \left[f_{\langle\mathscr{D}\rangle}(x) + \sum_{i=1}^{m} g_{i\langle\mathscr{D}\rangle}(A_i x) \right],\end{aligned}$$

because $\mathscr{S}_{\langle\mathscr{D}\rangle} \neq \varnothing$ holds under weaker assumption than $\mathscr{S}_0 \neq \varnothing$, and $\mathscr{S}_{\langle\mathscr{D}\rangle} = \mathscr{S}_0$ holds true if $\mathscr{S}_0 \neq \varnothing$. However, the well-established data analytic strategies only for consistent data $\mathscr{D}$ in the form of (12) have often been modified, with the so-called *tuning parameter* $\mathfrak{C} > 0$, to

$$\text{find } \tilde{x}^\star \in \operatorname*{argmin}_{x\in\mathscr{X}} \left[\frac{1}{\mathfrak{C}}\Psi(x) + f_{\langle\mathscr{D}\rangle}(x) + \sum_{i=1}^{m} g_{i\langle\mathscr{D}\rangle}(A_i x) \right], \qquad (13)$$

which is not really an extension of (12) because the model (13) unfortunately has no guarantee to produce $x^\star$ in (12) even if $\mathscr{D}$ satisfies $\mathscr{S}_0 \neq \varnothing$.

Scenario 2 Suppose that we are interested in an estimation problem of a desired vector in $\mathscr{X}$ at which the functions $f, g_i \circ A_i \in \Gamma_0(\mathscr{X})$ $(i = 1,\dots,m)$ in (2) are known to achieve small values and therefore the model (2) has been employed as an estimation strategy. Suppose also that we newly found another effective criterion $\Psi \in \Gamma_0(\mathscr{X})$ which likely to achieve small values around the desired vector to be estimated. In such a case, our common utilization of Ψ, for improvement of the previous strategy, has often been modeled as a new optimization problem:

$$\text{find } \tilde{x}^\star \in \widetilde{\mathscr{S}} := \operatorname*{argmin}_{x\in\mathscr{X}} \left[f(x) + \sum_{i=1}^{m} g_i(A_i x) + \Psi(x) \right] \neq \varnothing. \qquad (14)$$

[1] There are many practical conditions for (f,g,A) to guarantee $\mathscr{S}_p \neq \varnothing$, see, e.g., [9, 153] and Fact 2 in Section 2.1.

However, it is essentially hard to tell which is better between the estimation strategies (2) and (14) because the criteria in these optimizations are different. Indeed, $\tilde{x}^{\star}$ does not necessarily achieve best in the sense of the model (2) while $x^{\star}$ certainly achieves best in the sense of the model (2). On the other hand, if we formulate a new optimization problem, from a hierarchical optimization point of view, e.g., as

$$\text{find } x^{\star\star} \in \operatorname*{argmin}_{x^{\star} \in \mathscr{S}} \Psi(x^{\star}), \text{ where } \mathscr{S} := \operatorname*{argmin}_{x \in \mathscr{X}} \left[f(x) + \sum_{i=1}^{m} g_i(A_i x) \right] \neq \varnothing, \tag{15}$$

its solution $x^{\star\star}$ certainly meets more faithfully all the requirements than $x^{\star} \in \mathscr{S}$ because both $x^{\star\star}, x^{\star} \in \mathscr{S}$ and $\Psi(x^{\star\star}) \leq \Psi(x^{\star})$ are achieved.

The following examples suggest that the hierarchical optimization has been offering well-grounded direction for advancement of computational strategies in inverse problems and data sciences.

Example 1 *(Hierarchical Convex Optimizations in Real World Applications*[2]*)*

(a) *(Generalized inverse / Moore-Penrose inverse [9, 12, 100, 111, 118]) Let $\mathscr{X}$ and $\mathscr{K}$ be real Hilbert spaces, let $A \in \mathscr{B}(\mathscr{X},\mathscr{K})$ be such that* $\operatorname{ran}(A) := \{A(x) \in \mathscr{K} \mid x \in \mathscr{X}\}$ *is closed. Then for every $y \in \mathscr{K}$, $C_y := \{x \in \mathscr{X} \mid \|Ax-y\|_{\mathscr{K}} = \min_{z\in\mathscr{X}} \|Az-y\|_{\mathscr{K}}\} = \{x \in \mathscr{X} \mid A^*A(x) = A^*(y)\} \neq \varnothing$. The generalized inverse (in the sense of Moore-Penrose) $A^\dagger \in \mathscr{B}(\mathscr{K},\mathscr{X})$ is defined as $A^\dagger : \mathscr{K} \to \mathscr{X} : y \mapsto P_{C_y}(0)$,where P_{C_y} is the orthogonal projection onto C_y. $A^\dagger(y)$ can be seen as the unique solution to the hierarchical convex optimization problem (15) for $f(z) := \|A(z)-y\|_{\mathscr{K}}$, $g_i(z) := 0$ ($i = 1,2,\ldots,m$) and $\Psi(z) = \frac{1}{2}\|z\|^2_{\mathscr{X}}$. The* Moore-Penrose inverse *$A^\dagger \in \mathscr{B}(\mathscr{K},\mathscr{X})$ of $A \in \mathscr{B}(\mathscr{X},\mathscr{K})$ has been serving as one of the most natural generalizations of the inverse of A, typically in Scenario 1, under the situations where the existence of $A^{-1} \in \mathscr{B}(\mathscr{K},\mathscr{X})$ is not guaranteed. In particular, for finite dimensional settings, there are many ways to express $A^\dagger$. These include the singular value decomposition of $A^\dagger$ in terms of the singular value decomposition of A.*

(b) *(Tikhonov approximation [3, 9, 55, 133]) Let $\Psi, f \in \Gamma_0(\mathscr{X})$ and* $\operatorname{argmin}(f) \cap \operatorname{dom}(\Psi) \neq \varnothing$ *where Ψ is coercive and strictly convex. Then Ψ admits a unique minimizer x_0 over* $\operatorname{argmin}(f)$*. This x_0 can be seen as the solution of the hierarchical convex optimization in (15) for $g_i(z) := 0$ ($i = 1,2,\ldots,m$). Moreover, if we define $x_\varepsilon \in \mathscr{X}$ as the unique minimizer of the regularized problem*

$$\text{miminize } f(x) + \varepsilon\Psi(x) \text{ subject to } x \in \mathscr{X} \tag{16}$$

for every $\varepsilon > 0$, the desired x_0 can be approximated as (i) $x_\varepsilon \rightharpoonup x_0$ (as $\varepsilon \downarrow 0$) and (ii) $\Psi(x_\varepsilon) \to \Psi(x_0)$ (as $\varepsilon \downarrow 0$). This fact suggests a strategy for approximating x_0 if we have a practical way of computing x_{ε_n} for positive sequence $(\varepsilon_n)_{n=1}^\infty$ satisfying $\varepsilon_n \downarrow 0$ (as $n \to \infty$). Many computational approaches to the hierarchical convex optimization seem to have been designed along this strategy.[3] *We remark that many formulations of type* (13) *in Scenario 1 can be seen as instances of* (16) *with $\varepsilon = \frac{1}{\mathfrak{C}}$. However, in general, the hierarchical optimality can never be guaranteed by the solution of* (16) *for a fixed constant $\varepsilon > 0$.*

(c) *Assuming differentiability, the iteration of (7) for $\mathscr{M} = 1$ can also be interpreted as an implicit discretization of the continuous dynamical system:*

$$\dot{x}(t) + \nabla\Phi_0(x(t)) + \varepsilon(t)\nabla\Phi_1(x(t)) = 0, \quad t \geq 0, \tag{17}$$

where $\varepsilon : \mathbb{R}_+ \to \mathbb{R}$ is a control parameter tending to 0 when $t \to \infty$. This observation has been motivating explicit discretization of (17) for iterative approximation of point in S_1, e.g. by

$$x_{n+1} \in x_n + \lambda_n \partial(\Phi_0 + \varepsilon_n \Phi_0)(x_n),$$

and its variations (see, e.g., [78, 79, 127, 128]), where λ_n is a nonnegative stepsize. However this class of algorithms cannot exploit recent advanced proximal splitting techniques for dealing with the constrained set S_0.

(d) *Under the assumption that (i) Ψ is* Gâteaux *differentiable with Lipschitzian gradient $\nabla\Psi : \mathscr{X} \to \mathscr{X}$, and (ii)* $\operatorname{prox}_f : \mathscr{X} \to \mathscr{X}$ *is available as a computable tool, the* inertial forward-backward algorithm with vanishing Tikhonov regularization *was proposed [4], along in the frame of accelerated forward-backward methods*[4]*,*

[2] To the best of the authors' knowledge, little has been reported on the hierarchical *nonconvex* optimization. We remark that the MV-PURE (Minimum-variance pseudo-unbiased reduced-rank estimator) (see, e.g., [112, 113, 144]), for the unknown vector possibly subjected to linear constraints, is defined by a closed form solution of a certain hierarchical nonconvex optimization problem which characterizes a natural reduced rank extension of the Gauss-Markov (BLUE) estimator [85, 93] to the case of reduced-rank estimator. It was shown in [113] that specializations of the MV-PURE include Marquardt's reduced rank estimator [97], Chipman-Rao estimator [29], and Chipman's reduced rank estimator [28]. In Section 5.2 of this paper, we newly present a special instance of a hierarchical *nonconvex* optimization problem which can be solved through multiple hierarchical *convex* optimization subproblems.

[3] The behavior of $(x_\varepsilon)_{\varepsilon\in(0,1)} \subset \mathscr{X}$ can be analyzed in the context of *approximating curve* for monotone inclusion problem. For recent results combined with Yosida regularization, see [37].

[4] See [4] on the stream of research, to name but a few, [11, 24], originated from Nesterov's seminal paper [103].

for an iterative approximation of the solution of a hierarchical convex optimization in (15) for $g_i = 0$ ($i = 1,2,\ldots,m$).

(e) *In general, the convex optimization problems, especially in the convex feasibility problems [7, 22, 31], have infinitely many solutions that could be considerably different in terms of other criteria. However most iterative algorithms for convex optimization can approximate an anonymous solution of the problem. For pursuing a better solution in some other aspects,* superiorization *[21, 80, 104, 110] introduces proactively designed perturbations into the original algorithms with preserving preferable convergence properties. Essentially, by adopting another criterion Ψ, these methods aim to lower the value of Ψ with incorporating a perturbation involving the descent direction of Ψ. Apparently, as reported in [150], the hierarchical convex optimization can serve as one of the ideal formulations for the superiorization.*

(f) *Let $\Psi \in \Gamma_0(\mathscr{X})$ be* Gâteaux *differentiable and its gradient $\nabla\Psi : \mathscr{X} \to \mathscr{X}$ is Lipschitzian. Suppose that $f \in \Gamma_0(\mathscr{X})$ is also* Gâteaux *differentiable with Lipschitzian gradient $\nabla f : \mathscr{X} \to \mathscr{X}$ and admits* $\operatorname{argmin}(f + \iota_K) \neq \varnothing$ *for a nonempty closed convex set $K \subset \mathscr{X}$, where ι_K is the indicator function , i.e.,*

$$\iota_K(x) := \begin{cases} 0 & \textit{if } x \in K, \\ \infty & \textit{otherwise.} \end{cases}$$

Then

$$\text{minimize } \Psi(x) \text{ subject to } x \in \operatorname{argmin}(f + \iota_K) \tag{18}$$

can be seen as an instance of the hierarchical convex optimization in (15) for $g_1 := \iota_K$ and $g_i = 0$ ($i = 2,3,\ldots,m$). By applying the hybrid steepest descent method [52, 141, 142, 146–148] to several expressions of the set $\operatorname{argmin}(f + \iota_K)$ *as the fixed point set of certain computable nonexpansive operators $T : \mathscr{X} \to \mathscr{X}$ (see, e.g., [146, Proposition 2.5], [149, Example 17.6(b)]), practical algorithms have been established to produce a sequence $x_n \in \mathscr{X}$ ($n = 0,1,2,\ldots$) which is guaranteed to converge to a solution to Problem (18). These cover a version of Projected Landweber method [63, 115, 123] for $\Psi(x) := \frac{1}{2}\|x\|^2_{\mathscr{X}}$ and $f(x) := \|A(x) - b\|_{\mathscr{K}}$, where $A \in \mathscr{B}(\mathscr{X},\mathscr{K})$ and the metric projection $P_K : \mathscr{X} \to K$ is assumed available as a computational tool. As will be discussed below, the main idea of the present paper specialized for Problem (10) (or equivalently Problem (8)) is along this simple hierarchical optimization strategy [86, 107, 146, 149, 150] of applying the hybrid steepest descent method (HSDM: see Section 2.4) to the precise expressions of the solution sets of the convex optimization problems in terms of fixed point sets of computable nonexpansive operators defined on a certain real Hilbert space $\mathscr{H}$ which is not necessarily same as the original Hilbert space $\mathscr{X}$.*

Apparently, to tackle Problem (10) (or equivalently Problem (8)), we need to exploit full information on $\mathscr{S}_p$ which is an infinite set in general. Moreover, even by using the recently developed powerful proximal splitting algorithms, specially designed for (1), we can produce only some vector sequence that converges to just an anonymous point in $\mathscr{S}_p$, which implies that we need to add further a new twist to the well-known strategies applicable to Problem (1).

Fortunately, the unified perspective from the viewpoint of convex analysis and monotone operator theory (see, e.g., [9]) often enables us to enjoy notable characterizations of the solution set $\mathscr{S}_p$ in terms of the set of all fixed points of a computable nonexpansive operator defined on certain real Hilbert spaces. Indeed, almost all existing proximal splitting algorithms for Problem (1) more or less rely on the following type of characterizations of $\mathscr{S}_p$:

$$\mathscr{S}_p = \operatorname*{argmin}_{x \in \mathscr{X}} f(x) + g(Ax) = \Xi\left(\operatorname{Fix}(T)\right) := \bigcup_{z \in \operatorname{Fix}(T)} \Xi(z) \subset \mathscr{X}, \tag{19}$$

$$\operatorname{Fix}(T) := \{z \in \mathscr{H} \mid T(z) = z\} \qquad \text{(Fixed point set of } T\text{)}, \tag{20}$$

where $(\mathscr{H}, \langle\cdot,\cdot\rangle_{\mathscr{H}}, \|\cdot\|_{\mathscr{H}})$ is a certain real Hilbert space (not necessarily $\mathscr{H} = \mathscr{X}$), $T : \mathscr{H} \to \mathscr{H}$ is a computable nonexpansive operator, i.e., an operator satisfying

$$(\forall z_1, z_2 \in \mathscr{H}) \quad \|T(z_1) - T(z_2)\|_{\mathscr{H}} \leq \|z_1 - z_2\|_{\mathscr{H}}, \tag{21}$$

and $\Xi : \mathscr{H} \to 2^{\mathscr{X}}$ is a certain set valued operator. Examples of such characterizations are found in [62, 150] for the augmented Lagrangian method [81, 116], in [44, 45] for the forward-backward splitting approach [66, 109, 134], in [40, Proposition 18(iii)] for the Douglas-Rachford splitting approach (see Section 2.3) [91], in [61, 150] for the alternating direction method of multipliers (ADMM) [66, 76, 91], in [47, 139] for the primal-dual splitting method, and in [150] for a generalized version (see Section 2.3) of the linearized augmented Lagrangian method [151].

If we find a computable nonexpansive operator T satisfying (19) as well as a computationally tractable way to extract a point in $\Xi(z)(\subset \mathscr{X})$ for a given $z \in \mathrm{Fix}(T)$, we can realize an algorithmic solution to Problem (1) by applying the so-called *Krasnosel'skiĭ-Mann Iterative Process* (see Fact 6 in Section 2.2) to T, and can produce a weak convergent sequence to a fixed point $z \in \mathrm{Fix}(T)$, followed by a point extraction from $\Xi(z)$. Indeed, the powerful proximal splitting methodologies for Problem (2) seem to have been built more or less along this strategy through innovative designs of computable nonexpansive operators by using $\mathrm{prox}_f : \mathscr{X} \to \mathscr{X}$ and $\mathrm{prox}_{g_i} : \mathscr{K}_i \to \mathscr{K}_i$ $(i = 1, \ldots, m)$ as computational tools.

On the other hand, every nonexpansive operator $T : \mathscr{H} \to \mathscr{H}$ can also be plugged into the hybrid steepest descent method for minimizing $\Theta \in \Gamma_0(\mathscr{H})$, whose gradient $\nabla\Theta : \mathscr{H} \to \mathscr{H}$ is Lipschitz continuous, over the fixed point set $\mathrm{Fix}(T) \neq \varnothing$ (see Section2.4). [5] Moreover, for Problem (1), if such a computable nonexpansive operator T can be used to express $\mathscr{S}_p$ as in (19) but more nicely with some computable bounded linear operator $\Xi \in \mathscr{B}(\mathscr{H}, \mathscr{X})$, we can apply the hybrid steepest descent method to Problem (10) after translating it into

$$\text{find } z^\star \in \underset{z \in \mathrm{Fix}(T)}{\mathrm{argmin}}\, \Theta(z), \tag{22}$$

where $\Theta := \Psi \circ \Xi$, because $\Theta \in \Gamma_0(\mathscr{H})$ is certainly Gâteaux differentiable with Lipschitzian gradient $\nabla\Theta : z \mapsto \Xi^* \nabla\Psi(\Xi z)$ and $\Xi(z^\star) \in \mathscr{X}$ is a solution of (10).

The goal of this paper is to demonstrate that plugging the modern proximal splitting operators into the hybrid steepest descent method is a powerful computational strategy for solving highly valuable hierarchical convex optimization problems (8) in Scenario 1 and Scenario 2. The remainder of the paper is organized as follows. In the next section, as preliminaries, we introduce elements of convex analysis and fixed point theoretic view of the modern proximal splitting algorithms. These include key ideas behind fixed point characterizations of $\mathscr{S}_p$ in Problem (10) as well as the hybrid steepest descent method for nonexpansive operators. Section 3 contains the main idea of the hierarchical convex optimization based on the hybrid steepest descent method applied to modern proximal splitting operators. In Section 4, as a typical example of Scenario 1, we present an application of the proposed strategies to a hierarchical enhancement of *the support vector machine* [48, 135, 136] where we demonstrate how we can compute *the best linear classifier which achieves the maximal margin among all linear classifiers having least empirical hinge loss*. The proposed best linear classifier can be applied to general training data whether it is linearly separable or not. In particular, for linearly separable data, the proposed best linear classifier, *which does not require any parameter tuning*, is guaranteed to reproduce successfully the original support vector machine specially defined in [136]. To the best of the authors' knowledge, such a unified generalization of original support vector machine for linearly separable data has not been achieved by previously reported SVMs (see, e.g., [14, 25, 48, 73, 125, 126, 131] and Section 4.2). In Section 5, as a typical example along Scenario 2, we present an application of the proposed strategy to a hierarchical enhancement of Lasso [73, 132]. This enhancement is achieved by utilizing maximally the Douglas-Rachford splitting applied to a recently established proximity operator [35, 38] of a perspective function for the TREX problem [89] which is certainly *the state-of-the-art nonconvex formulation* for automatic sparsity control of Lasso. The proposed

[5] By extending the idea in [75], another algorithm, which we refer to as the *generalized Haugazeau's algorithm*, was developed for minimizing a *strictly convex* function in $\Gamma_0(\mathscr{H})$ over the fixed point set of a certain quasi-nonexpansive operator [33]. In particular, this algorithm was specialized in a clear way for finding the nearest fixed point of a certain quasi-nonexpansive operator [8] and applied successfully to an image recovery problem [39]. If we focus on the case of a nonstrictly convex function, the generalized Haugazeau's algorithm is not applicable, while some convergence theorems of the hybrid steepest descent method suggest its sound applicability *provided that the gradient of the function is Lipschitzian*.

application can optimize further an additional convex criterion over the all solutions of the TREX problem. Finally, in Section 6, we conclude this paper with some remarks on other possible advanced applications of the hybrid steepest descent method.

2 Preliminary

Let $\mathscr{X}$ be a real Hilbert space equipped with[6] an inner product $\langle\cdot,\cdot\rangle$ and its induced norm $\|\cdot\| = \sqrt{\langle\cdot,\cdot\rangle}$, which is denoted by $(\mathscr{X},\langle\cdot,\cdot\rangle,\|\cdot\|)$. Let $(\mathscr{K},\langle\cdot,\cdot\rangle_{\mathscr{K}},\|\cdot\|_{\mathscr{K}})$ be another real Hilbert space. Let $A\colon \mathscr{X}\to\mathscr{K}$ be a bounded linear operator of which the norm is defined by $\|A\|_{\mathrm{op}} := \sup_{x\in\mathscr{X}\colon \|x\|\leq 1}\|Ax\|_{\mathscr{K}}$. For a bounded linear operator $A\colon \mathscr{X}\to\mathscr{K}$, $A^*\colon \mathscr{K}\to\mathscr{X}$ denotes its adjoint or conjugate, i.e.,

$$(\forall(x,u)\in\mathscr{X}\times\mathscr{K}) \quad \langle x, A^*u\rangle = \langle Ax, u\rangle_{\mathscr{K}}.$$

2.1 Selected Elements of Convex Analysis and Optimization

For readers' convenience, we list minimum elements, in convex analysis, which will be used in the later sections (for their detailed accounts, see, e.g., [7, 9, 35, 38, 44, 64, 82, 122, 143, 152]).

(Convex Set) A set $C\subset\mathscr{X}$ is said to be convex if $\lambda x+(1-\lambda)y\in C$ for all $\lambda\in(0,1)$ and for all $x,y\in C$.

(Proper Lower Semicontinuous Convex Function; See, e.g., [9, Chapter 9] A function $f\colon \mathscr{X}\to(-\infty,\infty]$ is said to be proper if its effective domain $\mathrm{dom}(f) := \{x\in\mathscr{X}\mid f(x)<\infty\}$ is nonempty. A function $f\colon \mathscr{X}\to(-\infty,\infty]$ is said to be lower semicontinuous if its lower level set $\mathrm{lev}_{\leq\alpha} f := \{x\in\mathscr{X}\mid f(x)\leq\alpha\}(\subset\mathscr{X})$ is closed for every $\alpha\in\mathbb{R}$. A function $f\colon \mathscr{X}\to(-\infty,\infty]$ is said to be convex if $f(\lambda x+(1-\lambda)y)\leq\lambda f(x)+(1-\lambda)f(y)$ for all $\lambda\in(0,1)$ and for all $x,y\in\mathrm{dom}(f)$. In particular, f is said to be strictly convex if $f(\lambda x+(1-\lambda)y) < \lambda f(x)+(1-\lambda)f(y)$ for all $\lambda\in(0,1)$ and for all $x,y\in\mathrm{dom}(f)$ such that $x\neq y$. The set of all proper lower-semicontinuous convex functions defined over the real Hilbert space $\mathscr{X}$ is denoted by $\Gamma_0(\mathscr{X})$.

(Coercivity and Supercoercivity; See, e.g., [9, Chapter 11]) A function $f\colon \mathscr{X}\to(-\infty,\infty]$ is said to be coercive if

$$\|x\|\to\infty \;\Rightarrow\; f(x)\to\infty$$

and supercoercive if

$$\|x\|\to\infty \;\Rightarrow\; \frac{f(x)}{\|x\|}\to\infty.$$

Obviously, supercoercivity of f implies coercivity of f. Coercivity of $f\in\Gamma_0(\mathscr{X})$ implies that $\mathrm{lev}_{\leq\alpha} f = \{x\in\mathscr{X}\mid f(x)\leq\alpha\}$ is bounded for every $\alpha\in\mathbb{R}$ as well as $\mathrm{argmin}_{x\in\mathscr{X}} f(x)\neq\varnothing$. Strict convexity of $f\in\Gamma_0(\mathscr{X})$ implies that the set of minimizers is at most singleton.

Fact 2 (See, e.g., [9, Section 11.4]) *Let $f\in\Gamma_0(\mathscr{X})$, $g\in\Gamma_0(\mathscr{K})$ and $A\in\mathscr{B}(\mathscr{X},\mathscr{K})$ such that* $\mathrm{dom}(f)\cap\mathrm{dom}(g\circ A)\neq\varnothing$. *Then the following conditions*

(a) $\mathrm{argmin}(f+g\circ A)(\mathscr{X})$ *is nonempty, closed, and bounded;*

(b) $f+g\circ A$ *is coercive;*

(c) f *is coercive, and* g *is bounded below;*

(d) f *is super-coercive;*

[6] Often $\langle\cdot,\cdot\rangle_{\mathscr{X}}$ denotes $\langle\cdot,\cdot\rangle$ to explicitly describe its domain.

satisfy that $((d)$ *or* $(c)) \Rightarrow (b) \Rightarrow (a)$.

(Gâteaux Differential; See, e.g., [9, Section 2.6]) Let U be an open subset of $\mathscr{X}$. Then a function $f\colon U \to \mathbb{R}$ is said to be Gâteaux differentiable at $x \in U$ if there exists $a(x) \in \mathscr{X}$ such that

$$\lim_{\delta \to 0} \frac{f(x+\delta h) - f(x)}{\delta} = \langle a(x), h\rangle \quad (\forall h \in \mathscr{X}).$$

In this case, $\nabla f(x) := a(x)$ is called Gâteaux gradient (or gradient) of f at x. Let $f \in \Gamma_0(\mathscr{X})$ be Gâteaux differentiable at $x_\star \in \mathscr{X}$. Then $x_\star$ is a minimizer of f if and only if $\nabla f(x_\star) = 0$.
(Subdifferential; See, e.g., [9, Chapter 16]) For a function $f \in \Gamma_0(\mathscr{X})$, the subdifferential of f is defined as the set valued operator

$$\partial f\colon \mathscr{X} \to 2^{\mathscr{X}} : x \mapsto \{u \in \mathscr{X} \mid \langle y-x, u\rangle + f(x) \leq f(y), \forall y \in \mathscr{X}\}.$$

Every element $u \in \partial f(x)$ is called a subgradient of f at x. For a given function $f \in \Gamma_0(\mathscr{X})$, $x_\star \in \mathscr{X}$ is a minimizer of f if and only if $0 \in \partial f(x_\star)$. Note that if $f \in \Gamma_0(\mathscr{X})$ is Gâteaux differentiable at $x \in \mathscr{X}$, then $\partial f(x) := \{\nabla f(x)\}$.
(Conjugate Function; See, e.g., [9, Chapter 13 and Chapter 16]) For a function $f \in \Gamma_0(\mathscr{X})$, the conjugate of f is defined by

$$f^*\colon \mathscr{X} \to [-\infty, \infty] : u \mapsto \sup_{x \in \mathscr{X}} (\langle x, u\rangle - f(x)) = \sup_{x \in \mathrm{dom}(f)} (\langle x, u\rangle - f(x)).$$

Let $f \in \Gamma_0(\mathscr{X})$. Then $f^* \in \Gamma_0(\mathscr{X})$ and $f^{**} = f$ are guaranteed. Moreover, we have

$$(\forall (x,u) \in \mathscr{X} \times \mathscr{X}) \quad u \in \partial f(x) \Leftrightarrow f(x) + f^*(u) = \langle x, u\rangle \Leftrightarrow x \in \partial f^*(u),$$

which implies that $(\partial f)^{-1}(u) := \{x \in \mathscr{X} \mid u \in \partial f(x)\} = \partial f^*(u)$ and $(\partial f^*)^{-1}(x) := \{u \in \mathscr{X} \mid x \in \partial f^*(u)\} = \partial f(x)$. Often

$$(\forall x \in \mathrm{dom}(\partial f))(\forall u \in \partial f(x)) \quad f(x) + f^*(u) = \langle x, u\rangle \tag{23}$$

is referred to as Fenchel-Young identity.

For hierarchical enhancement of Lasso in Section 5, we exploit the following nontrivial example.

Example 3 *(Subdifferential of Perspective; See [38, Lemma 2.3])*
For given supercoercive $\varphi \in \Gamma_0(\mathbb{R}^N)$, *the function*

$$\widetilde{\varphi} : \mathbb{R} \times \mathbb{R}^N \to (-\infty, \infty] : (\eta, \mathbf{y}) \mapsto \begin{cases} \eta \varphi(\mathbf{y}/\eta), & \textit{if } \eta > 0; \\ \sup_{\mathbf{x} \in \mathrm{dom}(\varphi)} [\varphi(\mathbf{x}+\mathbf{y}) - \varphi(\mathbf{x})], & \textit{if } \eta = 0; \\ +\infty, & \textit{otherwise.} \end{cases} \tag{24}$$

satisfies $\widetilde{\varphi} \in \Gamma_0\left(\mathbb{R} \times \mathbb{R}^N\right)$ *and is called the perspective of* φ.
The subdifferential of $\widetilde{\varphi}$ *is given by*

$$\partial \widetilde{\varphi}(\eta, \mathbf{y}) = \begin{cases} \left\{(\varphi(\mathbf{y}/\eta) - \langle \mathbf{y}/\eta, \mathbf{u}\rangle, \mathbf{u}) \in \mathbb{R} \times \mathbb{R}^N \mid \mathbf{u} \in \partial \varphi(\mathbf{y}/\eta)\right\}, & \textit{if } \eta > 0; \\ \{(\mu, \mathbf{u}) \in \mathbb{R} \times \mathbb{R}^N \mid \mu + \varphi^*(\mathbf{u}) \leq 0\}, & \textit{if } \eta = 0 \textit{ and } \mathbf{y} = \mathbf{0}; \\ \varnothing, & \textit{otherwise.} \end{cases} \tag{25}$$

(Conical Hull, Span, Convex Sets; See, e.g., [9, Chapter 6]) For a given nonempty set $C \subset \mathscr{X}$, $\mathrm{cone}(C) := \{\lambda x \mid \lambda > 0, x \in C\}$ is called the conical hull of C, and $\mathrm{span}(C)$ denotes the intersection of all the linear subspaces of $\mathscr{X}$ containing C. The closure of $\mathrm{span}(C)$ is denoted by $\overline{\mathrm{span}}(C)$. The strong relative interior of a convex set $C \subset \mathscr{X}$ is defined by

$$\mathrm{sri}(C) := \{x \in C \mid \mathrm{cone}(C-x) = \overline{\mathrm{span}}(C-x)\},$$

where $C - x := \{y - x \in \mathscr{X} \mid y \in C\}$.
Similarly, the relative interior of a convex set $C \subset \mathscr{X}$ is defined by

$$\mathrm{ri}(C) := \{x \in C \mid \mathrm{cone}(C-x) = \mathrm{span}(C-x)\}.$$

By $\mathrm{cone}(C-x) \subset \mathrm{span}(C-x) \subset \overline{\mathrm{span}}(C-x)$ for every $x \in C$, we have $\mathrm{sri}(C) \subset \mathrm{ri}(C)$. Moreover, $\mathrm{sri}(C) = \mathrm{ri}(C)$ if $\mathrm{span}(C-x) = \overline{\mathrm{span}}(C-x)$ for every $x \in C$, which implies

$$\dim(\mathscr{X}) < \infty \;\Rightarrow\; \mathrm{sri}(C) = \mathrm{ri}(C). \tag{26}$$

(Indicator Function) For a nonempty closed convex set $C \subset \mathscr{X}$, the indicator function of C is defined by

$$\iota_C \colon \mathscr{X} \to (-\infty, \infty] : x \mapsto \begin{cases} 0, & \text{if } x \in C; \\ +\infty, & \text{otherwise,} \end{cases}$$

which belongs to $\Gamma_0(\mathscr{X})$. In particular, for a closed subspace $V \subset \mathscr{X}$,

$$u \in \partial \iota_V(x) \;\Leftrightarrow x \in V \text{ and } u \in V^{\perp} := \{y \in \mathscr{X} \mid (\forall v \in V)\ \langle v, y \rangle = 0\}. \tag{27}$$

Furthermore, the indicator function $\iota_{\{0\}} \in \Gamma_0(\mathscr{X})$ of $\{0\} \subset \mathscr{X}$ has the following properties: for all $x, u \in \mathscr{X}$

$$\partial \iota_{\{0\}}(x) = \begin{cases} \mathscr{X}, & \text{if } x = 0; \\ \varnothing, & \text{otherwise,} \end{cases} \tag{28}$$

$$\iota_{\{0\}}^*(u) = \sup_{y \in \mathscr{X}} (\langle y, u \rangle - \iota_{\{0\}}(y)) = 0, \tag{29}$$

$$\partial \iota_{\{0\}}^*(u) = \{0\}.$$

(Fenchel-Rockafellar Duality for Convex Optimization Problem Involving Linear Operator; See, e.g., [9, Definition 15.19]) Let $f \in \Gamma_0(\mathscr{X})$, $g \in \Gamma_0(\mathscr{K})$, and $A \in \mathscr{B}(\mathscr{X}, \mathscr{K})$. The primal problem associated with the composite function $f + g \circ A$ is

$$\mathrm{minimize}_{x \in \mathscr{X}}\ f(x) + g(Ax), \tag{30}$$

its dual problem is

$$\mathrm{minimize}_{u \in \mathscr{K}}\ f^*(A^* u) + g^*(-u), \tag{31}$$

$\mu := \inf_{x \in \mathscr{X}} (f(x) + g(Ax))$ is called the primal optimal value, and $\mu^* := \inf_{u \in \mathscr{K}} (f^*(A^* u) + g^*(-u))$ the dual optimal value.

Fact 4 (See, e.g., [9, Theorem 15.23, Theorem 16.47, Corollary 16.53]) *The condition*

$$\left.\begin{array}{l} 0 \in \mathrm{sri}(\mathrm{dom}(g) - A\,\mathrm{dom}(f)) \\ \textit{(}\mathrm{sri}\textit{ can be replaced by } \mathrm{ri} \textit{ in the case of } \dim(\mathscr{K}) < \infty\textit{, see } (26)\textit{)} \end{array}\right\} \tag{32}$$

is the so-called qualification condition *for problem* (30).

(a) *The condition* (32) *guarantees that the dual problem* (31) *has a minimizer and satisfies*

$$\mu = \inf_{x \in \mathscr{X}} (f(x) + g(Ax)) = -\min_{u \in \mathscr{K}} (f^*(A^* u) + g^*(-u)) = -\mu^*;$$

(b) *The condition* (32) *guarantees that the subdifferential of* $f + g \circ A$ *can be decomposed as*

$$\partial(f+g\circ A)=\partial f+A^*\circ(\partial g)\circ A;$$

(c) *The qualification condition* (32) *with* $f\equiv 0$ *becomes* $0\in \operatorname{sri}(\operatorname{dom}(g)-\operatorname{ran}(A))$, *where* $\operatorname{ran}(A):=A(\mathscr{X}):=\{Ax\in\mathscr{K}\mid x\in\mathscr{X}\}$. *Under this condition, (a), (b), and* (29) *guarantee*

$$\left[\begin{array}{l}\mu=\inf_{x\in\mathscr{X}} g(Ax)=\inf_{x\in\mathscr{X}} \iota^*_{\{0\}}(x)+g(Ax)=-\min_{u\in\mathscr{K}}(\iota_{\{0\}}(A^*u)+g^*(-u))\\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad =-\min_{u\in\mathscr{N}(A^*)} g^*(-u)=-\mu^*\\ \partial(g\circ A)=A^*\circ\partial g\circ A.\end{array}\right.$$

Fact 5 ([9, Theorem 19.1]) *Suppose that* $0\in\operatorname{dom}(g)-A\operatorname{dom}(f)$ *(Note: This condition is not sufficient for* (32)*). Let* $(x,u)\in\mathscr{X}\times\mathscr{K}$. *Then the following are equivalent:*

(i) x *is a solution of the primal problem* (30), u *is a solution of the dual problem* (31), *and* $\mu=-\mu^*$.
(ii) $A^*u\in\partial f(x)$ *and* $-u\in\partial g(Ax)$.
(iii) $x\in\partial f^*(A^*u)\cap A^{-1}(\partial g^*(-u))$.

2.2 Selected Elements of Fixed Point Theory of Nonexpansive Operators for Application to Hierarchical Convex Optimization

For readers' convenience, we list minimum elements in fixed point theory of nonexpansive mapping specially for application to hierarchical convex optimization in this paper (for their detailed accounts, see, e.g., [7, 9, 38, 42, 44, 51, 82, 122, 130, 143]).
(Monotone Operator; See, e.g., [9, Section 20.1]) A set-valued operator $T:\mathscr{X}\to 2^{\mathscr{X}}$ is said to be monotone over $S(\subset\mathscr{X})$ if

$$(\forall x,y\in S)(\forall u\in Tx)(\forall v\in Ty)\quad \langle u-v,x-y\rangle\geq 0.$$

In particular, it is said to be η-strongly monotone over S if

$$(\exists\eta>0)(\forall x,y\in S)(\forall u\in Tx)(\forall v\in Ty)\quad \langle u-v,x-y\rangle\geq \eta\|x-y\|^2.$$

(Nonexpansive Operator; See, e.g., [7] and [9, Chapter 4]) An operator $T:\mathscr{X}\to\mathscr{X}$ is said to be Lipschitz continuous with Lipschitz constant $\kappa>0$ (or κ-Lipschitzian) if

$$(\forall x,y\in\mathscr{X})\quad \|Tx-Ty\|\leq\kappa\|x-y\|.$$

In particular, an operator $T:\mathscr{X}\to\mathscr{X}$ is said to be nonexpansive if it is 1-Lipschitzian, i.e.,

$$(\forall x,y\in\mathscr{X})\quad \|Tx-Ty\|\leq\|x-y\|.$$

A nonexpansive operator T is said to be α-averaged (or averaged with constant α) [5, 9] if there exist $\alpha\in(0,1)$ and a nonexpansive operator $\widehat{T}:\mathscr{X}\to\mathscr{X}$ such that

$$T=(1-\alpha)\mathrm{I}+\alpha\widehat{T}, \tag{33}$$

i.e., T is an *average* of the identity operator I and some nonexpansive operator $\widehat{T}$. If (33) holds for $\alpha=1/2$, T is said to be firmly nonexpansive. A nonexpansive operator T is α-averaged if and only if

$$(\forall x,y\in\mathscr{X})\quad \|Tx-Ty\|^2\leq\|x-y\|^2-\frac{1-\alpha}{\alpha}\|(x-Tx)-(y-Ty)\|^2. \tag{34}$$

Suppose that a nonexpansive operator T has the fixed point set $\mathrm{Fix}(T) := \{x \in \mathscr{X} \mid Tx = x\} \neq \varnothing$. Then $\mathrm{Fix}(T)$ can be expressed as the intersection of closed halfspaces:

$$\mathrm{Fix}(T) = \bigcap_{y\in\mathscr{X}} \left\{ x \in \mathscr{X} \mid \langle y - T(y), x\rangle \leq \frac{\|y\|^2 - \|T(y)\|^2}{2} \right\}$$

and therefore $\mathrm{Fix}(T)$ is closed and convex (see, e.g., [70, Proposition 5.3], [142, Fact 2.1(a)], and [9, Corollary 4.24]). In addition, a nonexpansive operator T with $\mathrm{Fix}(T) \neq \varnothing$ is said to be attracting [7] if

$$(\forall x \notin \mathrm{Fix}(T))(\forall z \in \mathrm{Fix}(T)) \quad \|Tx - z\| < \|x - z\|.$$

The condition (34) implies that α-averaged nonexpansive operator T is attracting if $\mathrm{Fix}(T) \neq \varnothing$. Note that other useful properties on α-averaged nonexpansive operators are found, e.g., in [20, 45, 105].

Fact 6 (Krasnosel'skiĭ-Mann (KM) Iteration [71] (See Also [9, Section 5.2],[20, 56, 88, 96, 119])) *For a nonexpansive operator $T\colon \mathscr{X} \to \mathscr{X}$ with $\mathrm{Fix}(T) \neq \varnothing$ and any initial point $x_0 \in \mathscr{X}$, the sequence $(x_n)_{n\in\mathbb{N}}$ generated by*

$$x_{n+1} = (1 - \alpha_n)x_n + \alpha_n T x_n$$

converges weakly[7] *to a point in $\mathrm{Fix}(T)$ if $(\alpha_n)_{n\in\mathbb{N}} \subset [0,1]$ satisfies $\sum_{n\in\mathbb{N}} \alpha_n(1-\alpha_n) = \infty$ (Note: The weak limit of $(x_n)_{n\in\mathbb{N}}$ depends on the choices of x_0 and $(\alpha_n)_{n\in\mathbb{N}}$).*[8] *In particular, if T is α-averaged for some $\alpha \in (0,1)$ (see (33)), a simple iteration*

$$x_{n+1} = Tx_n = (1-\alpha)x_n + \alpha \widehat{T} x_n \tag{35}$$

converges weakly to a point in $\mathrm{Fix}(T) = \mathrm{Fix}(\widehat{T})$.

(Proximity Operator [101, 102] (See Also [9, Chapter 24])) The proximity operator of $f \in \Gamma_0(\mathscr{X})$ is defined by

$$\mathrm{prox}_f\colon \mathscr{X} \to \mathscr{X} : x \mapsto \operatorname*{argmin}_{y\in\mathscr{X}} f(y) + \frac{1}{2}\|y - x\|^2.$$

Note that $\mathrm{prox}_f(x) \in \mathscr{X}$ is well defined for all $x \in \mathscr{X}$ due to the coercivity and the strict convexity of $f(\cdot) + \frac{1}{2}\|\cdot - x\|^2 \in \Gamma_0(\mathscr{X})$. It is also well known that prox_f is nothing but the resolvent of ∂f, i.e., $\mathrm{prox}_f = (\mathrm{I} + \partial f)^{-1} =: J_{\partial f}$, which implies that

$$\begin{aligned} & x \in \mathrm{Fix}(\mathrm{prox}_f) \ \Leftrightarrow\ \mathrm{prox}_f(x) = x \ \Leftrightarrow\ (\mathrm{I} + \partial f)^{-1}(x) = x \\ & \Leftrightarrow\ x \in (\mathrm{I} + \partial f)(x) \ \Leftrightarrow\ 0 \in \partial f(x) \ \Leftrightarrow\ x \in \operatorname*{argmin}_{y\in\mathscr{X}} f(y). \end{aligned} \tag{36}$$

Thanks to this fact, the set of all minimizers of $f \in \Gamma_0(\mathscr{X})$ can be characterized in terms of a single-valued map, i.e., prox_f. Moreover, since the proximity operator is $1/2$-averaged nonexpansive, i.e., $\mathrm{rprox}_f := 2\mathrm{prox}_f - \mathrm{I}$ is nonexpansive, the iteration

[7] (Strong and weak convergences) A sequence $(x_n)_{n\in\mathbb{N}} \subset \mathscr{X}$ is said to converge strongly to a point $x \in \mathscr{X}$ if the real number sequence $(\|x_n - x\|)_{n\in\mathbb{N}}$ converges to 0, and to converge weakly to $x \in \mathscr{X}$ if for every $y \in \mathscr{X}$ the real number sequence $(\langle x_n - x, y\rangle)_{n\in\mathbb{N}}$ converges to 0. If $(x_n)_{n\in\mathbb{N}}$ converges strongly to x, then $(x_n)_{n\in\mathbb{N}}$ converges weakly to x. The converse is true if $\mathscr{X}$ is finite dimensional, hence in finite dimensional case we do not need to distinguish these convergences.
(Sequential cluster point) If a sequence $(x_n)_{n\in\mathbb{N}} \subset \mathscr{X}$ possesses a subsequence that strongly (weakly) converges to a point $x \in \mathscr{X}$, then x is a strong (weak) sequential cluster point of $(x_n)_{n\in\mathbb{N}}$. For weak topology of real Hilbert space in the context of Hausdorff space, see [9, Lemma 2.30].

[8] Some extensions to uniformly convex Banach spaces are found in [71, 119].

$$x_{n+1} = \mathrm{prox}_f(x_n) \tag{37}$$

converges weakly to a point in $\mathrm{argmin}_{x\in\mathscr{X}} f(x) = \mathrm{Fix}(\mathrm{prox}_f)$ by (35) in Fact 6. The iterative algorithm (37) is known as *proximal point algorithm* [121] (see (6)).
In this paper, $f \in \Gamma_0(\mathscr{X})$ is said to be *proximable* if prox_f is available as a computable operator. Note that if $f \in \Gamma_0(\mathscr{X})$ is proximable, so is $f^* \in \Gamma_0(\mathscr{X})$. This is verified by

$$\mathrm{prox}_{f^*} = J_{\partial f^*} = J_{(\partial f)^{-1}} = \mathrm{I} - J_{\partial f} = \mathrm{I} - \mathrm{prox}_f,$$

which is a special example of *the inverse resolvent identity* [9, Proposition 23.20]. Note that the sum of two proximable convex functions is not necessarily proximable. Moreover, for $A \in \mathscr{B}(\mathscr{X},\mathscr{K})$, the composition $g\circ A \in \Gamma_0(\mathscr{X})$ for a proximable function $g \in \Gamma_0(\mathscr{K})$ is not necessarily proximable. There are many useful formula to compute the proximity operator (see, e.g., [9, Chapter 24], [42]).

Example 7

(a) **(Indicator function; see, e.g., [9, Example 12.25])** *For a nonempty closed convex set $C \subset \mathscr{X}$,*

$$(\forall x \in \mathscr{X}) \quad \mathrm{prox}_{\iota_C}(x) = \underset{y\in\mathscr{X}}{\mathrm{argmin}} \left(\iota_C(y) + \frac{1}{2}\|y-x\|^2\right) = \underset{y\in C}{\mathrm{argmin}}\ \frac{1}{2}\|y-x\|^2 =: P_C(x)$$

holds, which implies that prox_{ι_C} *is identical to the* metric projection *onto C. In particular, if ι_C is proximable, C is said to be* simple.

(b) **(Semi-orthogonal linear transform of proximable function; see, e.g., [9, Proposition 24.14] and [42, Table 10.1])**
For $g \in \Gamma_0(\mathscr{K})$ and $A \in \mathscr{B}(\mathscr{X},\mathscr{K})$ such that $AA^* = \nu\mathrm{I}$ with some $\nu > 0$,

$$(\forall x \in \mathscr{X})\ \mathrm{prox}_{g\circ A}(x) = x + \nu^{-1}A^*(\mathrm{prox}_{\nu g}(Ax) - Ax). \tag{38}$$

(c) **(Hinge loss function; see, e.g., [1] and [9, Example 24.36])** *For $\gamma > 0$ and*

$$h : \mathbb{R} \to [0,\infty) : t \mapsto \max\{0, 1-t\}, \tag{39}$$

$$(\forall t \in \mathbb{R}) \quad \mathrm{prox}_{\gamma h}(t) = \min\{t+\gamma, \max\{t,1\}\}. \tag{40}$$

(d) **(ℓ_1 norm; see, e.g., [9, 44])** *For $\gamma \geq 0$ and the ℓ_1 norm $\|\cdot\|_1 \in \Gamma_0(\mathbb{R}^N)$*

$$(\forall \mathbf{x} = (x_1,x_2,\ldots,x_N) \in \mathbb{R}^N) \quad \|\mathbf{x}\|_1 := \sum_{j=1}^{N} |x_i|,$$

the i-th component of the proximity operator of $\gamma\|\cdot\|_1$ is given as

$$(\forall \mathbf{x} = (x_1,x_2,\ldots,x_N) \in \mathbb{R}^N) \quad [\mathrm{prox}_{\gamma\|\cdot\|_1}(\mathbf{x})]_i = \begin{cases} x_i - \mathrm{sgn}(x_i)\gamma, & \text{if } |x_i| > \gamma; \\ 0, & \text{otherwise,} \end{cases}$$

where $\mathrm{sgn}\colon \mathbb{R} \to \mathbb{R}$ *is the signum function, i.e.,* $\mathrm{sgn}(x) = 0$ *if $x = 0$ and* $\mathrm{sgn}(x) = x/|x|$ *otherwise.* $\mathrm{prox}_{\gamma\|\cdot\|_1}$ *is also known as soft-thresholding [53, 54].*

(e) **(Proximity operator of *perspective* of $\|\cdot\|^q$; see, e.g., [38])** *Let $\beta > 0$ and $q > 1$. The perspective $\widetilde{\varphi}_q$ of $\varphi_q(\cdot) := \|\cdot\|^q/\beta$ (see also* (24) *in Example 3) is given by*

$$\widetilde{\varphi}_q : \mathbb{R}\times\mathbb{R}^N \to (-\infty,\infty] : (\eta,\mathbf{y}) \mapsto \begin{cases} \frac{\|\mathbf{y}\|^q}{\beta\eta^{q-1}}, & \textit{if } \eta > 0; \\ 0, & \textit{if } \eta = 0 \textit{ and } \mathbf{y} = \mathbf{0}; \\ +\infty, & \textit{otherwise,} \end{cases} \tag{41}$$

and its proximity operator can be expressed as

$$\begin{aligned}\operatorname{prox}_{\widetilde{\varphi}_q} &: \mathbb{R}\times\mathbb{R}^N \to \mathbb{R}\times\mathbb{R}^N\\ &: (\eta,\mathbf{y}) \mapsto \begin{cases} \left(\eta + \frac{\rho}{q^*}\|\mathbf{p}\|^{q^*}, \mathbf{y}-\mathbf{p}\right), & \textit{if } q^*\eta + \rho\|\mathbf{y}\|^{q^*} > 0;\\ (0,\mathbf{0}), & \textit{if } q^*\eta + \rho\|\mathbf{y}\|^{q^*} \le 0, \end{cases}\\ \textit{where} \quad & q^* := \frac{q}{q-1},\ \rho := (\beta(1-1/q^*))^{q^*-1},\ \mathbf{p} := \begin{cases} \tau\frac{\mathbf{y}}{\|\mathbf{y}\|}, & \textit{if } \mathbf{y} \neq \mathbf{0};\\ \mathbf{0}, & \textit{if } \mathbf{y} = \mathbf{0}, \end{cases}\end{aligned}$$

and $\tau \in (0,\infty)$ is uniquely determined as the solution to the equation:

$$\tau^{2q^*-1} + \frac{q^*\eta}{\rho}\tau^{q^*-1} + \frac{q^*\|\mathbf{y}\|}{\rho^2} = 0.$$

The the proximity operator of the translation, by $(a,\mathbf{b}) \in \mathbb{R}\times\mathbb{R}^N$, of $\widetilde{\varphi}_q$

$$\tau_{(a,\mathbf{b})}\widetilde{\varphi}_q : \mathbb{R}\times\mathbb{R}^N \to \mathbb{R}\times\mathbb{R}^N : (\eta,\mathbf{y}) \mapsto \widetilde{\varphi}_q(\eta-a,\mathbf{y}-\mathbf{b}),$$

which can be expressed as

$$\operatorname{prox}_{\tau_{(a,\mathbf{b})}\widetilde{\varphi}_q} : \mathbb{R}\times\mathbb{R}^N \to \mathbb{R}\times\mathbb{R}^N : (\eta,\mathbf{y}) \mapsto (a,\mathbf{b}) + \operatorname{prox}_{\widetilde{\varphi}_q}(\eta-a,\mathbf{y}-\mathbf{b}),$$

will play an important role in Section 5.

Fact 8 (Moreau Envelope (See, e.g., [9, Section 12.4], [101, 102]))
For $f \in \Gamma_0(\mathscr{X})$,

$$^{\gamma}f\colon \mathscr{X} \to \mathbb{R} : x \mapsto \min_{y\in\mathscr{X}}\left(f(y) + \frac{1}{2\gamma}\|x-y\|^2\right)$$

is called the Moreau envelope *(or* Moreau-Yosida regularization) *[101, 102] of f of the index $\gamma > 0$. The function $^{\gamma}f$ is Gâteaux differentiable convex with Lipschitzian gradient*

$$\nabla^{\gamma}f\colon \mathscr{X} \to \mathscr{X} : x \mapsto \frac{1}{\gamma}(\mathrm{I} - \operatorname{prox}_{\gamma f})(x).$$

The Moreau envelope of f converges pointwise to f on $\operatorname{dom}(f)$ *as $\gamma \downarrow 0$ (see, e.g., [9, Proposition 12.33(ii)]), i.e.,* $\lim_{\gamma\downarrow 0} {}^{\gamma}f(x) = f(x)$ $(\forall x \in \operatorname{dom}(f))$.

2.3 Proximal Splitting Algorithms and Their Fixed Point Characterizations

In this section, we introduce the Douglas-Rachford splitting method[9] (see, e.g., [9, 10, 34, 40, 61, 91]) and the linearized augmented Lagrangian method (see, e.g., [150, 151]) as examples of *the proximal splitting algorithms* built on computable nonexpansive operators with a great deal of potential in their applications to the hierarchical

[9] See [10, 42] for the history of the Douglas-Rachford splitting method, originated from Douglas-Rachford's seminal paper [57] for solving matrix equations of the form $u = Ax + Bx$, where A and B are positive-definite matrices (see also [137]). For recent applications, of the Douglas-Rachford splitting method, to image recovery, see, e.g., [26, 40, 58, 60], and to data sciences, see, e.g., [38, 67, 68]. Lastly, we remark that it was shown in [61] that *the alternating direction method of multipliers (ADMM)* [17, 62, 66, 91, 150] can be seen as a dual variant of the Douglas-Rachford splitting method.

convex optimization problem. As explained briefly just after (19–21), these proximal splitting algorithms are essentially realized by applying Fact 6 (see Section 2.2) to certain computable nonexpansive operators.

Proposition 9 (DRS Operator and Douglas-Rachford Splitting Method[10]) *Let* $(\mathscr{X}, \langle\cdot,\cdot\rangle_{\mathscr{X}}, \|\cdot\|_{\mathscr{X}})$ *be a real Hilbert space and* $f, g \in \Gamma_0(\mathscr{X})$*. Suppose that*

$$\operatorname{argmin}(f+g)(\mathscr{X}) \neq \varnothing, \tag{42}$$
$$\operatorname{argmin}(f^* + g^* \circ (-\mathrm{I}))(\mathscr{X}) \neq \varnothing, \tag{43}$$
$$\min(f+g)(\mathscr{X}) = -\min(f^* + g^* \circ (-\mathrm{I}))(\mathscr{X}). \tag{44}$$

Then the DRS operator

$$T_{\mathrm{DRS}} := (2\operatorname{prox}_f - \mathrm{I}) \circ (2\operatorname{prox}_g - \mathrm{I}) \tag{45}$$

satisfies:

(a) $\operatorname{prox}_g(\operatorname{Fix}(T_{\mathrm{DRS}})) = \operatorname{argmin}(f+g)(\mathscr{X})$*;*
(b) T_{DRS} *is nonexpansive;*
(c) *By using* $(\alpha_n)_{n\in\mathbb{N}} \subset [0,1]$ *satisfying* $\sum_{n\in\mathbb{N}} \alpha_n(1-\alpha_n) = \infty$ *in Fact 6 (see Section 2.2), the sequence* $(y_n)_{n\in\mathbb{N}} \subset \mathscr{X}$ *generated by*

$$y_{n+1} = (1-\alpha_n)y_n + \alpha_n T_{\mathrm{DRS}}(y_n) \tag{46}$$

converges weakly to a point in $\operatorname{Fix}(T_{\mathrm{DRS}})$*. Moreover,* $(\operatorname{prox}_g(y_n))_{n\in\mathbb{N}}$ *converges weakly to a point in* $\operatorname{argmin}(f+g)(\mathscr{X})$*.*

The iterative algorithm to produce $(\operatorname{prox}_g(y_n))_{n\in\mathbb{N}}$ with (46) can be seen as a simplest example of the so-called Douglas-Rachford splitting method.

The proof of Proposition 9(a) is given in Appendix A because the conditions (42–44) are newly imposed for applications of T_{DRS} (in (45)) to hierarchical convex optimizations in Theorem 15 and in Theorem 17 (see Remark 16(b) and Remark 18(b) in Section 3.1) and different from [40, Condition (6)] which is also in the context of convex optimization. Proposition 9(b) is obvious from the properties of the proximity operator just after (36). For weak convergence of $(\operatorname{prox}_g(y_n))_{n\in\mathbb{N}}$ in Proposition 9(c), see, e.g., [9, Corollary 28.3(iii)] while the weak convergence of $(y_n)_{n\in\mathbb{N}}$ is obvious from Fact 6.

The linearized augmented Lagrangian method (LALM) seems to have been proposed originally as an algorithmic solution to the minimization of the nuclear norm of a matrix subject to a linear constraint [151]. Inspired by the operator defined as the iterative update [151, (3.7)] in the method for this special convex optimization problem, we extended in [150] the operator to T_{LAL} in (50) to be applicable to the general convex optimization problem (1) and showed the nonexpansiveness of T_{LAL} for solving efficiently the hierarchical convex optimization (10) by plugging the extended operator T_{LAL} into the HSDM.

Proposition 10 (LAL Operator and Linearized Augmented Lagrangian Method) *Let* $(\mathscr{X}, \langle\cdot,\cdot\rangle_{\mathscr{X}}, \|\cdot\|_{\mathscr{X}})$ *and* $(\mathscr{K}, \langle\cdot,\cdot\rangle_{\mathscr{K}}, \|\cdot\|_{\mathscr{K}})$ *be real Hilbert spaces. Suppose that* $f \in \Gamma_0(\mathscr{X})$*,* $g = \iota_{\{0\}} \in \Gamma_0(\mathscr{K})$ *and* $A \in \mathscr{B}(\mathscr{X}, \mathscr{K})$ *satisfy*

$$\mathscr{S}_{\mathrm{pLAL}} := \operatorname{argmin}(f + \iota_{\{0\}} \circ A)(\mathscr{X}) \neq \varnothing, \tag{47}$$
$$\mathscr{S}_{\mathrm{dLAL}} := \operatorname{argmin}(f^* \circ A^*)(\mathscr{K}) \neq \varnothing, \tag{48}$$
$$\min(f + \iota_{\{0\}} \circ A)(\mathscr{X}) = -\min(f^* \circ A^*)(\mathscr{K}), \tag{49}$$

[10] We should remark that Proposition 9 can also be reproduced from [9, Proposition 26.1(iii) and Theorem 26.11(i)(iii)] in the context of the monotone inclusion problems. For completeness, we present Proposition 1 and its proof in the scenario of convex optimization.

where $\mathscr{S}_{\mathrm{pLAL}}$ is the solution set of the primal problem and $\mathscr{S}_{\mathrm{dLAL}}$ is the solution set of the dual problem. Define the LAL operator $T_{\mathrm{LAL}}\colon \mathscr{X}\times\mathscr{K}\to\mathscr{X}\times\mathscr{K}\colon (x,\nu)\mapsto(x_T,\nu_T)$ by

$$\begin{bmatrix} x_T := \mathrm{prox}_f(x - A^*Ax + A^*\nu) \\ \nu_T := \nu - Ax_T. \end{bmatrix} \tag{50}$$

Then

(a) $\mathrm{Fix}(T_{\mathrm{LAL}}) = \mathscr{S}_{\mathrm{pLAL}}\times\mathscr{S}_{\mathrm{dLAL}}$*;*
(b) T_{LAL} *is nonexpansive if* $\|A\|_{\mathrm{op}}\le 1$*;*
(c) *By using* $(\alpha_n)_{n\in\mathbb{N}}\subset[0,1]$ *satisfying* $\sum_{n\in\mathbb{N}}\alpha_n(1-\alpha_n)=\infty$ *in Fact 6 (see Section 2.2), the sequence* $(x_n,\nu_n)_{n\in\mathbb{N}}\subset\mathscr{X}\times\mathscr{K}$ *generated by*

$$(x_{n+1},\nu_{n+1}) = (1-\alpha_n)(x_n,\nu_n) + \alpha_n T_{\mathrm{LAL}}(x_n,\nu_n) \tag{51}$$

converges weakly to a point in $\mathscr{S}_{\mathrm{pLAL}}\times\mathscr{S}_{\mathrm{dLAL}}$ *if* $\|A\|_{\mathrm{op}}\le 1$*;*
(d) *If* $\|A\|_{\mathrm{op}}<1$*, the sequence* $(x_n,\nu_n)_{n\in\mathbb{N}}\subset\mathscr{X}\times\mathscr{K}$ *generated by* (51) *with* $\alpha_n=1$ $(n\in\mathbb{N})$ *converges weakly to a point in* $\mathscr{S}_{\mathrm{pLAL}}\times\mathscr{S}_{\mathrm{dLAL}}$.

The iterative algorithms, in Proposition 10(c) and (d), to produce $(x_n)_{n\in\mathbb{N}}$ with (51) can be seen as simplest examples of the so-called linearized augmented Lagrangian method.

The proof of Proposition 10(a) is given in Appendix B for completeness because the conditions (47–49) are newly imposed for applications of T_{LAL} to hierarchical convex optimizations in Theorem 19 and in Theorem 21 (see Remark 20(b) and Remark 22(a) in Section 3.2) and different from [150, (32)]. For the proof of Proposition 10(b), see [150]. Proposition 10(c) is a straightforward application of Fact 6 to Proposition 10(b). The proof of Proposition 10(d) is given in Appendix B.

Remark 11 *A primitive idea behind the update of the LAL operator T_{LAL} is found in minimization of the augmented Lagrangian function [81, 116]:*

$$\mathscr{L}\colon \mathscr{X}\times\mathscr{K}\to(-\infty,\infty]:(x,\nu)\mapsto f(x)-\langle\nu,Ax\rangle_{\mathscr{K}}+\frac{1}{2}\|Ax\|^2_{\mathscr{K}}. \tag{52}$$

Indeed, by introducing

$$(\forall\hat{x}\in\mathscr{X})(\forall\hat{\nu}\in\mathscr{K})\begin{bmatrix}\mathscr{L}_1^{(\hat{\nu})}\colon\mathscr{X}\to(-\infty,\infty]\colon x\mapsto\mathscr{L}(x,\hat{\nu}); \\ \mathscr{L}_2^{(\hat{x})}\colon\mathscr{K}\to(-\infty,\infty]\colon \nu\mapsto\mathscr{L}(\hat{x},\nu),\end{bmatrix}$$

the zero $(x_\star,\nu_\star)\in\mathscr{X}\times\mathscr{K}$ of the partial subdifferentials of (52) *is characterized as*

$$\begin{aligned}\begin{bmatrix}0\in\partial\mathscr{L}_1^{(\nu_\star)}(x_\star)\\ 0\in\partial\mathscr{L}_2^{(x_\star)}(\nu_\star)\end{bmatrix} &\Leftrightarrow \begin{bmatrix}0\in\partial f(x_\star)-A^*\nu_\star+A^*(Ax_\star)\\ 0=-Ax_\star\end{bmatrix}\\ &\Leftrightarrow \begin{bmatrix}x_\star=\mathrm{prox}_f(x_\star-A^*Ax_\star+A^*\nu_\star)\\ \nu_\star=\nu_\star-Ax_\star\end{bmatrix}\\ &\Leftrightarrow (x_\star,\nu_\star)\in\mathrm{Fix}(T_{\mathrm{LAL}}).\end{aligned}$$

2.4 Hybrid Steepest Descent Method

Consider the problem

$$\text{find } x_\star \in \operatorname*{argmin}_{x\in \mathrm{Fix}(T)} \Theta(x) =: \Omega \neq \varnothing, \tag{53}$$

where $\Theta \in \Gamma_0(\mathscr{H})$ is Gâteaux differentiable over $T(\mathscr{H})$ and $T : \mathscr{H} \to \mathscr{H}$ is a nonexpansive operator with $\mathrm{Fix}(T) \neq \varnothing$. The hybrid steepest descent method (HSDM)

$$x_{n+1} = T(x_n) - \lambda_{n+1}\nabla\Theta(T(x_n)) \tag{54}$$

generates a sequence $(x_n)_{n\in\mathbb{N}}$ to approximate successively a solution of Problem (53).

Fact 12 (Hybrid Steepest Descent Method for Nonexpansive Operators)

I. *[142, special case of Theorems 3.2 and 3.3 for more general variational inequality problems] Let $T : \mathscr{H} \to \mathscr{H}$ be a nonexpansive mapping with* $\mathrm{Fix}(T) \neq \varnothing$. *Suppose that the gradient $\nabla\Theta$ is κ-Lipschitzian and η-strongly monotone over $T(\mathscr{H}) := \{T(x) \in \mathscr{H} \mid x \in \mathscr{H}\}$, which guarantees $|\Omega| = 1$. Then, by using any sequence $(\lambda_{n+1})_{n\in\mathbb{N}} \subset [0,\infty)$ satisfying (W1) $\lim_{n\to+\infty}\lambda_n = 0$, (W2) $\sum_{n\in\mathbb{N}}\lambda_{n+1} = +\infty$, (W3) $\sum_{n\in\mathbb{N}}|\lambda_{n+1} - \lambda_{n+2}| < \infty$ [or $(\lambda_{n+1})_{n\in\mathbb{N}} \subset (0,\infty)$ satisfying (L1) $\lim_{n\to+\infty}\lambda_n = 0$, (L2) $\sum_{n\in\mathbb{N}}\lambda_{n+1} = +\infty$, (L3) $\lim_{n\to+\infty}(\lambda_n - \lambda_{n+1})\lambda_{n+1}^{-2} = 0$], the sequence $(x_n)_{n\in\mathbb{N}} \subset \mathscr{H}$ generated, for arbitrary $x_0 \in \mathscr{H}$, by (54) converges strongly to the uniquely existing solution of Problem (53).*

II. *(Nonstrictly convex case [105, 106, 149]) Assume that* $\dim(\mathscr{H}) < \infty$. *Suppose that (i) $T : \mathscr{H} \to \mathscr{H}$ is an attracting nonexpansive operator with bounded* $\mathrm{Fix}(T) \neq \varnothing$, *(ii) $\nabla\Theta$ is κ-Lipschitzian over $T(\mathscr{H})$, which guarantees $\Omega \neq \varnothing$. Then, by using*[11] *$(\lambda_{n+1})_{n\in\mathbb{N}} \in \ell^2_+ \setminus \ell^1_+$, the sequence $(x_n)_{n\in\mathbb{N}}$ generated by (54), for arbitrary $x_0 \in \mathscr{H}$, satisfies* $\lim_{n\to\infty} d_\Omega(x_n) = 0$, *where $d_\Omega(x_n) := \min_{y\in\Omega}\|x_n - y\|$.*

Remark 13

(a) **(Comparison between Fact 12(I) and Fact 6)** *Fact 6 in Section 2.2 is available for generation of a weak convergent sequence to a point in* $\mathrm{Fix}(T)$, *where the weak limit depends on the choices of x_0 and $(\alpha_n)_{n\in\mathbb{N}}$. Fact 12(I) guarantees* the strong convergence *of $(x_n)_{n\in\mathbb{N}}$ to a point in* $\mathrm{Fix}(T)$, *where the strong limit is optimal in* $\mathrm{Fix}(T)$ *because it minimizes Θ able to be designed strategically for many applications. Note that, thanks to Fact 12(I), we present that the LAL operator plugged into the HSDM yields an iterative approximation, of a solution of Problem* (53), *whose the strong convergence is guaranteed if Θ has the strongly monotone Lipschitzian gradient over $\mathscr{H}$ (see Theorem 19 below).*

(b) **(Boundedness assumption of $\mathrm{Fix}(T)$ in Fact 12(II))** *For readers who get worried about the boundedness assumption in Fact 12(II), we present some sufficient conditions, in Section 3.3, to guarantee the boundedness for* $\mathrm{Fix}(T)$ *in the context of DRS operators and LAL operators. These conditions hold automatically in the application to the hierarchical enhancement of the Lasso, in Section 5.2. However, the boundedness assumption in Fact 12(II) may not be restrictive for most practitioners by just modifying our original target* (53) *into*

$$\text{minimize}\,\Theta(x) \text{ subject to } x \in \overline{B}(0,r) \cap \mathrm{Fix}(T) \neq \varnothing \tag{55}$$

with a sufficiently large closed ball $\overline{B}(0,r)$. Note that Fact 12(II) is applicable to (55) *because $P_{\overline{B}(0,r)} \circ T$ is nonexpansive and satisfies* $\mathrm{Fix}(P_{\overline{B}(0,r)} \circ T) = \overline{B}(0,r) \cap \mathrm{Fix}(T)$ *(see [145, Proposition 1(d)]). Similar strategy will be utilized in the application to the hierarchical enhancement of the SVM in Section 4.2.*

(c) **(Conditions for Θ)** *The condition for $\Theta \in \Gamma_0(\mathscr{H})$ in* (53), *where it is required to have the Lipschitzian gradient $\nabla\Theta$, may not be restrictive as well for practitioners just by passing through the smooth regularizations, e.g., Moreau-Yosida regularization (see* (9) *and Fact 8 in Section 2.2).*

Remark 14 (On the Hybrid Steepest Descent Method)

(a) *The HSDM was established originally as a generalization of the so-called Halpern-type iteration (or anchor method) [6, 72, 90] for iteratively computing $P_{\mathrm{Fix}(T)}(x)$ for a nonexpansive operator $T : \mathscr{H} \to \mathscr{H}$ and $x \in \mathscr{H}$. Indeed, by choosing $\Psi(\cdot) := \frac{1}{2}\|\cdot - x\|^2$, the iteration* (54) *is reduced to the Halpern-type iteration.*

[11] ℓ^1_+ denotes the set of all summable nonnegative sequences. ℓ^2_+ denotes the set of all square-summable nonnegative sequences.

(b) *One can relax (L3) to* $\lim_{n\to\infty}\frac{\lambda_n}{\lambda_{n+1}}=1$ *in [140]. Moreover, if T is an averaged nonexpansive operator it was shown in [83] that only (W1) and (W2) are required to guarantee the strong convergence.*

(c) *The HSDM can be robustified against the numerical errors produced possibly in the computation of T [146].*

(d) *Parallel versions of the HSDM were developed in [129]. Specifically, convex optimization over the Cartesian product of the intersections of the fixed point sets of nonexpansive operators is considered, where strong convergence theorems are established under a certain contraction assumption with respect to the weighted maximum norm.*

(e) *The HSDM has been extended for the variational inequality problems over the fixed point set of certain class of quasi-nonexpansive operators including subgradient projection operators [145, 149] and has been applied to signal processing problems (see, e.g., [108, 149]).*

(f) *The mathematical properties of the HSDM, e.g., in [142, 145] have been studied extensively in various directions by many mathematicians (see, e.g., [27, 94] for extensions in Banach spaces).*

3 Hierarchical Convex Optimization with Proximal Splitting Operators

In this section, we present our central strategy for iterative approximation of the solution of the hierarchical convex optimization (10) by plugging proximal splitting operators into the HSDM. For simplicity, we focus on the DRS and the LAL operators as such proximal splitting operators.[12] Assume that Problem (10) has a solution, i.e., there exists at least one minimizer of Ψ over $\mathscr{S}_p$, and that (f,g,A) satisfies its qualification condition (32) (Note: The condition (32) holds automatically for many instances of (1), see, e.g., Section 4.2 [(125)] and Section 5.2 [Lemma 27 and (A.26)]). As explained briefly just around (22) in Section 1, for applications of the HSDM (54) to Problem (10), we need characterization of the constraint set as $\mathscr{S}_p=\Xi(\mathrm{Fix}(T))$ with a computable nonexpansive operator $T\colon\mathscr{H}\to\mathscr{H}$ and with a bounded linear operator $\Xi\in\mathscr{B}(\mathscr{H},\mathscr{X})$ which ensures the Gâteaux differentiability of $\Theta:=\Psi\circ\Xi\in\Gamma_0(\mathscr{H})$ with Lipschitzian gradient $\nabla\Theta$. In the following, we introduce three examples of such pair of computable nonexpansive operator $T\colon\mathscr{H}\to\mathscr{H}$ and $\Xi\in\mathscr{B}(\mathscr{H},\mathscr{X})$.

3.1 Plugging DRS Operators into Hybrid Steepest Descent Method

We introduce a nonexpansive operator called $\mathbf{T}_{\mathrm{DRS}_{\mathrm{I}}}$ of Type-I, as an instance of the DRS operator, that can characterize $\mathscr{S}_p$ (see (67)) and demonstrate how this nonexpansive operator can be plugged into the HSDM for (10).

Theorem 15 (HSDM with the DRS Operator in Product Space of Type-I). *Let $f\in\Gamma_0(\mathscr{X})$, $g\in\Gamma_0(\mathscr{K})$, and $A\in\mathscr{B}(\mathscr{X},\mathscr{K})$ in Problem* (10) *satisfy $\mathscr{S}_p\neq\varnothing$ and the qualification condition* (32)*. Suppose that $\Psi\in\Gamma_0(\mathscr{X})$ is Gâteaux differentiable with Lipschitzian gradient $\nabla\Psi$ over $\mathscr{X}$ and that $\Omega:=\operatorname*{argmin}_{x^\star\in\mathscr{S}_p}\Psi(x^\star)\neq\varnothing$. Then the operator*

$$\mathbf{T}_{\mathrm{DRS}_{\mathrm{I}}}\colon\mathscr{X}\times\mathscr{K}\to\mathscr{X}\times\mathscr{K}\colon(x,y)\mapsto(x_T,y_T),\tag{56}$$

where

[12] In [149, Section 17.5], the authors introduced briefly the central strategy of plugging the Douglas-Rachford splitting operator into the HSDM for hierarchical convex optimization. For applications of the HSDM to other proximal splitting operators, e.g., the forward-backward splitting operator [44], the primal-dual splitting operator [47, 139] for the hierarchical convex optimization of different types from (10), see [107, 149].

$$\left[\begin{array}{l} p = x - A^*(\mathrm{I}+AA^*)^{-1}(Ax-y) \\ (x_{1/2}, y_{1/2}) = (2p - x, 2Ap - y) \\ (x_T, y_T) = (2\operatorname{prox}_f(x_{1/2}) - x_{1/2}, 2\operatorname{prox}_g(y_{1/2}) - y_{1/2}), \end{array}\right. \tag{57}$$

can be plugged into the HSDM (54)*, with any* $\alpha \in (0,1)$ *and any* $(\lambda_{n+1})_{n\in\mathbb{N}} \in \ell_+^2 \setminus \ell_+^1$*, as*

$$\left[\begin{array}{l} (x_{n+1/2}, y_{n+1/2}) = (1-\alpha)(x_n, y_n) + \alpha \mathbf{T}_{\mathrm{DRS_I}}(x_n, y_n) \\ x_{n+1}^\star = x_{n+1/2} - A^*(\mathrm{I}+AA^*)^{-1}(Ax_{n+1/2} - y_{n+1/2}) \\ x_{n+1} = x_{n+1/2} - \lambda_{n+1}(\mathrm{I} - A^*(\mathrm{I}+AA^*)^{-1}A) \circ \nabla\Psi(x_{n+1}^\star) \\ y_{n+1} = y_{n+1/2} - \lambda_{n+1}((\mathrm{I}+AA^*)^{-1}A) \circ \nabla\Psi(x_{n+1}^\star). \end{array}\right. \tag{58}$$

The algorithm (58) *generates, for any* $(x_0, y_0) \in \mathscr{X} \times \mathscr{K}$*, a sequence* $(x_{n+1}^\star)_{n\in\mathbb{N}} \subset \mathscr{X}$ *which satisfies*

$$\lim_{n\to\infty} d_\Omega(x_n^\star) = 0 \tag{59}$$

if $\dim(\mathscr{X} \times \mathscr{K}) < \infty$ *and* $\operatorname{Fix}(\mathbf{T}_{\mathrm{DRS_I}})$ *is bounded.*

Remark 16 (Idea Behind the Derivation of Theorem 15)

(a) *The operator* $\mathbf{T}_{\mathrm{DRS_I}}$ *in* (56) *can be expressed as*[13]

$$\mathbf{T}_{\mathrm{DRS_I}} = (2\operatorname{prox}_F - \mathrm{I}) \circ (2\operatorname{prox}_{\iota_{\mathscr{N}(\check{A})}} - \mathrm{I}) = (2\operatorname{prox}_F - \mathrm{I}) \circ (2P_{\mathscr{N}(\check{A})} - \mathrm{I}) \tag{60}$$

which is nothing but the DRS operator in the sense of Proposition 9 (see Section 2.3) specialized for

$$\text{minimize } (F + \iota_{\mathscr{N}(\check{A})})(\mathscr{X} \times \mathscr{K}), \tag{61}$$

where

$$F\colon\ \mathscr{X} \times \mathscr{K} \to (-\infty, \infty]\colon (x,y) \mapsto f(x) + g(y), \tag{62}$$

$$\check{A}\colon\ \mathscr{X} \times \mathscr{K} \to \mathscr{K}\colon (x,y) \mapsto Ax - y, \tag{63}$$

and $\mathscr{N}(\check{A})$ *stands for the null space of* $\check{A} \in \mathscr{B}(\mathscr{X} \times \mathscr{K}, \mathscr{K})$*. Note that exactly in the same way as in* (11)*,* $\operatorname{prox}_F\colon \mathscr{X} \times \mathscr{K} \to \mathscr{X} \times \mathscr{K}\colon (x,y) \mapsto (\operatorname{prox}_f(x), \operatorname{prox}_g(y))$ *can be used as a computational tool if* prox_f *and* prox_g *are available. Moreover,* $\operatorname{prox}_{\iota_{\mathscr{N}(\check{A})}} = P_{\mathscr{N}(\check{A})}\colon \mathscr{X} \times \mathscr{K} \to \mathscr{N}(\check{A})\colon (x,y) \mapsto (p, Ap)$ *is also available if* p *in* (57) *is computable, hence Problem* (61) *is minimization of the sum of two proximable functions. Obviously, Problem* (61) *is a reformulation of Problem* (10) *in a higher dimensional space in the sense of*

$$\mathscr{S}_p\text{[in (10)]} = \mathscr{Q}_{\mathscr{X}}\left[\operatorname*{argmin}_{(x,y)\in\mathscr{X}\times\mathscr{K}} (F(x,y) + \iota_{\mathscr{N}(\check{A})}(x,y))\right],$$

where

$$\mathscr{Q}_{\mathscr{X}}\colon\ \mathscr{X} \times \mathscr{K} \to \mathscr{X}\colon (x,y) \mapsto x, \tag{64}$$

which is verified by

[13] The use of the DRS operator in a product space as in (60) is found explicitly or implicitly in various applications, mainly for solving (2) (see, e.g., [23, 41, 43, 59, 67, 68, 117]).

$$\begin{aligned}
&\operatorname{argmin}_{x\in\mathscr{X}} f(x)+g(Ax)\\
&= \mathscr{Q}_{\mathscr{X}}\left[\operatorname{argmin}_{(x,y)\in\mathscr{X}\times\mathscr{K}} f(x)+g(y)+\iota_{\{0\}}(Ax-y)\right]\\
&= \mathscr{Q}_{\mathscr{X}}\left[\operatorname{argmin}_{(x,y)\in\mathscr{X}\times\mathscr{K}} F(x,y)+\iota_{\mathscr{N}(\check{A})}(x,y)\right].
\end{aligned}$$

(b) *For application of the HSDM (based on Fact 12(II) in Section 2.4), Theorem 15 uses the convenient expression:*

$$\mathscr{S}_p\textit{[in }(10)\textit{]} \overset{\text{see below}}{=} \mathscr{Q}_{\mathscr{X}}(\operatorname{prox}_{\iota_{\mathscr{N}(\check{A})}}(\operatorname{Fix}(\mathbf{T}_{\mathrm{DRS_I}})) \tag{65}$$

$$= \mathscr{Q}_{\mathscr{X}}(P_{\mathscr{N}(\check{A})}(\operatorname{Fix}(\mathbf{T}_{\mathrm{DRS_I}})) \tag{66}$$

$$= \Xi_{\mathrm{DRS_I}}(\operatorname{Fix}(\mathbf{T}_{\mathrm{DRS_I}})) = \Xi_{\mathrm{DRS_I}}(\operatorname{Fix}((1-\alpha)\mathrm{I}+\alpha\mathbf{T}_{\mathrm{DRS_I}})) \tag{67}$$

in terms of attracting operator $(1-\alpha)\mathrm{I}+\alpha\mathbf{T}_{\mathrm{DRS_I}}$ *with* $\alpha\in(0,1)$ *(see* (34)*), where*

$$\Xi_{\mathrm{DRS_I}} := \mathscr{Q}_{\mathscr{X}}\circ P_{\mathscr{N}(\check{A})} \in \mathscr{B}(\mathscr{X}\times\mathscr{K},\mathscr{X}). \tag{68}$$

Note that the characterization (67) *is illustrated in Figure 3 (see Section 5.1) and is utilized, in Section 5.2, in the context of the hierarchical enhancement of Lasso. To prove* (65) *based on Proposition 9(a) in Section 2.3, we need:*

Claim 15: If $f\in\Gamma_0(\mathscr{X})$, $g\in\Gamma_0(\mathscr{K})$, and $A\in\mathscr{B}(\mathscr{X},\mathscr{K})$ in Problem (10) satisfy $\mathscr{S}_p\neq\varnothing$ and the qualification condition (32), we have

$$\operatorname{argmin}(F+\iota_{\mathscr{N}(\check{A})})(\mathscr{X}\times\mathscr{K})\neq\varnothing, \tag{69}$$

$$\operatorname{argmin}(F^*+\iota^*_{\mathscr{N}(\check{A})}\circ(-\mathrm{I}))(\mathscr{X}\times\mathscr{K})\neq\varnothing, \tag{70}$$

$$\min(F+\iota_{\mathscr{N}(\check{A})})(\mathscr{X}\times\mathscr{K}) = -\min(F^*+\iota^*_{\mathscr{N}(\check{A})}\circ(-\mathrm{I}))(\mathscr{X}\times\mathscr{K}). \tag{71}$$

Note that (69–71) correspond to (42–44) in Proposition 9 for minimization of $F+\iota_{\mathscr{N}(\check{A})}$ *and therefore Claim 15 is the main step in the proof of Theorem 15.*

(c) *To plug the operator* $\mathbf{T}_{\mathrm{DRS_I}}:\mathscr{H}\to\mathscr{H}$*, with* $\mathscr{H}:=\mathscr{X}\times\mathscr{K}$*, into the HSDM based on Fact 12(II) in Section 2.4, the characterization* $\mathscr{S}_p=\Xi_{\mathrm{DRS_I}}(\operatorname{Fix}((1-\alpha)\mathrm{I}+\alpha\mathbf{T}_{\mathrm{DRS_I}}))$ *in* (67) *is utilized in the translation [exactly in the same way as in* (22)*]:*

$$\Omega\textit{[in Theorem 15]} = \Xi_{\mathrm{DRS_I}}(\Omega_{\mathrm{DRS_I}}), \tag{72}$$

$$\textit{where } \Omega_{\mathrm{DRS_I}} := \operatorname*{argmin}_{\mathbf{z}\in\operatorname{Fix}(\mathbf{T}_{\mathrm{DRS_I}})} \Theta_{\mathrm{DRS_I}}(\mathbf{z}) = \operatorname*{argmin}_{\mathbf{z}\in\operatorname{Fix}((1-\alpha)\mathrm{I}+\alpha\mathbf{T}_{\mathrm{DRS_I}})} \Theta_{\mathrm{DRS_I}}(\mathbf{z}), \tag{73}$$

and $\Theta_{\mathrm{DRS_I}}=\Psi\circ\Xi_{\mathrm{DRS_I}}\in\Gamma_0(\mathscr{X}\times\mathscr{K})$*.*

(d) *Application of the HSDM to* (73) *yields*

$$\left[\begin{aligned}
\mathbf{z}_{n+1/2} &= [(1-\alpha)\mathrm{I}+\alpha\mathbf{T}_{\mathrm{DRS_I}}](\mathbf{z}_n),\\
\mathbf{z}_{n+1} &= \mathbf{z}_{n+1/2}-\lambda_{n+1}\nabla\Theta_{\mathrm{DRS_I}}(\mathbf{z}_{n+1/2})\\
&= \mathbf{z}_{n+1/2}-\lambda_{n+1}\Xi^*_{\mathrm{DRS_I}}\nabla\Psi(\Xi_{\mathrm{DRS_I}}\mathbf{z}_{n+1/2}),
\end{aligned}\right. \tag{74}$$

where $\Xi^*_{\mathrm{DRS_I}}$ *is the conjugate of* $\Xi_{\mathrm{DRS_I}}$ *in* (68)*. By letting* $\mathbf{z}_n=:(x_n,y_n)\in\mathscr{X}\times\mathscr{K}$*,* $\mathbf{z}_{n+1/2}=:(x_{n+1/2},y_{n+1/2})\in\mathscr{X}\times\mathscr{K}$*, and* $x^\star_{n+1}:=\Xi_{\mathrm{DRS_I}}\mathbf{z}_{n+1/2}\in\mathscr{X}$*, as well as, by noting*

$$\Xi^*_{\mathrm{DRS_I}} = P_{\mathscr{N}(\check{A})}\circ\mathscr{Q}^*_{\mathscr{X}}:\ \mathscr{X}\to\mathscr{X}\times\mathscr{K}: x\mapsto((\mathrm{I}-A^*(\mathrm{I}+AA^*)^{-1}A)x,(\mathrm{I}+AA^*)^{-1}Ax),$$

we can verify the equivalence between (74) *and* (58)*.*

(e) *Fact 12(II) in Section 2.4 guarantees* $\lim_{n\to\infty} d_{\Omega_{\mathrm{DRS_I}}}(\mathbf{z}_n) = 0$. *Moreover, by noting that* $\Xi_{\mathrm{DRS_I}} P_{\Omega_{\mathrm{DRS_I}}}(\mathbf{z}_{n+1/2}) \in \Omega$ *(see* (72)*) and* $\Omega_{\mathrm{DRS_I}} \subset \mathrm{Fix}((1-\alpha)\mathrm{I} + \alpha \mathbf{T}_{\mathrm{DRS_I}})$ *(see* (73)*),* (59) *is verified as*

$$\begin{aligned} & d_\Omega(x^\star_{n+1}) = d_\Omega(\Xi_{\mathrm{DRS_I}} \mathbf{z}_{n+1/2}) \\ \le\ & \|\Xi_{\mathrm{DRS_I}} \mathbf{z}_{n+1/2} - \Xi_{\mathrm{DRS_I}} P_{\Omega_{\mathrm{DRS_I}}}(\mathbf{z}_{n+1/2})\|_{\mathscr{X}} \\ \le\ & \|\Xi_{\mathrm{DRS_I}}\|_{\mathrm{op}} \|\mathbf{z}_{n+1/2} - P_{\Omega_{\mathrm{DRS_I}}}(\mathbf{z}_{n+1/2})\|_{\mathscr{H}} \\ \le\ & \|\Xi_{\mathrm{DRS_I}}\|_{\mathrm{op}} d_{\Omega_{\mathrm{DRS_I}}}(\mathbf{z}_n) \to 0 \ (n \to \infty). \end{aligned}$$

(The proof of Theorem 15 is given in Appendix C).

Next, we introduce another nonexpansive operator called $\mathbf{T}_{\mathrm{DRS_{II}}}$ of Type-II, as an instance of the DRS operator, that can characterize $\mathscr{S}_p$ (see (84)) and demonstrate how this nonexpansive operator can be plugged into the HSDM for (10). The operator $\mathbf{T}_{\mathrm{DRS_{II}}}$ is designed based on Example 7(b) in Section 2.2.

Theorem 17 (HSDM with the DRS Operator in Product Space of Type-II). *Let* $\mathscr{X} = \mathbb{R}^m$. *Let* $f \in \Gamma_0(\mathscr{X})$, $g = \bigoplus_{i=1}^m g_i \in \Gamma_0(\mathscr{K})$, $A\colon \mathscr{X} \to \mathscr{K} : x \mapsto Ax = (A_1 x, A_2 x, \ldots, A_m x)$ *with* $A_i \in \mathscr{B}(\mathscr{X}, \mathbb{R}) \setminus \{0\}$ $(i = 1, 2, \ldots, m)$ *in Problem* (10) *satisfy* $\mathscr{S}_p \neq \varnothing$ *and the qualification condition* (32). *Suppose that* $\Psi \in \Gamma_0(\mathscr{X})$ *is Gâteaux differentiable with Lipschitzian gradient* $\nabla\Psi$ *over* $\mathscr{X}$ *and that* $\Omega := \operatorname*{argmin}_{x^\star \in \mathscr{S}_p} \Psi(x^\star) \neq \varnothing$. *Then the operator*

$$\mathbf{T}_{\mathrm{DRS_{II}}} \colon \mathscr{X}^{m+1} \to \mathscr{X}^{m+1} : (x^{(1)}, x^{(2)}, \ldots, x^{(m+1)}) \mapsto (x_T^{(1)}, x_T^{(2)}, \ldots, x_T^{(m+1)}), \tag{75}$$

where

$$\left[\begin{array}{l} \bar{x} = \frac{1}{m+1} \sum_{j=1}^{m+1} x^{(j)} \\ x_T^{(i)} = (2\bar{x} - x^{(i)}) + 2(A_i A_i^*)^{-1} A_i^* (\mathrm{prox}_{(A_i A_i^*) g_i} [A_i(2\bar{x} - x^{(i)})] - A_i(2\bar{x} - x^{(i)})) \quad (i = 1, 2, \ldots, m) \\ x_T^{(m+1)} = 2\,\mathrm{prox}_f (2\bar{x} - x^{(m+1)}) - (2\bar{x} - x^{(m+1)}), \end{array}\right.$$

can be plugged into the HSDM (54)*, with any* $\alpha \in (0,1)$ *and any* $(\lambda_{n+1})_{n\in\mathbb{N}} \in \ell^2_+ \setminus \ell^1_+$, *as*

$$\left[\begin{array}{l} \left(x_{n+1/2}^{(1)}, \ldots, x_{n+1/2}^{(m+1)}\right) = (1-\alpha)\left(x_n^{(1)}, \ldots, x_n^{(m+1)}\right) + \alpha \mathbf{T}_{\mathrm{DRS_{II}}} \left(x_n^{(1)}, \ldots, x_n^{(m+1)}\right) \\ x^\star_{n+1} = \frac{1}{m+1} \sum_{j=1}^{m+1} x_{n+1/2}^{(j)} \\ x_{n+1}^{(i)} = x_{n+1/2}^{(i)} - \frac{\lambda_{n+1}}{m+1} \nabla\Psi(x^\star_{n+1}) \quad (i = 1, 2, \ldots, m+1). \end{array}\right. \tag{76}$$

The algorithm (76) *generates, for any* $\left(x_0^{(1)}, \ldots, x_0^{(m+1)}\right) \in \mathscr{X}^{m+1}$, *a sequence* $(x^\star_{n+1})_{n\in\mathbb{N}} \subset \mathscr{X}$ *which satisfies*

$$\lim_{n\to\infty} d_\Omega(x^\star_n) = 0 \tag{77}$$

if $\dim(\mathscr{X}) < \infty$ *and* $\mathrm{Fix}(\mathbf{T}_{\mathrm{DRS_{II}}})$ *is bounded.*

Remark 18 (Idea Behind the Derivation of Theorem 17)

(a) *The operator* $\mathbf{T}_{\mathrm{DRS_{II}}}$ *in* (75) *can be expressed as*

$$\mathbf{T}_{\mathrm{DRS_{II}}} = (2\,\mathrm{prox}_H - \mathrm{I}) \circ (2\,\mathrm{prox}_{\iota_D} - \mathrm{I}) = (2\,\mathrm{prox}_H - \mathrm{I}) \circ (2P_D - \mathrm{I}) \tag{78}$$

which is the DRS operator in the sense of Proposition 9 (see Section 2.3) specialized for

$$\text{minimize } (H + \iota_D)(\mathscr{X}^{m+1}), \tag{79}$$

where

$$H\colon \mathscr{X}^{m+1} \to (-\infty,\infty]\colon (x^{(1)},\dots,x^{(m+1)}) \mapsto \sum_{i=1}^{m} g_i(A_i x^{(i)}) + f(x^{(m+1)}), \tag{80}$$

$$D := \{(x^{(1)},\dots,x^{(m+1)}) \in \mathscr{X}^{m+1} \mid x^{(i)} = x^{(j)}\ (i,j = 1,2,\dots,m+1)\}. \tag{81}$$

Note that exactly in the same way as in (11)*,*

$$\begin{aligned} &\operatorname{prox}_H(x^{(1)},x^{(2)},\dots,x^{(m+1)}) \\ &= (\operatorname{prox}_{g_1\circ A_1}(x^{(1)}),\dots,\operatorname{prox}_{g_m\circ A_m}(x^{(m)}),\operatorname{prox}_f(x^{(m+1)})) \end{aligned}$$

can be used with (38)*, in Example 7(b) (see Section 2.2), as a computational tool if* prox_f *and* $\operatorname{prox}_{A_iA_i^*g}$ $(i = 1,2,\dots,m)$ *are available. Moreover,* $\operatorname{prox}_{\iota_D} = P_D\colon\ \mathscr{X}^{m+1} \to \mathscr{X}^{m+1}\colon (x^{(1)},x^{(2)},\dots,x^{(m+1)}) \mapsto (\bar{x},\dots,\bar{x})$ *with* $\bar{x} := \frac{1}{m+1}\sum_{i=1}^{m+1} x^{(i)}$ *is also available. Hence Problem* (79) *is minimization of the sum of two proximable functions (Note: Thanks to* $A_iA_i^* \in \mathbb{R}_{++} := \{r \in \mathbb{R} \mid r > 0\}$*, the computation of* $\mathbf{T}_{\mathrm{DRS}_{\mathrm{II}}}$ *in* (78) *does not require any matrix inversion). Obviously, Problem* (79) *is a reformulation of Problem* (10) *in a higher dimensional space in the sense of*

$$\mathscr{S}_p\textit{[in (10)]} = \mathscr{Q}_{\mathscr{X}^{(1)}}\left[\operatorname*{argmin}_{(x^{(1)},\dots,x^{(m+1)})\in\mathscr{X}^{m+1}} (H+\iota_D)(x^{(1)},\dots,x^{(m+1)}))\right],$$

where

$$\mathscr{Q}_{\mathscr{X}^{(1)}}\colon \mathscr{X}^{m+1} \to \mathscr{X}\colon (x^{(1)},\dots,x^{(m+1)}) \mapsto x^{(1)},$$

which is verified by

$$\begin{aligned} &\operatorname{argmin}_{x\in\mathscr{X}}\, g(Ax) + f(x) \\ &= \operatorname{argmin}_{x\in\mathscr{X}} \sum_{i=1}^{m} g_i(A_i x) + f(x) \\ &= \mathscr{Q}_{\mathscr{X}^{(1)}}\left[\operatorname{argmin}_{(x^{(1)},\dots,x^{(m+1)})\in\mathscr{X}^{m+1}} (H+\iota_D)(x^{(1)},\dots,x^{(m+1)})\right]. \end{aligned}$$

(b) *For application of the HSDM (based on Fact 12(II) in Section 2.4), Theorem 17 uses the convenient expression:*

$$\begin{aligned} &\mathscr{S}_p\textit{[in (10)]} \\ &\overset{\text{see below}}{=} \mathscr{Q}_{\mathscr{X}^{(1)}}(\operatorname{prox}_{\iota_D}(\operatorname{Fix}(\mathbf{T}_{\mathrm{DRS}_{\mathrm{II}}})) && (82) \\ &= \mathscr{Q}_{\mathscr{X}^{(1)}}(P_D(\operatorname{Fix}(\mathbf{T}_{\mathrm{DRS}_{\mathrm{II}}})) && (83) \\ &= \Xi_{\mathrm{DRS}_{\mathrm{II}}}(\operatorname{Fix}(\mathbf{T}_{\mathrm{DRS}_{\mathrm{II}}})) = \Xi_{\mathrm{DRS}_{\mathrm{II}}}(\operatorname{Fix}((1-\alpha)\mathrm{I}+\alpha\mathbf{T}_{\mathrm{DRS}_{\mathrm{II}}})) && (84) \end{aligned}$$

in terms of attracting operator $(1-\alpha)\mathrm{I}+\alpha\mathbf{T}_{\mathrm{DRS}_{\mathrm{II}}}$ *with* $\alpha \in (0,1)$ *(see* (34)*), where*

$$\Xi_{\mathrm{DRS}_{\mathrm{II}}} := \mathscr{Q}_{\mathscr{X}^{(1)}} \circ P_D \in \mathscr{B}(\mathscr{X}^{m+1},\mathscr{X}). \tag{85}$$

To prove (82) *based on Proposition 9(a) in Section 2.3, we need:*

Claim 17: If $\dim(\mathscr{K}) < \infty$, $f \in \Gamma_0(\mathscr{X})$, $g = \bigoplus_{i=1}^{m} g_i \in \Gamma_0(\mathscr{K})$, $A\colon\ \mathscr{X} \to \mathscr{K}\colon x \mapsto Ax = (A_1x, A_2x, \dots, A_mx)$ with $A_i \in \mathscr{B}(\mathscr{X},\mathbb{R}) \setminus \{0\}$ $(i = 1,2,\dots,m)$ in Problem (10) satisfy $\mathscr{S}_p \neq \varnothing$ and the qualification condition (32), we have

$$\operatorname{argmin}(H+\iota_D)(\mathscr{X}^{m+1}) \neq \varnothing, \tag{86}$$

$$\operatorname{argmin}(H^* + \iota_D^* \circ (-\mathrm{I}))(\mathscr{X}^{m+1}) \neq \varnothing, \tag{87}$$

$$\min(H+\iota_D)(\mathscr{X}^{m+1}) = -\min(H^* + \iota_D^* \circ (-\mathrm{I}))(\mathscr{X}^{m+1}). \tag{88}$$

Note that (86–88) correspond to (42–44) in Proposition 9 for minimization of $H+\iota_D$ and therefore Claim 17 is the main step in the proof of Theorem 17.

(c) *To plug the operator $\mathbf{T}_{\mathrm{DRS}_{\mathrm{II}}}: \mathscr{H} \to \mathscr{H}$, with $\mathscr{H} := \mathscr{X}^{m+1}$, into the HSDM based on Fact 12(II) in Section 2.4, the characterization $\mathscr{S}_p = \Xi_{\mathrm{DRS}_{\mathrm{II}}}(\mathrm{Fix}((1-\alpha)\mathrm{I}+\alpha\mathbf{T}_{\mathrm{DRS}_{\mathrm{II}}}))$ in* (84) *is utilized in the translation [exactly in the same way as in* (22)*]:*

$$\begin{aligned}&\Omega\text{[in Theorem 17]} = \Xi_{\mathrm{DRS}_{\mathrm{II}}}(\Omega_{\mathrm{DRS}_{\mathrm{II}}}),\\ &\text{where } \Omega_{\mathrm{DRS}_{\mathrm{II}}} := \operatorname*{argmin}_{\mathbf{X}\in\mathrm{Fix}(\mathbf{T}_{\mathrm{DRS}_{\mathrm{II}}})} \Theta_{\mathrm{DRS}_{\mathrm{II}}}(\mathbf{X}) = \operatorname*{argmin}_{\mathbf{X}\in\mathrm{Fix}((1-\alpha)\mathrm{I}+\alpha\mathbf{T}_{\mathrm{DRS}_{\mathrm{II}}})} \Theta_{\mathrm{DRS}_{\mathrm{II}}}(\mathbf{X}),\end{aligned} \tag{89}$$

and $\Theta_{\mathrm{DRS}_{\mathrm{II}}} = \Psi \circ \Xi_{\mathrm{DRS}_{\mathrm{II}}} \in \Gamma_0(\mathscr{X}^{m+1})$.

(d) *Application of the HSDM to* (89) *yields*

$$\left[\begin{aligned}\mathbf{X}_{n+1/2} &= [(1-\alpha)\mathrm{I}+\alpha\mathbf{T}_{\mathrm{DRS}_{\mathrm{II}}}](\mathbf{X}_n),\\ \mathbf{X}_{n+1} &= \mathbf{X}_{n+1/2} - \lambda_{n+1}\nabla\Theta_{\mathrm{DRS}_{\mathrm{II}}}(\mathbf{X}_{n+1/2})\\ &= \mathbf{X}_{n+1/2} - \lambda_{n+1}\Xi_{\mathrm{DRS}_{\mathrm{II}}}^*\nabla\Psi(\Xi_{\mathrm{DRS}_{\mathrm{II}}}\mathbf{X}_{n+1/2}),\end{aligned}\right. \tag{90}$$

where $\Xi_{\mathrm{DRS}_{\mathrm{II}}}^*$ is the conjugate of $\Xi_{\mathrm{DRS}_{\mathrm{II}}}$ in (85)*. By letting $\mathbf{X}_n =: (x_n^{(1)}, \ldots, x_n^{(m+1)}) \in \mathscr{X}^{m+1}$, $\mathbf{X}_{n+1/2} =: (x_{n+1/2}^{(1)}, \ldots, x_{n+1/2}^{(m+1)}) \in \mathscr{X}^{m+1}$, and $x_{n+1}^\star := \Xi_{\mathrm{DRS}_{\mathrm{II}}}\mathbf{X}_{n+1/2} \in \mathscr{X}$, as well as, by noting*

$$\Xi_{\mathrm{DRS}_{\mathrm{II}}}^* = P_D \circ \mathscr{Q}_{\mathscr{X}^{(1)}}^* : \mathscr{X} \to \mathscr{X}^{m+1} : x \mapsto \frac{1}{m+1}(x, x, \ldots, x),$$

we can verify the equivalence between (90) *and* (76)*.*

(e) *In the same way as in Remark 16(e), Fact 12(II) in Section 2.4 guarantees* $\lim_{n\to\infty} d_{\Omega_{\mathrm{DRS}_{\mathrm{II}}}}(\mathbf{X}_n) = 0$ *and* (77)*.*

(The proof of Theorem 17 is given in Appendix D).

3.2 Plugging LAL Operator into Hybrid Steepest Descent Method

We introduce a nonexpansive operator called $\mathbf{T}_{\mathrm{LAL}}$, as an instance of the LAL operator, that can characterize $\mathscr{S}_p$ (see (96)) and demonstrate how this nonexpansive operator can be plugged into the HSDM for (10). In particular, if $\nabla\Psi$ is strongly monotone over $\mathscr{X}$, $\mathbf{T}_{\mathrm{LAL}}$ can be plugged into the HSDM based on Fact 12(I) in Section 2.4, which results in a strongly convergent iterative algorithm for (10) (see Theorem 19). Of course, $\mathbf{T}_{\mathrm{LAL}}$ can also be plugged into the HSDM based on Fact 12(II) (see Theorem 21).

Theorem 19 (Strong Convergence Achieved by HSDM with LAL Operator). *Let $f \in \Gamma_0(\mathscr{X})$, $g \in \Gamma_0(\mathscr{K})$ and $A \in \mathscr{B}(\mathscr{X}, \mathscr{K})$ in Problem* (10) *satisfy not only $\mathscr{S}_p \neq \varnothing$ and the qualification condition* (32) *but also $\|\breve{A}\|_{\mathrm{op}} \leq \frac{1}{\mathrm{u}}$ $(\exists \mathrm{u} > 0)$ with $\breve{A}$ in* (63)*. Suppose also that $\Psi \in \Gamma_0(\mathscr{X})$ is Gâteaux differentiable with Lipschitzian as well as strongly monotone gradient $\nabla\Psi$ over $\mathscr{X}$. Then the operator*

$$\begin{aligned}\mathbf{T}_{\mathrm{LAL}} :\ & \mathscr{X}\times\mathscr{K}\times\mathscr{K} \to \mathscr{X}\times\mathscr{K}\times\mathscr{K}\\ :\ & \begin{pmatrix} x\\ y\\ \nu\end{pmatrix} \mapsto \begin{pmatrix} x_T\\ y_T\\ \nu_T\end{pmatrix} = \begin{pmatrix}\mathrm{prox}_f(x-\mathrm{u}^2(A^*Ax-A^*y)+\mathrm{u}A^*\nu)\\ \mathrm{prox}_g(y-\mathrm{u}^2(-Ax+y)-\mathrm{u}\nu)\\ \nu-\mathrm{u}(Ax_T-y_T)\end{pmatrix}\end{aligned} \tag{91}$$

can be plugged into HSDM (54)*, with any* $\alpha \in (0,1]$ *and any* $\eta_{xy}, \eta_\nu > 0$*, as*

$$\left[\begin{array}{l} (x_{n+1/2}, y_{n+1/2}, \nu_{n+1/2}) = (1-\alpha)(x_n, y_n, \nu_n) + \alpha \mathbf{T}_{\mathrm{LAL}}(x_n, y_n, \nu_n) \\ x_{n+1} = x_{n+1/2} - \lambda_{n+1}(\nabla \Psi(x_{n+1/2}) + \eta_{xy} A^*(Ax_{n+1/2} - y_{n+1/2})) \\ y_{n+1} = y_{n+1/2} + \lambda_{n+1}\eta_{xy}(Ax_{n+1/2} - y_{n+1/2}) \\ \nu_{n+1} = \nu_{n+1/2} - \lambda_{n+1}\eta_\nu \nu_{n+1/2}. \end{array}\right. \tag{92}$$

The algorithm (92) *generates, for any* $(x_0, y_0, \nu_0) \in \mathscr{X} \times \mathscr{K} \times \mathscr{K}$*, a sequence* $(x_n)_{n\in\mathbb{N}} \subset \mathscr{X}$ *which converges strongly to the uniquely existing solution of Problem (10) if* $(\lambda_{n+1})_{n\in\mathbb{N}} \subset [0,\infty)$ *satisfies conditions (W*1*–W*3*) [or* $(\lambda_{n+1})_{n\in\mathbb{N}} \subset (0,\infty)$ *satisfies (L*1*–L*3*)] in Fact 12(I) in Section 2.4.*

Remark 20 (Idea Behind the Derivation of Theorem 19)

(a) *The operator* $\mathbf{T}_{\mathrm{LAL}}$ *in* (91) *can be expressed as*

$$(\mathbf{z}, \nu) \mapsto (\mathbf{z}_T, \nu_T) \textit{ with } \begin{cases} \mathbf{z}_T = \mathrm{prox}_F(\mathbf{z} - (\mathfrak{u}\check{A})^*(\mathfrak{u}\check{A})\mathbf{z} + (\mathfrak{u}\check{A})^*\nu) \\ \nu_T = \nu - \mathfrak{u}\check{A}\mathbf{z}_T \end{cases} \tag{93}$$

by introducing $\mathbf{z} := (x,y)$ *and* $\mathbf{z}_T := (x_T, y_T)$*, which is the LAL operator of Proposition 10 (see Section 2.3) specialized for*

$$\text{minimize } (F + \iota_{\{0\}} \circ (\mathfrak{u}\check{A}))(\mathscr{X} \times \mathscr{K}), \tag{94}$$

where F and $\check{A}$ *are defined, respectively, in* (62) *and in* (63)*. Note that exactly in the same way as in* (11)*,* $\mathrm{prox}_F \colon \mathscr{X} \times \mathscr{K} \to \mathscr{X} \times \mathscr{K} \colon (x,y) \mapsto (\mathrm{prox}_f(x), \mathrm{prox}_g(y))$ *can be used as a computational tool if* prox_f *and* prox_g *are available. Obviously, Problem* (94) *is a reformulation of Problem* (10) *in a higher dimensional space in the sense of*

$$\mathscr{S}_p\textit{[in (10)]} = \mathscr{Q}_{\mathscr{X}}\left[\underset{(x,y)\in\mathscr{X}\times\mathscr{K}}{\mathrm{argmin}} \; F(x,y) + \iota_{\{0\}}(\mathfrak{u}\check{A}(x,y))\right],$$

where $\mathscr{Q}_{\mathscr{X}}$ *is defined as in* (64)*, which is verified by*

$$\begin{aligned} \mathscr{S}_p &= \mathrm{argmin}_{x\in\mathscr{X}} f(x) + g(Ax) \\ &= \mathscr{Q}_{\mathscr{X}}\left[\mathrm{argmin}_{(x,y)\in\mathscr{X}\times\mathscr{K}} f(x) + g(y) + \iota_{\{0\}}(Ax - y)\right] \\ &= \mathscr{Q}_{\mathscr{X}}\left[\mathrm{argmin}_{(x,y)\in\mathscr{X}\times\mathscr{K}} F(x,y) + \iota_{\{0\}}(\mathfrak{u}\check{A}(x,y))\right]. \end{aligned}$$

(b) *For application of the HSDM (based on Fact 12(I) in Section 2.4), Theorem 19 uses the convenient expression:*

$$\begin{aligned} &\mathscr{S}_p\textit{[in (10)]} \\ &= \mathscr{Q}_{\mathscr{X}} \circ \mathscr{Q}_{\mathscr{X}\times\mathscr{K}}\left[\mathrm{argmin}(F + \iota_{\{0\}} \circ (\mathfrak{u}\check{A}))(\mathscr{X}\times\mathscr{K}) \times \mathrm{argmin}(F^* \circ (\mathfrak{u}\check{A})^*)(\mathscr{K})\right] \\ &\overset{\text{see below}}{=} \mathscr{Q}_{\mathscr{X}} \circ \mathscr{Q}_{\mathscr{X}\times\mathscr{K}}(\mathrm{Fix}(\mathbf{T}_{\mathrm{LAL}})) && (95) \\ &= \Xi_{\mathrm{LAL}}(\mathrm{Fix}(\mathbf{T}_{\mathrm{LAL}})) = \Xi_{\mathrm{LAL}}(\mathrm{Fix}((1-\alpha)\mathrm{I} + \alpha\mathbf{T}_{\mathrm{LAL}})) && (96) \end{aligned}$$

in terms of nonexpansive operator $(1-\alpha)\mathrm{I} + \alpha\mathbf{T}_{\mathrm{LAL}}$ *with* $\alpha \in (0,1]$ *(Note: The nonexpansiveness of* $\mathbf{T}_{\mathrm{LAL}}$ *is ensured by Proposition 10(b) in Section 2.3 with* $\|\mathfrak{u}\check{A}\|_{\mathrm{op}} \leq 1$*)*

$$\begin{aligned} &\mathscr{Q}_{\mathscr{X}\times\mathscr{K}} \colon \mathscr{X}\times\mathscr{K}\times\mathscr{K} \to \mathscr{X}\times\mathscr{K} \colon (x,y,\nu) \mapsto (x,y) \\ &\Xi_{\mathrm{LAL}} := \mathscr{Q}_{\mathscr{X}} \circ \mathscr{Q}_{\mathscr{X}\times\mathscr{K}} \in \mathscr{B}(\mathscr{X}\times\mathscr{K}\times\mathscr{K}, \mathscr{X}), && (97) \end{aligned}$$

where $\mathscr{Q}_{\mathscr{X}}$ is defined as in (64). *To prove* (95) *based on Proposition 10(a) in Section 2.3, we need:*

Claim 19: If $f \in \Gamma_0(\mathscr{X})$, $g \in \Gamma_0(\mathscr{K})$ and $A \in \mathscr{B}(\mathscr{X},\mathscr{K})$ in Problem (10) satisfy not only $\mathscr{S}_p \neq \varnothing$ and the qualification condition (32) but also $\|\breve{A}\|_{\mathrm{op}} \leq \frac{1}{\mathrm{u}}$ $(\exists \mathrm{u} > 0)$ with $\breve{A}$ in (63), we have

$$\operatorname{argmin}(F + \iota_{\{0\}} \circ (\mathrm{u}\breve{A}))(\mathscr{X} \times \mathscr{K}) \neq \varnothing, \tag{98}$$

$$\operatorname{argmin}(F^* \circ (\mathrm{u}\breve{A})^*)(\mathscr{K}) \neq \varnothing, \tag{99}$$

$$\min(F + \iota_{\{0\}} \circ (\mathrm{u}\breve{A}))(\mathscr{X} \times \mathscr{K}) = -\min(F^* \circ (\mathrm{u}\breve{A})^*)(\mathscr{K}). \tag{100}$$

Note that (98–100) correspond to (47–49) in Proposition 10 for minimization of $F + \iota_{\{0\}} \circ (\mathrm{u}\breve{A})$ and therefore Claim 19 is the main step in the proof of Theorem 19. In Claim 19, we also remark that $(\mathrm{u}\breve{A})^*$ in (99) *is the conjugate of $\mathrm{u}\breve{A}$ and given by*

$$(\mathrm{u}\breve{A})^* \colon \mathscr{K} \to \mathscr{X} \times \mathscr{K} \colon \nu \mapsto (\mathrm{u}A^*\nu, -\mathrm{u}\nu). \tag{101}$$

(c) *To plug the operator $\mathbf{T}_{\mathrm{LAL}} \colon \mathscr{H} \to \mathscr{H}$, with $\mathscr{H} := \mathscr{X} \times \mathscr{K} \times \mathscr{K}$, into the HSDM based on Fact 12(I) in Section 2.4, the characterization $\mathscr{S}_p = \Xi_{\mathrm{LAL}}(\mathrm{Fix}((1-\alpha)\mathrm{I} + \alpha\mathbf{T}_{\mathrm{LAL}}))$ in* (96) *is utilized in the translation:*

$$\Omega \text{\textit{[in Theorem 19]}} = \Xi_{\mathrm{LAL}}(\Omega_{\mathrm{LAL}}^{\mathrm{reg}}), \tag{102}$$

$$\text{\textit{where }} \Omega_{\mathrm{LAL}}^{\mathrm{reg}} := \operatorname*{argmin}_{\mathbf{w} \in \mathrm{Fix}(\mathbf{T}_{\mathrm{LAL}})} \Theta_{\mathrm{LAL}}^{\mathrm{reg}}(\mathbf{w}) = \operatorname*{argmin}_{\mathbf{w} \in \mathrm{Fix}((1-\alpha)\mathrm{I} + \alpha\mathbf{T}_{\mathrm{LAL}})} \Theta_{\mathrm{LAL}}^{\mathrm{reg}}(\mathbf{w}), \tag{103}$$

$$\begin{aligned} \Theta_{\mathrm{LAL}}^{\mathrm{reg}} &\colon \mathscr{X} \times \mathscr{K} \times \mathscr{K} \to \mathbb{R} \\ &\colon \mathbf{w}_\star \mapsto \Psi(\Xi_{\mathrm{LAL}}\mathbf{w}_\star) + \frac{\eta_{xy}}{2}\|\breve{A} \circ \mathscr{Q}_{\mathscr{X}\times\mathscr{K}}\mathbf{w}_\star\|_{\mathscr{K}}^2 + \frac{\eta_\nu}{2}\|\mathscr{Q}_{\mathscr{K}}\mathbf{w}_\star\|_{\mathscr{K}}^2, \end{aligned}$$

for $\eta_{xy}, \eta_\nu > 0$ with $\mathscr{Q}_{\mathscr{K}} \colon \mathscr{X} \times \mathscr{K} \times \mathscr{K} \colon (x,y,\nu) \mapsto \nu$. Note that, since $\nabla\Psi$ is strongly monotone over $\mathscr{X}$, the gradient $\nabla\Theta_{\mathrm{LAL}}^{\mathrm{reg}}$ is strongly monotone over $\mathscr{X} \times \mathscr{K} \times \mathscr{K}$ (for the proof, see [150, Theorem 2(d)]).

(d) *Application of the HSDM to* (103) *yields*

$$\left[\begin{aligned} \mathbf{w}_{n+1/2} &= [(1-\alpha)\mathrm{I} + \alpha\mathbf{T}_{\mathrm{LAL}}](\mathbf{w}_n) \\ \mathbf{w}_{n+1} &= \mathbf{w}_{n+1/2} - \lambda_{n+1}\nabla\Theta_{\mathrm{LAL}}^{\mathrm{reg}}(\mathbf{w}_{n+1/2}). \end{aligned}\right. \tag{104}$$

By letting $\mathbf{w}_n := (x_n, y_n, \nu_n) \in \mathscr{X} \times \mathscr{K} \times \mathscr{K}$ and $\mathbf{w}_{n+1/2} := (x_{n+1/2}, y_{n+1/2}, \nu_{n+1/2}) \in \mathscr{X} \times \mathscr{K} \times \mathscr{K}$, as well as, by noting

$$\begin{aligned} \Xi_{\mathrm{LAL}}^* = \mathscr{Q}_{\mathscr{X}\times\mathscr{K}}^* \circ \mathscr{Q}_{\mathscr{X}}^* &\colon \mathscr{X} \to \mathscr{X} \times \mathscr{K} \times \mathscr{K} \colon x \mapsto (x,0,0), \\ (\breve{A} \circ \mathscr{Q}_{\mathscr{X}\times\mathscr{K}})^* &\colon \mathscr{K} \to \mathscr{X} \times \mathscr{K} \times \mathscr{K} \colon y \mapsto (A^*y, -y, 0), \\ \mathscr{Q}_{\mathscr{K}}^* &\colon \mathscr{K} \to \mathscr{X} \times \mathscr{K} \times \mathscr{K} \colon \nu \mapsto (0,0,\nu), \end{aligned} \tag{105}$$

we can verify the equivalence between (104) *and* (92).

(e) *Fact 12(I) in Section 2.4 guarantees that $(\mathbf{w}_n)_{n\in\mathbb{N}}$ converges strongly to a point in $\Omega_{\mathrm{LAL}}^{\mathrm{reg}}$. Hence, $(\Xi_{\mathrm{LAL}}\mathbf{w}_n(= x_n))_{n\in\mathbb{N}}$ also converges strongly to a point in Ω (see* (102)*).*

(The proof of Theorem 19 is given in Appendix E).

Theorem 21 (HSDM with the LAL Operator Based on Fact 12(II)). *Let $f \in \Gamma_0(\mathscr{X})$, $g \in \Gamma_0(\mathscr{K})$ and $A \in \mathscr{B}(\mathscr{X},\mathscr{K})$ in Problem* (10) *satisfy not only $\mathscr{S}_p \neq \varnothing$ and the qualification condition* (32) *but also $\|\breve{A}\|_{\mathrm{op}} \leq \frac{1}{\mathrm{u}}$ $(\exists \mathrm{u} > 0)$ with $\breve{A}$ in* (63). *Suppose also that $\Psi \in \Gamma_0(\mathscr{X})$ is Gâteaux differentiable with Lipschitzian gradient $\nabla\Psi$ over $\mathscr{X}$ and that $\Omega := \operatorname*{argmin}_{x^\star \in \mathscr{S}_p} \Psi(x^\star) \neq \varnothing$. Then the operator $\mathbf{T}_{\mathrm{LAL}}$ in* (91) *can be plugged into HSDM* (54), *with any $\alpha \in (0,1)$ and any $(\lambda_{n+1})_{n\in\mathbb{N}} \in \ell_+^2 \setminus \ell_+^1$, as*

$$\left[\begin{array}{l}(x_{n+1/2}, y_{n+1}, \nu_{n+1}) = (1-\alpha)(x_n^\star, y_n, \nu_n) + \alpha \mathbf{T}_{\mathrm{LAL}}(x_n^\star, y_n, \nu_n) \\ x_{n+1}^\star = x_{n+1/2} - \lambda_{n+1} \nabla \Psi(x_{n+1/2}).\end{array}\right. \tag{106}$$

The algorithm (106) *generates, for any* $(x_0^\star, y_0, \nu_0) \in \mathscr{X} \times \mathscr{K} \times \mathscr{K}$*, a sequence* $(x_n^\star)_{n\in\mathbb{N}} \subset \mathscr{X}$ *which satisfies*

$$\lim_{n\to\infty} d_\Omega(x_n^\star) = 0 \tag{107}$$

if $\dim(\mathscr{X} \times \mathscr{K} \times \mathscr{K}) < \infty$ *and* $\mathrm{Fix}(\mathbf{T}_{\mathrm{LAL}})$ *is bounded.*

Remark 22 (Idea Behind the Derivation of Theorem 21)

(a) *Following Remark 20(a)(b), we obtain the characterization*

$$\mathscr{S}_p\textit{[in (10)]} = \Xi_{\mathrm{LAL}}(\mathrm{Fix}((1-\alpha)\mathrm{I} + \alpha \mathbf{T}_{\mathrm{LAL}})),$$

in (96) *(see also* Ξ_{LAL} *in* (97)*), with the attracting operator* $(1-\alpha)\mathrm{I} + \alpha \mathbf{T}_{\mathrm{LAL}}$ *for* $\alpha \in (0,1)$ *(see* (34)*). This characterization is utilized, to plug* $\mathbf{T}_{\mathrm{LAL}} : \mathscr{H} \to \mathscr{H}$ *(*$\mathscr{H} := \mathscr{X} \times \mathscr{K} \times \mathscr{K}$*) into the HSDM based on Fact 12(II) in Section 2.4, in the translation [see also* (22)*]:*

$$\begin{aligned}&\Omega\textit{[in Theorem 21]} = \Xi_{\mathrm{LAL}}(\Omega_{\mathrm{LAL}}), \\ &\textit{where } \Omega_{\mathrm{LAL}} := \operatorname*{argmin}_{\mathbf{w}\in\mathrm{Fix}(\mathbf{T}_{\mathrm{LAL}})} \Theta_{\mathrm{LAL}}(\mathbf{w}) = \operatorname*{argmin}_{\mathbf{w}\in\mathrm{Fix}((1-\alpha)\mathrm{I}+\alpha\mathbf{T}_{\mathrm{LAL}})} \Theta_{\mathrm{LAL}}(\mathbf{w})\end{aligned} \tag{108}$$

and $\Theta_{\mathrm{LAL}} := \Psi \circ \Xi_{\mathrm{LAL}} \in \Gamma_0(\mathscr{X} \times \mathscr{K} \times \mathscr{K})$.

(b) *Application of the HSDM to* (108) *yields*

$$\left[\begin{array}{l}\mathbf{w}_{n+1/2} = [(1-\alpha)\mathrm{I} + \alpha \mathbf{T}_{\mathrm{LAL}}](\mathbf{w}_n), \\ \mathbf{w}_{n+1} = \mathbf{w}_{n+1/2} - \lambda_{n+1} \nabla \Theta_{\mathrm{LAL}}(\mathbf{w}_{n+1/2}) \\ \qquad\ = \mathbf{w}_{n+1/2} - \lambda_{n+1} \Xi_{\mathrm{LAL}}^* \nabla \Psi(\Xi_{\mathrm{LAL}} \mathbf{w}_{n+1/2}),\end{array}\right. \tag{109}$$

where Ξ_{LAL}^* *is the conjugate of* Ξ_{LAL}*. By letting* $\mathbf{w}_n =: (x_n^\star, y_n, \nu_n) \in \mathscr{X} \times \mathscr{K} \times \mathscr{K}$ *and* $\mathbf{w}_{n+1/2} =: (x_{n+1/2}, y_{n+1/2}, \nu_{n+1/2}) \in \mathscr{X} \times \mathscr{K} \times \mathscr{K}$*, as well as, by noting* (105)*, we can verify the equivalence between* (109) *and* (106).

(c) *In the same way as in Remark 16(e), Fact 12(II) in Section 2.4 guarantees* $\lim_{n\to\infty} d_{\Omega_{\mathrm{LAL}}}(\mathbf{w}_n) = 0$ *and* (107).

(The proof of Theorem 21 is omitted, see Remark 22).

3.3 Conditions for Boundedness of Fixed Point Sets of DRS and LAL Operators

In Theorems 15, 17, and 21, the boundednesses of $\mathrm{Fix}(\mathbf{T}_{\mathrm{DRS_I}})$, $\mathrm{Fix}(\mathbf{T}_{\mathrm{DRS_{II}}})$, and $\mathrm{Fix}(\mathbf{T}_{\mathrm{LAL}})$ are required for the algorithms (58), (76), and (106) to produce $(x_{n+1}^\star)_{n\in\mathbb{N}}$ satisfying $\lim_{n\to\infty} d_\Omega(x_n^\star) = 0$. Theorem 23 below presents sufficient conditions for the boundednesses of these fixed point sets. Corollary 24 below presents a stronger condition which will be used in Section 5.2 to guarantee the boundedness of $\mathrm{Fix}(\mathbf{T}_{\mathrm{DRS_I}})$ in the context of the hierarchical enhancement of Lasso.

Theorem 23. *Let* $f \in \Gamma_0(\mathscr{X})$*,* $g \in \Gamma_0(\mathscr{K})$ *and* $A \in \mathscr{B}(\mathscr{X}, \mathscr{K})$ *in Problem* (10) *satisfy* $\mathscr{S}_p \neq \varnothing$ *and the qualification condition* (32)*. Let* $(A^*)^{-1} : \mathscr{X} \to 2^{\mathscr{K}} : x \mapsto \{y \in \mathscr{K} \mid x = A^* y\}$*. Then we have*

(a) $\mathscr{S}_p$*,* $\partial f(\mathscr{S}_p)$*, and* $\bigcup_{x\in\mathscr{S}_p} \left([-(A^*)^{-1}(\partial f(x))] \cap \partial g(Ax)\right) \subset \mathscr{K}$ *are bounded*

$$\Rightarrow\ \mathrm{Fix}\left(\mathbf{T}_{\mathrm{DRS_I}}\right) \subset \mathscr{X} \times \mathscr{K} \textit{ in Theorem 15 is bounded.}$$

(b) *If* $\check{A}$ *in* (63) *satisfies* $\|\check{A}\|_{\mathrm{op}} \leq \frac{1}{\mathrm{u}}$ $(\exists \mathrm{u} > 0)$*, then*

$$\mathscr{S}_p \textit{ and } \bigcup_{x\in\mathscr{S}_p} \left([-(A^*)^{-1}(\partial f(x))]\cap \partial g(Ax)\right) \subset \mathscr{K} \textit{ are bounded}$$
$$\Rightarrow\ \mathrm{Fix}\,(\mathbf{T}_{\mathrm{LAL}}) \subset \mathscr{X}\times\mathscr{K}\times\mathscr{K} \textit{ in Theorem 21 is bounded.}$$

(c) *If, in particular,* $\mathscr{K}=\mathbb{R}^m$, $g=\bigoplus_{i=1}^m g_i \in \Gamma_0(\mathbb{R}^m)$, $A\colon\ \mathscr{X}\to\mathbb{R}^m\colon x\mapsto Ax=(A_1x,A_2x,\ldots,A_mx)$ *with* $A_i\in\mathscr{B}(\mathscr{X},\mathbb{R})\setminus\{0\}$ $(i=1,2,\ldots,m)$ *in Problem* (10)*, then*
$$\mathscr{S}_p \textit{ and } \bigcup_{x\in\mathscr{S}_p}\left(\left[\bigtimes_{j=1}^m A_j^*\partial g_j(A_jx)\right]\times[\partial f(x)\cap(-\textstyle\sum_{i=1}^m A_i^*\partial g_i(A_ix))]\right) \textit{ are bounded}$$
$$\Rightarrow\ \mathrm{Fix}\,(\mathbf{T}_{\mathrm{DRS_{II}}}) \subset \mathscr{X}^{m+1} \textit{ in Theorem 17 is bounded.}$$

(The proof of Theorem 23 is given in Appendix F.)

The following simple relations

$$\begin{aligned}&\textstyle\bigcup_{x\in\mathscr{S}_p}\left([-(A^*)^{-1}(\partial f(x))]\cap\partial g(Ax)\right) \text{ [in Theorem 23(a)(b)]}\\ &\quad\subset\left[(-(A^*)^{-1}(\partial f(\mathscr{S}_p)))\cap\partial g(\mathscr{K})\right] \text{ and}\\ &\textstyle\bigcup_{x\in\mathscr{S}_p}\left(\left[\bigtimes_{j=1}^m A_j^*\partial g_j(A_jx)\right]\times[\partial f(x)\cap(-\sum_{i=1}^m A_i^*\partial g_i(A_ix))]\right) \text{ [in Theorem 23(c)]}\\ &\quad\subset\left[\textstyle\bigtimes_{j=1}^m A_j^*\partial g_j(\mathbb{R})\right]\times[-\textstyle\sum_{i=1}^m A_i^*\partial g_i(\mathbb{R})]\end{aligned}$$

lead to the corollary below.

Corollary 24 *Let* $f\in\Gamma_0(\mathscr{X})$, $g\in\Gamma_0(\mathscr{K})$, $A\in\mathscr{B}(\mathscr{X},\mathscr{K})$ *in Problem* (10)*, and* $\breve{A}$ *in* (63) *satisfy* $\mathscr{S}_p\neq\varnothing$*, the qualification condition* (32)*, and* $\|\breve{A}\|_{\mathrm{op}}\leq\frac{1}{\mathfrak{u}}$ $(\exists\mathfrak{u}>0)$*. Then we have*
(a) $\mathscr{S}_p, \partial f(\mathscr{S}_p)$ *and* $(-(A^*)^{-1}(\partial f(\mathscr{S}_p)))\cap\partial g(\mathscr{K})$ *are bounded*

$$\Rightarrow\ \left[\begin{array}{l}\mathrm{Fix}\,(\mathbf{T}_{\mathrm{DRS_I}})\subset\mathscr{X}\times\mathscr{K} \textit{ in Theorem 15 is bounded;}\\ \mathrm{Fix}\,(\mathbf{T}_{\mathrm{LAL}})\subset\mathscr{X}\times\mathscr{K}\times\mathscr{K} \textit{ in Theorem 21 is bounded.}\end{array}\right.$$

(b) *If, in particular,* $\mathscr{K}=\mathbb{R}^m$, $g=\bigoplus_{i=1}^m g_i\in\Gamma_0(\mathbb{R}^m)$, $A\colon\ \mathscr{X}\to\mathbb{R}^m\colon x\mapsto Ax=(A_1x,A_2x,\ldots,A_mx)$ *with* $A_i\in\mathscr{B}(\mathscr{X},\mathbb{R})\setminus\{0\}$ $(i=1,2,\ldots,m)$ *in Problem* (10)*, then*

$$\mathscr{S}_p \textit{ and } \partial g_i(\mathbb{R})\ (i=1,2,\ldots,m) \textit{ are bounded}$$
$$\Rightarrow\ \mathrm{Fix}\,(\mathbf{T}_{\mathrm{DRS_{II}}})\subset\mathscr{X}^{m+1} \textit{ in Theorem 17 is bounded.}$$

4 Application to Hierarchical Enhancement of Support Vector Machine

4.1 Support Vector Machine

Consider a supervised learning problem for estimating a binary function

$$\mathfrak{L}:\mathbb{R}^p\to\{-1,1\}$$

with a given training dataset $\mathscr{D}:=\{(\mathbf{x}_i,y_i)\in\mathbb{R}^p\times\{-1,1\}\mid i=1,2,\ldots,N\}$, where y_i is a possibly corrupted version of the label $\mathfrak{L}(\mathbf{x}_i)$ of the point $\mathbf{x}_i$. The *support vector machine* (SVM) has been recognized as one of the most successful supervised machine learning algorithms for such a learning problem. For simplicity, we focus on the linear SVM because the nonlinear SVM exploiting the so-called *Kernel trick* can be viewed as an instance of the linear classifiers in the *Reproducing Kernel Hilbert Spaces* (RKHS).

The dataset $\mathscr{D}$ is said to be *linearly separable* if there exists $(\mathbf{w},b)\in(\mathbb{R}^p\setminus\{\mathbf{0}\})\times\mathbb{R}$ defining a $(p-1)$-dimensional hyperplane

$$\Pi_{(\mathbf{w},b)}:=\{\mathbf{x}\in\mathbb{R}^p\mid\mathbf{w}^\top\mathbf{x}-b=0\}=\Pi_{t(\mathbf{w},b)}\quad(\forall t>0)\tag{110}$$

which satisfies

$$\left.\begin{array}{l}\{\mathbf{x}_i \in \mathbb{R}^p \mid (\mathbf{x}_i, 1) \in \mathscr{D}\} \subset \Pi^{+}_{(\mathbf{w},b)} := \{\mathbf{x} \in \mathbb{R}^p \mid \mathbf{w}^\top \mathbf{x} - b > 0\} \\ \{\mathbf{x}_i \in \mathbb{R}^p \mid (\mathbf{x}_i, -1) \in \mathscr{D}\} \subset \Pi^{-}_{(\mathbf{w},b)} := \{\mathbf{x} \in \mathbb{R}^p \mid \mathbf{w}^\top \mathbf{x} - b < 0\}\end{array}\right\}. \tag{111}$$

In such a case, the so-called linear classifier is defined as a mapping

$$\mathscr{L}_{(\mathbf{w},b)} : \mathbb{R}^p \to \{-1,1\} : \mathbf{x} \mapsto \begin{cases} +1 & \text{if } \mathbf{x} \in \Pi^{+}_{(\mathbf{w},b)}, \\ -1 & \text{if } \mathbf{x} \in \Pi^{-}_{(\mathbf{w},b)}, \end{cases} \tag{112}$$

which is hopefully a good approximation of the function $\mathfrak{L}$ observed partially through the training dataset $\mathscr{D}$. If $\mathscr{D}$ is linearly separable, there also exists infinitely many $(\mathbf{w}, b) \in \mathbb{R}^p \times \mathbb{R}$ satisfying

$$\left.\begin{array}{l}\mathscr{D}_+ := \{\mathbf{x}_i \in \mathbb{R}^p \mid (\mathbf{x}_i, 1) \in \mathscr{D}\} \subset \Pi^{\geq 1}_{(\mathbf{w},b)} := \{\mathbf{x} \in \mathbb{R}^p \mid \mathbf{w}^\top \mathbf{x} - b \geq 1\} \\ \mathscr{D}_- := \{\mathbf{x}_i \in \mathbb{R}^p \mid (\mathbf{x}_i, -1) \in \mathscr{D}\} \subset \Pi^{\leq -1}_{(\mathbf{w},b)} := \{\mathbf{x} \in \mathbb{R}^p \mid \mathbf{w}^\top \mathbf{x} - b \leq -1\}\end{array}\right\}, \tag{113}$$

which is confirmed by rescaling $(\mathbf{w}, b) \in \mathbb{R}^p \times \mathbb{R}$ in (110) with a constant $t \geq 1/\min\{|\mathbf{w}^\top \mathbf{x}_i - b|\}_{i=1}^N > 0$.

The half-spaces $\Pi^{\geq 1}_{(\mathbf{w},b)}$ and $\Pi^{\leq -1}_{(\mathbf{w},b)}$ defined in (113) are main players in the following consideration on the linear classifier $\mathscr{L}_{(\mathbf{w},b)}$ even for linearly non-separable data $\mathscr{D}$. In this paper, the *margin* of the linear classifier $\mathscr{L}_{(\mathbf{w},b)}$ in (112) is defined by

$$\frac{1}{2}\operatorname{dist}\left(\Pi^{\geq 1}_{(\mathbf{w},b)}, \Pi^{\leq -1}_{(\mathbf{w},b)}\right) = \frac{1}{2} \min_{\mathbf{x}_+ \in \Pi^{\geq 1}_{(\mathbf{w},b)}, \mathbf{x}_- \in \Pi^{\leq -1}_{(\mathbf{w},b)}} \|\mathbf{x}_+ - \mathbf{x}_-\| = \frac{1}{\|\mathbf{w}\|}.$$

By using the function h in (39) and

$$(\forall \mathbf{z} \in \mathbb{R}^p) \quad \left[\begin{array}{l} \mathrm{d}\left(\mathbf{z}, \Pi^{\geq 1}_{(\mathbf{w},b)}\right) = \begin{cases} \frac{|\mathbf{w}^\top \mathbf{z} - b - 1|}{\|\mathbf{w}\|} = \frac{1 - (\mathbf{w}^\top \mathbf{z} - b)}{\|\mathbf{w}\|} & \text{if } \mathbf{z} \notin \Pi^{\geq 1}_{(\mathbf{w},b)}, \\ 0 & \text{otherwise,} \end{cases} \\ \mathrm{d}\left(\mathbf{z}, \Pi^{\leq -1}_{(\mathbf{w},b)}\right) = \begin{cases} \frac{|\mathbf{w}^\top \mathbf{z} - b + 1|}{\|\mathbf{w}\|} = \frac{1 + (\mathbf{w}^\top \mathbf{z} - b)}{\|\mathbf{w}\|} & \text{if } \mathbf{z} \notin \Pi^{\leq -1}_{(\mathbf{w},b)}, \\ 0 & \text{otherwise,} \end{cases} \end{array}\right.$$

we deduce

$$\|\mathbf{w}\| \left[\sum_{\mathbf{z} \in \mathscr{D}_+} \mathrm{d}\left(\mathbf{z}, \Pi^{\geq 1}_{(\mathbf{w},b)}\right) + \sum_{\mathbf{z} \in \mathscr{D}_-} \mathrm{d}\left(\mathbf{z}, \Pi^{\leq -1}_{(\mathbf{w},b)}\right)\right] = \sum_{i=1}^N h\left(y_i \left(\mathbf{w}^\top \mathbf{x}_i - b\right)\right) \tag{114}$$

which clarifies the geometric interpretation of "the *empirical hinge loss* of $\mathscr{L}_{(\mathbf{w},b)}$" defined in the right hand side of (114) and ensures

$$\text{Condition (113)} \Leftrightarrow \sum_{i=1}^N h\left(y_i \left(\mathbf{w}^\top \mathbf{x}_i - b\right)\right) = 0. \tag{115}$$

For linearly separable data $\mathscr{D}$, among all $\mathscr{L}_{(\mathbf{w},b)}$ satisfying (113), the *Support Vector Machine (SVM)* $\mathscr{L}_{(\mathbf{w}^\star,b^\star)}$ was proposed in 1960s by Vapnik (see, e.g.,[135, 136]) as a special linear classifier which achieves maximal margin, i.e.,

$$\frac{1}{2}\operatorname{dist}\left(\Pi^{\geq 1}_{(\mathbf{w}^\star,b^\star)}, \Pi^{\leq -1}_{(\mathbf{w}^\star,b^\star)}\right) = \max_{(\mathbf{w},b) \text{ satisfies (115)}} \frac{1}{2}\operatorname{dist}\left(\Pi^{\geq 1}_{(\mathbf{w},b)}, \Pi^{\leq -1}_{(\mathbf{w},b)}\right). \tag{116}$$

Therefore the SVM $\mathscr{L}_{(\mathbf{w}^\star,b^\star)}$ for linearly separable $\mathscr{D}$ is given as the solution of the following convex optimization problem:

$$\text{minimize } \|\mathbf{w}\|^2 \text{subject to } \sum_{i=1}^{N} h\left(y_i\left(\mathbf{w}^\top \mathbf{x}_i - b\right)\right) = 0 \tag{117}$$

$$\Updownarrow$$

$$\text{minimize } \|\mathbf{w}\|^2 \text{subject to } (\mathbf{w},b) \in \operatorname*{argmin}_{(\hat{\mathbf{w}},\hat{b})\in\mathbb{R}^p\times\mathbb{R}} \sum_{i=1}^{N} h\left(y_i\left(\hat{\mathbf{w}}^\top \mathbf{x}_i - \hat{b}\right)\right), \tag{118}$$

where the last equivalence holds true under the linear separability of $\mathscr{D}$ because of the nonnegativity of h in (39).

The SVM defined equivalently in (116) or (117) or (118) for linearly separable training data has been extended for applications to even possibly linearly nonseparable training data $\mathscr{D}$ where the existence of $(\mathbf{w},b) \in \mathbb{R}^p \times \mathbb{R}$ satisfying (113) is no longer guaranteed. One of the most widely accepted extensions of (118) is known as *the soft margin hyperplane* [14, 25, 48, 73] which is characterized as a solution to the optimization problem:

$$\text{minimize, w.r.t. } (\mathbf{w},b), \quad \frac{1}{2}\|\mathbf{w}\|^2 + \mathfrak{C}\sum_{i=1}^{N} h\left(y_i(\mathbf{w}^\top \mathbf{x}_i - b)\right) \tag{119}$$

or equivalently

$$\begin{aligned} &\text{minimize, w.r.t. } (\mathbf{w},b,\xi), \qquad \frac{1}{2}\|\mathbf{w}\|^2 + \mathfrak{C}\sum_{i=1}^{N} \xi_i \\ &\text{subject to } y_i\left(\mathbf{w}^\top \mathbf{x} - b\right) \geq 1 - \xi_i \text{ and } \xi_i \geq 0 \ (i = 1,2,\ldots,N), \end{aligned} \tag{120}$$

where $\mathfrak{C} > 0$ is a tuning parameter, and ξ_i $(i = 1,2,\ldots,N)$ are slack variables.

Along the Cover's theorem (on the capacity of a space in linear dichotomies) [49], saying that the probability of any grouping of the points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N, \in \mathbb{R}^l$, in general position, into two classes to be linearly separable tends to unity as $l \to \infty$, another extension of the strategy $\mathscr{L}_{(\mathbf{w}^\star,b^\star)}$ of (116) into higher dimensional spaces was made in [16, 49], for application to possibly linearly nonseparable training data $\mathscr{D}$, by passing through a certain nonlinear transform $\mathfrak{N}\colon \mathbb{R}^p \to \mathbb{R}^l$ $(l \gg p)$ of the original training data $\mathscr{D} := \{(\mathbf{x}_i, y_i) \in \mathbb{R}^p \times \{-1,1\} \mid i = 1,2,\ldots,N\}$ to $\mathfrak{D} := \{(\mathfrak{N}(\mathbf{x}_i), y_i) \in \mathbb{R}^l \times \{-1,1\} \mid i = 1,2,\ldots,N\}$, where the nonlinear transform $\mathfrak{N}$ is defined usually in terms of kernel built in the theory of *the Reproducing Kernel Hilbert Space (RKHS)* [2, 124, 126] for exploiting the so-called kernel trick.

4.2 Optimal Margin Classifier with Least Empirical Hinge Loss

As suggested in [48, Section 3], the original goal behind the soft margin hyperplane in (119) or (120) seems to determine $(\mathbf{w}^{\star\star}, b^{\star\star}) \in (\mathbb{R}^p \setminus \{\mathbf{0}\}) \times \mathbb{R}$ as the solution of the following nonconvex hierarchical optimization:

$$\text{minimize } \frac{1}{2}\|\mathbf{w}^\star\|^2 \tag{121}$$

$$\text{subject to } (\mathbf{w}^\star, b^\star) \in \operatorname*{argmin}_{(\mathbf{w},b)} |\mathscr{E}(\mathbf{w},b)|, \tag{122}$$

where $|\cdot|$ stands for the cardinality of a set and $\mathscr{E}(\mathbf{w},b) \subset \mathscr{D}_+ \cup \mathscr{D}_-$ is the training error set defined as

$$\mathscr{E}(\mathbf{w},b) := \left\{\mathbf{z}\in\mathscr{D}_+ \mid \mathrm{d}\left(\mathbf{z},\Pi^{\geq 1}_{(\mathbf{w},b)}\right) > 0\right\} \cup \left\{\mathbf{z}\in\mathscr{D}_- \mid \mathrm{d}\left(\mathbf{z},\Pi^{\leq -1}_{(\mathbf{w},b)}\right) > 0\right\}, \tag{123}$$

i.e., determining a special hyperplane $(\mathbf{w}^{\star\star}, b^{\star\star})$, which achieves maximal margin in the set $\operatorname{argmin}_{(\mathbf{w},b)} |\mathscr{E}(\mathbf{w},b)|$, is desired.

Unfortunately, since the problem to determine $(\mathbf{w}^{\star}, b^{\star})$ in (122) is in general NP-hard [15, 48, 84] and since (114) implies that

$$(122) \Leftrightarrow (\mathbf{w}^{\star}, b^{\star}) \in \operatorname*{argmin}_{(\mathbf{w},b)} \sum_{i=1}^{N} \left[\lim_{\sigma\downarrow 0} h^{\sigma}\left(y_i(\mathbf{w}^{\top}\mathbf{x}_i - b)\right)\right],$$

the original goal set in (121–123) was replaced in [48, Section3] by a realistic goal (119) [or (120)] for a sufficiently large constant $\mathfrak{C} > 0$. However, unlike the desired solution of (121–123), the soft margin hyperplane in (119) applied to linearly separable data has no guarantee to reproduce the original SVM in (116).

The above observations induce a natural question:
Is the solution of (119) for general training data really a mathematically sound extension of the original SVM defined equivalently in (116) or (117) or (118) specialized for linearly separable training data ?[14]

Clearly, this question comes from essentially common concern as seen in Scenario 1, therefore, an alternative natural extension of the original SVM in (118) would be the solution of the optimization problem:

$$\text{minimize } \frac{1}{2}\|\mathbf{w}^{\star}\|^2 \text{ subject to } (\mathbf{w}^{\star}, b^{\star}) \in \Gamma := \operatorname*{argmin}_{(\mathbf{w},b)\in\mathbb{R}^p\times\mathbb{R}} \sum_{i=1}^{N} h\left(y_i(\mathbf{w}^{\top}\mathbf{x}_i - b)\right) \tag{124}$$

which does not seem different from (118) at a glance but is defined even possibly for linearly nonseparable training data $\mathscr{D}$. Remark that the hierarchical convex optimization problem (124) is a more faithful convex relaxation of (121–123) than the convex optimization (119) [or its equivalent formulation (120) with slack variables.[15]] for the soft margin hyperplane. This is because the solution of (124) for linearly separable data certainly reproduces the original SVM in (116). As remarked in Example 1(b) in Section 1, in general, the soft margin hyperplane via (119) for a fixed constant $\mathfrak{C} > 0$ does not achieve the hierarchical optimality in the sense of (124). Fortunately, Problem (124) falls in the class of the hierarchical convex optimization problems of type (10).

In the following, we demonstrate how Problem (124) can be solved by a proposed strategy in Section 3. Let $\mathscr{X} := \mathbb{R}^p \times \mathbb{R}$, $A_i\colon \mathscr{X} \to \mathbb{R}\colon (\mathbf{w},b) \mapsto y_i(\mathbf{x}_i^{\top}\mathbf{w} - b)$ $(i = 1,2,\ldots,N)$, $f\colon \mathscr{X} \to \mathbb{R}\colon (\mathbf{w},b) \mapsto h(A_N(\mathbf{w},b))$, $g := \bigoplus_{i=1}^{N-1} h$, and $A\colon \mathbb{R}^{p+1} \to \mathbb{R}^{N-1}\colon (\mathbf{w},b) \mapsto (A_i(\mathbf{w},b))_{i=1}^{N-1}$. By using these translations, we can express the hinge loss function, in the form of the first stage cost function of (10), as

$$\sum_{i=1}^{N} h\left(y_i(\mathbf{w}^{\top}\mathbf{x}_i - b)\right) = g \circ A(\mathbf{w},b) + f(\mathbf{w},b)$$

[14] This question is common even for the soft margin SVM applied to the transformed data $\mathfrak{D}$ employed in [16] because the linear separability of $\mathfrak{D}$ is not always guaranteed.

[15] In terms of slack variables, Problem (124) can also be restated as

$$\begin{aligned}
&\text{minimize } \frac{1}{2}\|\mathbf{w}^{\star}\|^2 \\
&\text{subject to } (\mathbf{w}^{\star}, b^{\star}, \xi^{\star}) \in \operatorname*{argmin}_{(\mathbf{w},b,\xi)\in\mathbb{R}^p\times\mathbb{R}\times\mathbb{R}^N} \sum_{i=1}^{N} \left[\xi_i + \iota_{S_i}(\mathbf{w},b,\xi) + \iota_{\mathfrak{g}_i}(\mathbf{w},b,\xi)\right], \\
&\text{where } S_i := \left\{(\mathbf{w},b,\xi) \in \mathbb{R}^p\times\mathbb{R}\times\mathbb{R}^N \mid y_i\left(\mathbf{w}^{\top}\mathbf{x}_i - b\right) \geq 1 - \xi_i\right\} \\
&\text{and } \mathfrak{g}_i := \left\{(\mathbf{w},b,\xi) \in \mathbb{R}^p\times\mathbb{R}\times\mathbb{R}^N \mid \xi_i \geq 0\right\} \quad (i = 1,2,\ldots,N).
\end{aligned}$$

and its associated qualification condition (see (32)) is verified by

$$\operatorname{ri}(\operatorname{dom}(g) - A\operatorname{dom}(f)) = \operatorname{ri}\left(\mathbb{R}^{N-1} - A\operatorname{dom}(f)\right) = \operatorname{ri}\left(\mathbb{R}^{N-1}\right) = \mathbb{R}^{N-1} \ni 0. \tag{125}$$

Note that, for any $\gamma \in \mathbb{R}_{++}$, the proximity operator $\operatorname{prox}_{\gamma h}$ can be computed as (40) in Example 7(c) (see Section 2.2) and therefore $\operatorname{prox}_f = \operatorname{prox}_{h\circ A_N}$ can also be computed by applying (38) and (40) in Example 7(b)(c). Moreover, by introducing $\Psi\colon \mathbb{R}^p \times \mathbb{R} \to \mathbb{R}\colon (\mathbf{w},b) \mapsto \frac{1}{2}\|\mathbf{w}\|^2$, we can regard Problem (124) as an instance of Problem (10) under the assumption of $\mathscr{S}_p := \Gamma \neq \varnothing$.[16] In fact, we can apply Theorem 15, Theorem 17, and Theorem 21 to (124) because Ψ is not strictly convex. In the following numerical experiment, we applied Theorem 17 to (124) with slight modification[16].

4.3 Numerical Experiment: Margin Maximization with Least Empirical Hinge Loss

We demonstrate that, as an extension of the original SVM in (116), the hierarchical enhancement of the SVM in (124) is more faithful to the original SVM than the soft margin SVM (119). In our experiment, we applied the original SVM in (116), the soft margin SVM in (119), and the proposed hierarchical enhancement of the SVM in (124) to the Iris dataset which is famous dataset used firstly in Fisher's paper [65]. This data set has 150 sample points, which are divided into three classes (I(setosa), II(versicolor), III(virginica)), and each sample point has four features (sepal length, sepal width, petal length, and petal width). From Iris dataset, we construct two datasets: separable $\mathscr{D}_{\mathrm{sep}} \subset \mathbb{R}^2 \times \{-1,1\}$ with $|\mathscr{D}_{\mathrm{sep}}| = 100$ comprising all the samples of Class I and Class II having only sepal length and sepal width; and non-separable $\mathscr{D}_{\mathrm{nsep}} \subset \mathbb{R}^2 \times \{-1,1\}$ with $|\mathscr{D}_{\mathrm{nsep}}| = 100$ comprising all the samples of Class II and Class III having only petal length and petal width. For each linear classifier $\mathscr{L}_{(\mathbf{w},b)}$ of our interest, the three hyperplanes $\Pi_{(\mathbf{w},b)}, \Pi_{(\mathbf{w},b+1)}, \Pi_{(\mathbf{w},b-1)}$ (see (111)) are drawn in Figure 1 and Figure 2, in cyan for "Original SVM", in green for "Soft Margin SVM", and in magenta for the proposed "M^{2}LEHL" (which stands for *the Margin Maximization with Least Empirical Hinge Loss*), respectively, where $(\mathbf{w},b)_{\mathrm{org}}$ is obtained by applying a quadratic programming solver `quadprog` in Matlab to (116), $(\mathbf{w},b)_{\mathrm{soft}}$ is obtained by applying a soft margin SVM solver `fitcsvm` (with the default setting, i.e., $\mathfrak{C} = 1$) in Matlab to (119), and $(\mathbf{w},b)_{\mathrm{M^2LEHL}}$ is obtained by applying the proposed algorithmic solution in Section 4.2, to (124), designed based on Theorem 17 with slight modification[16].

Figure 1 illustrates the resulting separating hyperplanes for the separable dataset $\mathscr{D}_{\mathrm{sep}}$. Since the magenta lines are completely overlapped with cyan lines, "M^{2}LEHL" reproduces "Original SVM", as explained in just after (124). "Soft Margin SVM" does not succeed in maximizing the margin, i.e., (124) is a more faithful extension of the original SVM in (116) than the soft margin SVM (119).

Figure 2 illustrates the resulting separating hyperplanes for the nonseparable dataset $\mathscr{D}_{\mathrm{nsep}}$. Since the original SVM (116) has no solution, "Original SVM" is not depicted. As the performance measure, we employ the number of errors $|\mathscr{E}(\cdot)|$ defined in (123) along the original goal (121) (as suggested in [48, Sec. 3]). Though "Soft Margin SVM" has 21 errors, "M^{2}LEHL" achieves only 6 errors, which demonstrates that (124) is more effective formulation for approaching to the original goal (121) than the soft margin SVM (119).

[16] If we need to guarantee $\mathscr{S}_p$[in (10)] $\neq \varnothing$, we recommend the following slight modification of (124):

$$\underset{\mathbf{w}^\star \in \widetilde{\Gamma}}{\text{minimize}}\ \frac{1}{2}\|\mathbf{w}^\star\|^2 \text{ subject to } \widetilde{\Gamma} := \underset{(\mathbf{w},b)\in\mathbb{R}^p\times\mathbb{R}}{\operatorname{argmin}} \left[\Phi(\mathbf{w},b) := \iota_{\overline{B}(0,r)}(\mathbf{w},b) + \sum_{i=1}^{N} h\left(y_i(\mathbf{w}^\top \mathbf{x}_i - b)\right)\right]$$

with a sufficiently large closed ball $\overline{B}(0,r)$, where $\mathscr{S}_p := \widetilde{\Gamma} \neq \varnothing$ is guaranteed due to the coercivity of Φ. Fortunately, our strategies in Section 3 are still applicable to this modified problem because it is also an instance of (8) which can be translated into (10) as explained in Section 1. In the application of Theorem 17 in Section 3.1 to this modification, the boundedness of $\operatorname{Fix}(\mathbf{T}_{\mathrm{DRS_{II}}})$ is automatically guaranteed because of Corollary 24(b) (see Section 3.3) and the boundedness of both $\widetilde{\Gamma} \subset \overline{B}(0,r)$ and $\partial h(\mathbb{R}) = \partial h(\mathbb{R}\setminus\{1\}) \cup \partial h(\{1\}) = [-1,0]$.

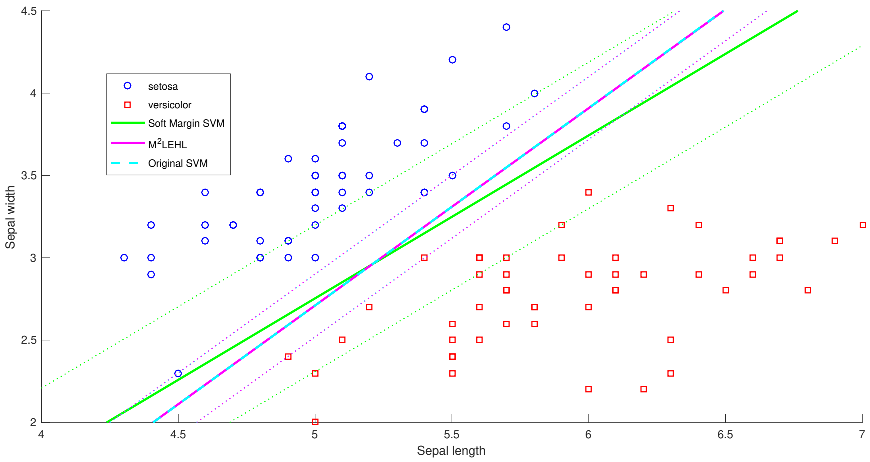


Fig. 1 Comparison between M^{2}LEHL, Original SVM, and Soft Margin SVM (Case of a separable training dataset $\mathscr{D}_{\mathrm{sep}}$): M^{2}LEHL reproduces Original SVM.

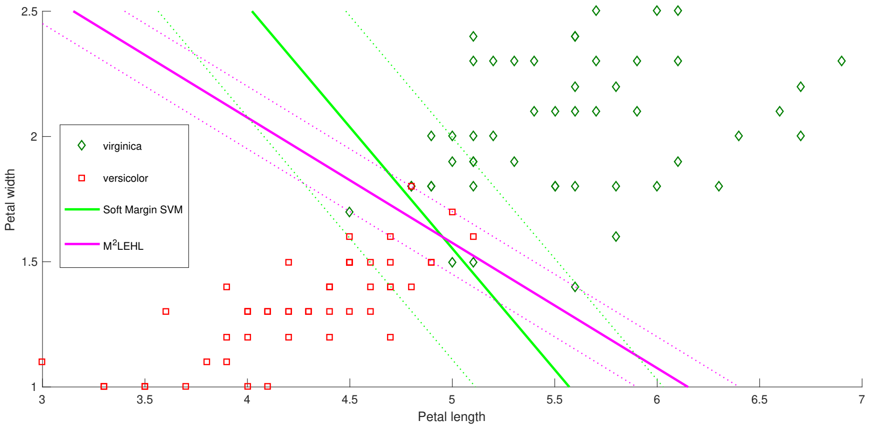


Fig. 2 Comparison between M^{2}LEHL and Soft Margin SVM (Case of a nonseparable training dataset $\mathscr{D}_{\mathrm{nsep}}$).

5 Application to Hierarchical Enhancement of Lasso

5.1 *TREX : A Nonconvex Automatic Sparsity Control of Lasso*

Consider the estimation of a sparse vector $\mathbf{b}^{\mathrm{tru}} \in \mathbb{R}^p$ in the standard linear regression model:

$$\mathbf{z} = \mathbf{X}\mathbf{b}^{\mathrm{tru}} + \sigma\mathbf{e}, \tag{126}$$

where $\mathbf{z} = (z_1, \ldots, z_N)^\top \in \mathbb{R}^N$ is a response vector, $\mathbf{X} \in \mathbb{R}^{N\times p}$ a design matrix, $\sigma > 0$ a constant, $\mathbf{e} = (\varepsilon_1, \ldots, \varepsilon_N)^\top$ the noise vector, each ε_i is the realization of a random variable with mean zero and variance 1.

The Lasso (Least Absolute Shrinkage and Selection Operator) [132] has been used widely as one of the most well-known sparsity aware statistical estimation methods [73, 74]. The Lasso for (126) is defined as a minimizer of the least squares criterion with ℓ_1 penalty, i.e.,

$$\mathbf{b}_{\mathrm{Lasso}}(\lambda) \in \operatorname*{argmin}_{\mathbf{b}\in\mathbb{R}^p} \frac{1}{2N}\|\mathbf{z}-\mathbf{X}\mathbf{b}\|_2^2+\lambda\|\mathbf{b}\|_1,$$

where the tuning parameter $\lambda > 0$ aims at controlling the sparsity of $\mathbf{b}_{\mathrm{Lasso}}(\lambda)$. However selection of $\lambda > 0$ is highly influential to $\mathbf{b}_{\mathrm{Lasso}}(\lambda)$ and therefore its reliable way of selection has been strongly desired. Among many efforts toward automatic sparsity control of Lasso, the following prediction bound offers a firm basis and has been applied widely in recent strategies including [50, 69, 77, 89].

Fact 25 *(A Prediction Bound of Lasso [87, 120]) For* $\lambda \geq \frac{2\|\mathbf{X}^\top(\mathbf{z}-\mathbf{X}\mathbf{b}^{\mathrm{tru}})\|_\infty}{N}$*, it holds* $\frac{\|\mathbf{X}\mathbf{b}_{\mathrm{Lasso}}(\lambda)-\mathbf{X}\mathbf{b}^{\mathrm{tru}}\|^2}{N} \leq 2\lambda\|\mathbf{b}^{\mathrm{tru}}\|_1$.

The TREX (Tuning-free Regression that adapts to the Entire noise $\sigma\mathbf{e}$ and the design matrix $\mathbf{X}$) [89] is one of the state-of-the-art strategies based on Fact 25. The TREX is defined as a solution of a nonconvex optimization problem:

$$\text{find } \mathbf{b}_{\mathrm{TREX}} \in \operatorname*{argmin}_{\mathbf{b}\in\mathbb{R}^p} \frac{\|\mathbf{X}\mathbf{b}-\mathbf{z}\|^2}{\left\|\mathbf{X}^\top(\mathbf{X}\mathbf{b}-\mathbf{z})\right\|_\infty}+\beta\|\mathbf{b}\|_1, \tag{127}$$

where $\left\|\mathbf{X}^\top(\mathbf{X}\mathbf{b}-\mathbf{z})\right\|_\infty = \max_{1\leq j\leq p}\left|\mathbf{X}_{:j}^\top(\mathbf{X}\mathbf{b}-\mathbf{z})\right|$, $\mathbf{X}_{:j}$ denotes the jth column of $\mathbf{X}$, and the parameter β can be set to a constant value ($\beta = 1/2$ being the default choice).

The authors in [13] cleverly decomposed the nonconvex optimization (127) into $2p$ subproblems:

$$\text{find } \mathbf{b}_{\mathrm{TREX}}^{(j)} \in \operatorname*{argmin}_{\substack{\mathbf{b}\in\mathbb{R}^p\\ \mathbf{x}_j^\top(\mathbf{X}\mathbf{b}-\mathbf{z})>0}} \left[\frac{\|\mathbf{X}\mathbf{b}-\mathbf{z}\|^2}{\beta\mathbf{x}_j^\top(\mathbf{X}\mathbf{b}-\mathbf{z})}+\|\mathbf{b}\|_1\right], \tag{128}$$

where

$$\mathbf{x}_j = \begin{cases}\mathbf{X}_{:j} & (j=1,2,\ldots,p);\\ -\mathbf{X}_{:j-p} & (j=p+1,p+2,\ldots,2p).\end{cases} \tag{129}$$

More precisely, $\mathbf{b}_{\mathrm{TREX}}$ in (127) is characterized as

$$\mathbf{b}_{\mathrm{TREX}} \in \widehat{\mathscr{Q}}_{\mathbb{R}^p}\left[\operatorname*{argmin}_{\substack{(\mathbf{b},j)\in\mathbb{R}^p\times\{1,2,\ldots,2p\}\\ \mathbf{x}_j^\top(\mathbf{X}\mathbf{b}-\mathbf{z})>0}}\left(\frac{\|\mathbf{X}\mathbf{b}-\mathbf{z}\|^2}{\beta\mathbf{x}_j^\top(\mathbf{X}\mathbf{b}-\mathbf{z})}+\|\mathbf{b}\|_1\right)\right], \tag{130}$$

where

$$\widehat{\mathscr{Q}}_{\mathbb{R}^p}: \mathbb{R}^p\times\{1,2,\ldots,2p\}\to\mathbb{R}^p: (\mathbf{b},j)\mapsto\mathbf{b}. \tag{131}$$

Remarkably, each subproblem (128) was shown to be a convex optimization and solved in [13] with a second-order cone program (SOCP) [92].

Recently, for sound extensions of the subproblem (128) as well as for sound applications of proximal splitting, a successful reformulation of (130) was made for general $q > 1$ in [38] as

$$\mathbf{b}_{\mathrm{TREX_q}} \in \mathscr{S}_{\mathrm{TREX_q}} := \widehat{\mathscr{Q}}_{\mathbb{R}^p}\left[\operatorname*{argmin}_{(\mathbf{b},j)\in\mathbb{R}^p\times\{1,2,\ldots,2p\}} g_{(j,q)}(\mathbf{M}_j\mathbf{b})+\|\mathbf{b}\|_1\right] \tag{132}$$

whose solution $\mathbf{b}_{\mathrm{TREX_q}}$ is given, by passing through $2p$ convex subproblems, as $\mathbf{b}_{\mathrm{TREX_q}}^{(j^\star)}$, where

$$\left[\begin{array}{l}\mathbf{b}^{(j)}_{\mathrm{TREX_q}} \in \mathscr{S}_{(j,q)} := \operatorname*{argmin}_{\mathbf{b}\in\mathbb{R}^p}\left[g_{(j,q)}(\mathbf{M}_j\mathbf{b}) + \|\mathbf{b}\|_1\right] \ (j = 1,2,\ldots,2p); \\ j^\star \in \operatorname*{argmin}_{j\in\{1,2,\ldots,2p\}}\left[g_{(j,q)}(\mathbf{M}_j\mathbf{b}^{(j)}_{\mathrm{TREX_q}}) + \|\mathbf{b}^{(j)}_{\mathrm{TREX_q}}\|_1\right],\end{array}\right. \tag{133}$$

$$g_{(j,q)} : \mathbb{R}\times\mathbb{R}^N \to (-\infty,\infty] : (\eta,\mathbf{y}) \mapsto \begin{cases} \frac{\|\mathbf{y}-\mathbf{z}\|^q}{\beta(\eta-\mathbf{x}_j^\top\mathbf{z})^{q-1}}, & \text{if } \eta > \mathbf{x}_j^\top\mathbf{z}; \\ 0, & \text{if } \mathbf{y}=\mathbf{z} \text{ and } \eta = \mathbf{x}_j^\top\mathbf{z}; \\ +\infty, & \text{otherwise}\end{cases} \tag{134}$$

is a proper lower semicontinuous convex function, and

$$\mathbf{M}_j : \mathbb{R}^p \to \mathbb{R}\times\mathbb{R}^N : \mathbf{b} \mapsto \left(\mathbf{x}_j^\top\mathbf{X}\mathbf{b}, \mathbf{X}\mathbf{b}\right) \tag{135}$$

is a bounded linear operator. The estimator $\mathbf{b}_{\mathrm{TREX_q}}$ in (132) is called the generalized TREX in [38] where, as its specialization, $\mathbf{b}_{\mathrm{TREX_2}}$ is also called TREX. Note that, in view of Example 7(d)(e) in Section 2.2 and a relation between $g_{(j,q)}$ and $\widetilde{\varphi}_q$ in (41) [38, in Section 4.3.2]:

$$(\forall(\eta,\mathbf{y})\in\mathbb{R}\times\mathbb{R}^N) \quad g_{(j,q)}(\eta,\mathbf{y}) = \tau_{\left(\mathbf{x}_j^\top\mathbf{z},\mathbf{z}\right)}\widetilde{\varphi}_q(\eta,\mathbf{y}) = \widetilde{\varphi}_q\left(\eta - \mathbf{x}_j^\top\mathbf{z}, \mathbf{y}-\mathbf{z}\right), \tag{136}$$

each convex subproblem in (133) is an instance of Problem (1). For the subproblem (133), the Douglas-Rachford splitting method (see Proposition 9 in Section 2.3) was successfully applied in [38]. For completeness, we reproduce this result in the style of (19–21) followed by application of Fact 6 (in Section 2.2) to the characterization (66). Suppose that for (133) the qualification condition (see (32))

$$0 \in \operatorname{ri}(\operatorname{dom}(g_{(j,q)}) - \mathbf{M}_j \operatorname{dom}(\|\cdot\|_1)) \tag{137}$$

holds[17]. Then, by using

$$\left[\begin{array}{l}\check{\mathbf{M}}_j : \mathbb{R}^p\times\mathbb{R}^{N+1} \to \mathbb{R}^{N+1} : (\mathbf{b},\mathbf{c}) \mapsto \mathbf{M}_j\mathbf{b} - \mathbf{c}, \\ \mathscr{Q}_{\mathbb{R}^p} : \mathbb{R}^p\times\mathbb{R}^{N+1} \to \mathbb{R}^p : (\mathbf{b},\mathbf{c}) \mapsto \mathbf{b},\end{array}\right.$$

we obtain

$$\mathscr{S}_{(j,q)} = \mathscr{Q}_{\mathbb{R}^p} \circ P_{\mathscr{N}(\check{\mathbf{M}}_j)}\left(\operatorname{Fix}\left(\mathbf{T}^{(j,q)}_{\mathrm{DRS_I}}\right)\right) \tag{138}$$

(which is a specialization of (66) for (133), see Figure 3), where

$$\mathbf{T}^{(j,q)}_{\mathrm{DRS_I}} : \mathbb{R}^p\times\mathbb{R}^{N+1} \to \mathbb{R}^p\times\mathbb{R}^{N+1} : (\mathbf{b},\mathbf{c}) \mapsto (\mathbf{b}_T, \mathbf{c}_T) \tag{139}$$

is the DRS operator of Type-I (c.f., (57) and (60)) specialized for (133) and is defined by

$$\left[\begin{array}{l}\mathbf{p} = \mathbf{b} - \mathbf{M}_j^*(\mathbf{I} + \mathbf{M}_j\mathbf{M}_j^*)^{-1}(\mathbf{M}_j\mathbf{b} - \mathbf{c}) \\ (\mathbf{b}_{1/2}, \mathbf{c}_{1/2}) = (2\mathbf{p} - \mathbf{b}, 2\mathbf{M}_j\mathbf{p} - \mathbf{c})) \\ (\mathbf{b}_T, \mathbf{c}_T) = (2\operatorname{prox}_{\|\cdot\|_1}(\mathbf{b}_{1/2}) - \mathbf{b}_{1/2}, 2\operatorname{prox}_{g_{(j,q)}}(\mathbf{c}_{1/2}) - \mathbf{c}_{1/2}),\end{array}\right.$$

or equivalently by

$$\mathbf{T}^{(j,q)}_{\mathrm{DRS_I}} := (2\operatorname{prox}_{F_{(j,q)}} - \mathbf{I}) \circ (2P_{\mathscr{N}(\check{\mathbf{M}}_j)} - \mathbf{I})$$

[17] In [38], the qualification condition (137) seems to be assumed implicitly. If we assume additionally that $\mathbf{X} \in \mathbb{R}^{N\times p}$ has no zero column, it is automatically guaranteed as will be shown in Lemma 27 in Section 5.2.

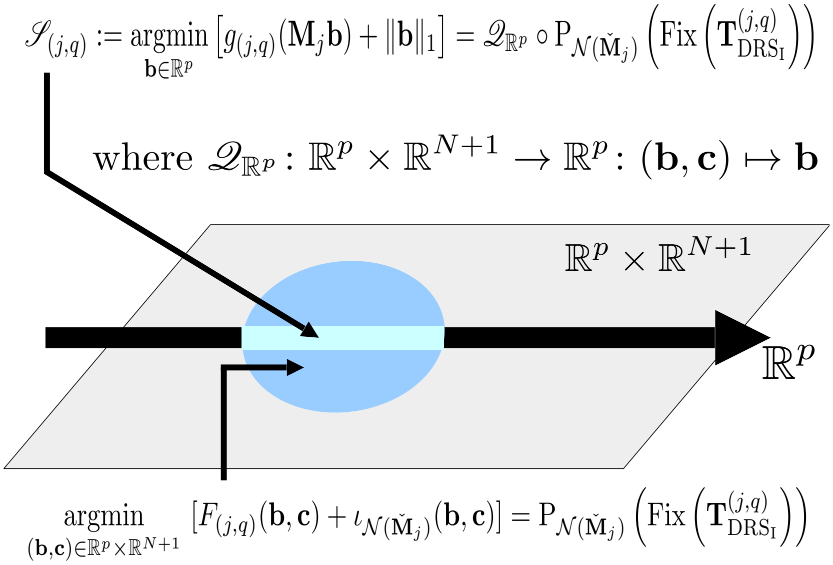

$\mathscr{S}_{(j,q)} := \operatorname*{argmin}_{\mathbf{b}\in\mathbb{R}^p}\left[g_{(j,q)}(\mathrm{M}_j\mathbf{b}) + \|\mathbf{b}\|_1\right] = \mathscr{Q}_{\mathbb{R}^p} \circ \mathrm{P}_{\mathcal{N}(\check{\mathrm{M}}_j)}\left(\mathrm{Fix}\left(\mathbf{T}^{(j,q)}_{\mathrm{DRS_I}}\right)\right)$
where $\mathscr{Q}_{\mathbb{R}^p} : \mathbb{R}^p \times \mathbb{R}^{N+1} \to \mathbb{R}^p : (\mathbf{b},\mathbf{c}) \mapsto \mathbf{b}$
$\mathbb{R}^p \times \mathbb{R}^{N+1}$
$\mathbb{R}^p$
$\operatorname*{argmin}_{(\mathbf{b},\mathbf{c})\in\mathbb{R}^p\times\mathbb{R}^{N+1}}\left[F_{(j,q)}(\mathbf{b},\mathbf{c}) + \iota_{\mathcal{N}(\check{\mathrm{M}}_j)}(\mathbf{b},\mathbf{c})\right] = \mathrm{P}_{\mathcal{N}(\check{\mathrm{M}}_j)}\left(\mathrm{Fix}\left(\mathbf{T}^{(j,q)}_{\mathrm{DRS_I}}\right)\right)$

Fig. 3 Illustration of the fixed point characterization of $\mathscr{S}_{(j,q)}$ in (133) via the Douglas-Rachford splitting operator $\mathbf{T}^{(j,q)}_{\mathrm{DRS_I}}$ in (139).

with $F_{(j,q)}: \mathbb{R}^p \times \mathbb{R}^{N+1} \to (-\infty,\infty]: (\mathbf{b},\mathbf{c}) \mapsto g_{(j,q)}(\mathbf{c}) + \|\mathbf{b}\|_1$. Note that $\mathbf{T}^{(j,q)}_{\mathrm{DRS_I}}$ can be computed efficiently if $(\mathbf{I}+\mathbf{M}_j\mathbf{M}_j^*)^{-1}$ is available as a computational tool.

The above characterization (138) and Fact 6 (see Section 2.2) lead to the following algorithmic solution of (133).

Fact 26 (Douglas-Rachford Splitting Method for Subproblems of Generalized TREX) *Under the qualification condition* (137) *for* $\mathscr{S}_{(j,q)}$ *in* (133)*, the sequence* $(\mathbf{b}_n,\mathbf{c}_n)_{n\in\mathbb{N}} \subset \mathbb{R}^p \times \mathbb{R}^{N+1}$ *generated, with* $(\alpha_n)_{n\in\mathbb{N}} \subset [0,1]$ *satisfying* $\sum_{n\in\mathbb{N}} \alpha_n(1-\alpha_n) = \infty$ *in Fact 6 (see Section 2.2) and* $(\mathbf{b}_0,\mathbf{c}_0) \in \mathbb{R}^p \times \mathbb{R}^{N+1}$*, by*

$$(\mathbf{b}_{n+1},\mathbf{c}_{n+1}) = (1-\alpha_n)(\mathbf{b}_n,\mathbf{c}_n) + \alpha_n\mathbf{T}^{(j,q)}_{\mathrm{DRS_I}}(\mathbf{b}_n,\mathbf{c}_n) \tag{140}$$

converges to a point $(\mathbf{b}_\star,\mathbf{c}_\star)$ *in* $\mathrm{Fix}\left(\mathbf{T}^{(j,q)}_{\mathrm{DRS_I}}\right)$ *as well as the sequence* $(\mathscr{Q}_{\mathbb{R}^p} \circ P_{\mathscr{N}(\check{\mathbf{M}}_j)}(\mathbf{b}_n,\mathbf{c}_n))_{n\in\mathbb{N}}$ *converges to* $\mathscr{Q}_{\mathbb{R}^p} \circ P_{\mathscr{N}(\check{\mathbf{M}}_j)}(\mathbf{b}_\star,\mathbf{c}_\star) \in \mathscr{S}_{(j,q)}$*, where*

$$\mathscr{Q}_{\mathbb{R}^p} \circ P_{\mathscr{N}(\check{\mathbf{M}}_j)}: \mathbb{R}^p \times \mathbb{R}^{N+1} \to \mathbb{R}^p: (\mathbf{b},\mathbf{c}) \mapsto \mathbf{b} - \mathbf{M}_j^*(\mathbf{I}+\mathbf{M}_j\mathbf{M}_j^*)^{-1}(\mathbf{M}_j\mathbf{b}-\mathbf{c}).$$

Note that the sequence $(\mathscr{Q}_{\mathbb{R}^p} \circ P_{\mathscr{N}(\check{\mathbf{M}}_j)}(\mathbf{b}_n,\mathbf{c}_n))_{n\in\mathbb{N}}$ can be generated efficiently by (140) if $(\mathbf{I}+\mathbf{M}_j\mathbf{M}_j^*)^{-1}$ is available as a computational tool.

5.2 Enhancement of Generalized TREX Solutions with Hierarchical Optimization

Along Scenario 2 in Section 1, suppose that we found newly an effective criterion $\Psi \in \Gamma_0(\mathbb{R}^p)$ whose gradient is Lipschitzian over $\mathbb{R}^p$ and we hope to select a most desirable vector, in the sense of Ψ, from the solution set $\mathscr{S}_{\mathrm{TREX_q}}$ in (132). This task is formulated as a *hierarchical nonconvex optimization problem* (see (131), (134), and (135) for $\widehat{\mathscr{Q}}_{\mathbb{R}^p}$, $g_{(j,q)}$, and $\mathbf{M}_j$):

$$\begin{aligned}&\text{minimize } \Psi(\mathbf{b}^\star) \\ &\text{subject to } \mathbf{b}^\star \in \mathscr{S}_{\mathrm{TREX_q}} = \widehat{\mathscr{Q}}_{\mathbb{R}^p}\left[\underset{(\mathbf{b},j)\in\mathbb{R}^p\times\{1,2,\ldots,2p\}}{\operatorname{argmin}} g_{(j,q)}(\mathbf{M}_j\mathbf{b}) + \|\mathbf{b}\|_1\right]\end{aligned} \tag{141}$$

whose solution $\mathbf{b}_{\mathrm{HTREX_q}}$ is given, by passing through $2p$ (*hierarchical convex* optimization) subproblems, as $\mathbf{b}_{\mathrm{HTREX_q}}^{(j^{\star\star})}$, where

$$\left[\begin{array}{l}\mathbf{b}_{\mathrm{HTREX_q}}^{(j)} \in \Omega_{\mathrm{DRS_I}}^{(j,q)} := \underset{\mathbf{b}^\star\in\mathscr{S}_{(j,q)}}{\operatorname{argmin}} \Psi(\mathbf{b}^\star), \\ \mathscr{S}_{(j,q)} = \underset{\mathbf{b}\in\mathbb{R}^p}{\operatorname{argmin}} \left[g_{(j,q)}(\mathbf{M}_j\mathbf{b}) + \|\mathbf{b}\|_1\right] \quad (j = 1,2,\ldots,2p) \text{ [in (133)]}; \\ \mathfrak{J}^\star := \underset{j\in\{1,2,\ldots,2p\}}{\operatorname{argmin}} \left[g_{(j,q)}(\mathbf{M}_j\mathbf{b}_{\mathrm{HTREX_q}}^{(j)}) + \|\mathbf{b}_{\mathrm{HTREX_q}}^{(j)}\|_1\right]; \\ j^{\star\star} \in \underset{j^\star\in\mathfrak{J}^\star}{\operatorname{argmin}} \Psi(\mathbf{b}_{\mathrm{HTREX_q}}^{(j^\star)}).\end{array}\right. \tag{142}$$

Note that the coercivity of $\|\cdot\|_1$ and the nonnegativity of $g_{(j,q)}$ ensure that $\mathscr{S}_{(j,q)}$ is nonempty and bounded (see Fact 2(c) in Section 2.1), which also guarantees $\Omega_{\mathrm{DRS_I}}^{(j,q)} = \operatorname{argmin}(\iota_{\mathscr{S}_{(j,q)}} + \Psi)(\mathbb{R}^p) \neq \varnothing$ $(j = 1,2,\ldots,2p)$ by the classical Weierstrass theorem.

In the following, we focus on how to compute the solution $\mathbf{b}_{\mathrm{HTREX_q}}^{(j)}$ $(j = 1,2,\ldots,2p)$ in (142) by a proposed strategy in Section 3. We assume that the design matrix $\mathbf{X} \in \mathbb{R}^{N\times p}$ in (126) has no zero column, to guarantee the qualification condition (137) for $\mathscr{S}_{(j,q)}$ in (133) for each $j = 1,2,\ldots,2p$.

Lemma 27 *Suppose that the design matrix $\mathbf{X} \in \mathbb{R}^{N\times p}$ has no zero column. Then the qualification condition* (137) *for $\mathscr{S}_{(j,q)}$ in* (133) *is guaranteed automatically for each $j = 1,2,\ldots,2p$.*

(The proof of Lemma 27 is given in Appendix G).

Theorem 28 (Algorithmic Solution to Hierarchical TREX$_q$). *Suppose that $\mathbf{X}$ has no zero column and $\Psi \in \Gamma_0(\mathbb{R}^p)$ is Gâteaux differentiable with Lipschitzian gradient $\nabla\Psi$ over $\mathbb{R}^p$. Then, for $\mathbf{T}_{\mathrm{DRS_I}}^{(j,q)}$ in* (139) *$(j = 1,2,\ldots,2p)$,*

(a) $\operatorname{Fix}(\mathbf{T}_{\mathrm{DRS_I}}^{(j,q)})$ *is bounded;*

(b) $\mathbf{T}_{\mathrm{DRS_I}}^{(j,q)}$ *can be plugged into the HSDM* (54)*, with any $\alpha \in (0,1)$ and $(\lambda_{n+1})_{n\in\mathbb{N}} \in \ell_+^2 \setminus \ell_+^1$, as*

$$\left[\begin{array}{l}(\mathbf{b}_{n+1/2},\mathbf{c}_{n+1/2}) = (1-\alpha)(\mathbf{b}_n,\mathbf{c}_n) + \alpha\mathbf{T}_{\mathrm{DRS_I}}^{(j,q)}(\mathbf{b}_n,\mathbf{c}_n) \\ \mathbf{b}_{n+1}^\star = \mathbf{b}_{n+1/2} - \mathbf{M}_j^*(\mathbf{I}+\mathbf{M}_j\mathbf{M}_j^*)^{-1}(\mathbf{M}_j\mathbf{b}_{n+1/2} - \mathbf{c}_{n+1/2}) \\ \mathbf{b}_{n+1} = \mathbf{b}_{n+1/2} - \lambda_{n+1}(\mathbf{I} - \mathbf{M}_j^*(\mathbf{I}+\mathbf{M}_j\mathbf{M}_j^*)^{-1}\mathbf{M}_j)\circ\nabla\Psi(\mathbf{b}_{n+1}^\star) \\ \mathbf{c}_{n+1} = \mathbf{c}_{n+1/2} - \lambda_{n+1}((\mathbf{I}+\mathbf{M}_j\mathbf{M}_j^*)^{-1}\mathbf{M}_j)\circ\nabla\Psi(\mathbf{b}_{n+1}^\star).\end{array}\right. \tag{143}$$

The algorithm (143) *generates, for any $(\mathbf{b}_0,\mathbf{c}_0) \in \mathbb{R}^p \times \mathbb{R}^{N+1}$, a sequence $(\mathbf{b}_{n+1}^\star)_{n\in\mathbb{N}} \subset \mathbb{R}^p$ which satisfies*

$$\lim_{n\to\infty} d_{\Omega_{\mathrm{DRS_I}}^{(j,q)}}(\mathbf{b}_n^\star) = 0,$$

where $\Omega_{\mathrm{DRS_I}}^{(j,q)} \neq \varnothing$ is defined in (142).

Remark 29 (Idea Behind Derivation of Theorem 28)

(a) *Recall that $\mathbf{T}_{\mathrm{DRS_I}}^{(j,q)}$ is a DRS operator of Type-I (see* (139) *and Theorem 15 in Section 3.1). By applying Corollary 24(a) in Section 3.3 to Lemma 27, the boundedness of $\mathscr{S}_{(j,q)} \neq \varnothing$, and the boundedness of the image of $\partial\|\cdot\|_1 \colon \mathbb{R}^p \to [-1,1]^p \colon \mathbf{b} = (b_1,b_2,\ldots,b_p) \mapsto \bigtimes_{i=1}^p \partial|\cdot|(b_i)$, we deduce the relation:*

$$\left[-(\mathbf{M}_j^\top)^{-1}(\partial\|\cdot\|_1(\mathscr{S}_{(j,q)}))\right]\cap\partial g_{(j,q)}(\mathbb{R}^{N+1}) \textit{ is bounded} \tag{144}$$
$$\Rightarrow \mathrm{Fix}(\mathbf{T}^{(j,q)}_{\mathrm{DRS_I}}) \textit{ is bounded,}$$

where $(\mathbf{M}_j^\top)^{-1}\colon \mathbb{R}^p \to 2^{\mathbb{R}^{N+1}}\colon \mathbf{b}\mapsto\{\mathbf{c}\in\mathbb{R}^{N+1} \mid \mathbf{b}=\mathbf{M}_j^\top\mathbf{c}\}$ *(see* (135) *for* $\mathbf{M}_j$*). Now, by* $\partial g_{(j,q)}(\mathbb{R}^{N+1}) = \partial\widetilde{\varphi}_q(\mathbb{R}^{N+1})$ *(due to* (136)*) and the supercoercivity of* φ_q *and* φ_q^* *(due to [9, Example 13.2 and Example 13.8]), for proving the boundedness of* $\mathrm{Fix}(\mathbf{T}^{(j,q)}_{\mathrm{DRS_I}})$ *from* (144)*, it is sufficient to show the following claim:*

Claim 28: Suppose that $\mathbf{X}$ has no zero column. Let $S\subset\mathbb{R}^p$ be bounded, and $\varphi\in\Gamma_0(\mathbb{R}^N)$ a supercoercive function having supercoercive $\varphi^*\in\Gamma_0(\mathbb{R}^N)$. Then $(\mathbf{M}_j^\top)^{-1}(S)\cap\partial\widetilde{\varphi}(\mathbb{R}^{N+1})$ is bounded.

Note that Claim 28 is the main step in the proof of Theorem 28.

(b) *We have already confirmed the qualification condition* (137) *in Lemma 27,* $\mathscr{S}_{(j,q)}\neq\varnothing$*, and* $\Omega^{(j,q)}_{\mathrm{DRS_I}}\neq\varnothing$ *(*$j=1,2,\ldots,2p$*) (see the short remark just after* (142)*). Therefore, application of Theorem 15 (in Section 3.1) to the subproblems to compute* $\mathbf{b}^{(j)}_{\mathrm{HTREX_q}}$ *(*$j=1,2,\ldots,2p$*) in* (142) *guarantees the statement of Theorem 28(b).*

(The proof of Theorem 28 is given in Appendix H).

5.3 Numerical Experiment: Hierarchical TREX$_2$

We demonstrate that the proposed estimator $\mathbf{b}_{\mathrm{HTREX}_2}$, i.e., Hierarchical TREX$_2$ in (141) (see Section 5.2) can enhance further the estimation accuracy achieved by $\mathbf{b}_{\mathrm{TREX}_2}$ in (132) if we can exploit another new criterion $\Psi\colon\mathbb{R}^p\to\mathbb{R}$ for promoting characteristics, of $\mathbf{b}^{\mathrm{tru}}$, which is not utilized in TREX$_2$. Consider the situation where we like to estimate unknown vector

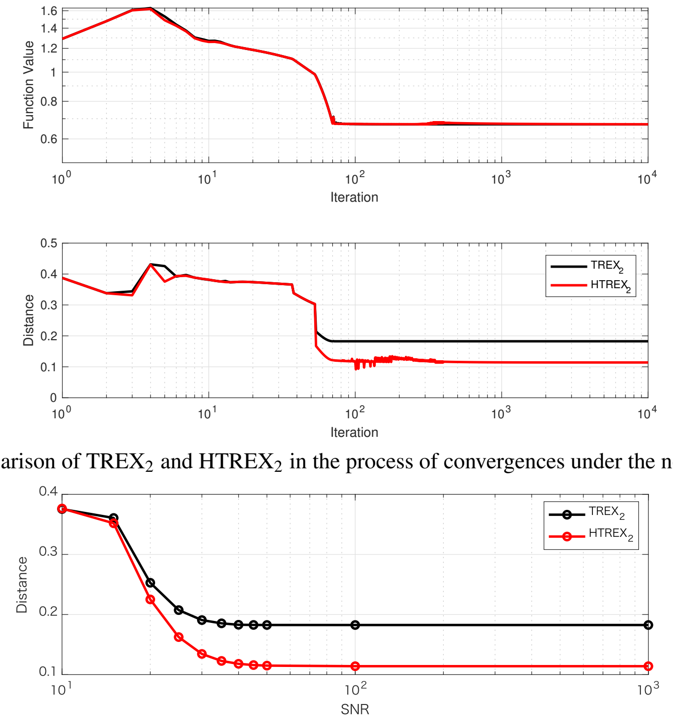
$$\mathbf{b}^{\mathrm{tru}} = \frac{1}{\sqrt{p}}(0,0,0,1,1,1,0,\ldots,0)^\top\in\mathbb{R}^p$$

from the noisy observation $\mathbf{z}\in\mathbb{R}^N$ in (126). We suppose to know that $\mathbf{b}^{\mathrm{tru}}$ is not only sparse but also *fairly flat*. Here, the fairly flatness of $\mathbf{b}^{\mathrm{tru}}$ means that the energy of oscillations (i.e., the sum of the squared gaps between the adjacent components) of $\mathbf{b}^{\mathrm{tru}}$ is small, which is supposed to be our additional knowledge not utilized in the TREX$_2$ and Lasso estimators. If we have such prior knowledge, suppression of

$$\Psi\colon\mathbb{R}^p\to\mathbb{R}\colon\mathbf{b}\mapsto\frac{1}{2}\|\mathbf{D}\mathbf{b}\|^2,\ \text{ with }\mathbf{D}:=\begin{pmatrix}-1&1&0&0&0&0\\0&-1&1&0&0&0\\&&\ddots&\ddots&&\\0&0&0&-1&1&0\\0&0&0&0&-1&1\end{pmatrix}\in\mathbb{R}^{(p-1)\times p}, \tag{145}$$

is expected to be effective for estimation of $\mathbf{b}^{\mathrm{tru}}$ because Ψ can distinguish $\mathbf{b}^{\mathrm{tru}}$ from $\widetilde{\mathbf{b}}:=\frac{1}{\sqrt{p}}(0,1,0,0,1,1,0,\ldots,0)^\top\in\mathbb{R}^p$ of the same sparsity as $\mathbf{b}^{\mathrm{tru}}$ (i.e., $\|\mathbf{b}^{\mathrm{tru}}\|_0=\|\widetilde{\mathbf{b}}\|_0$ and $\|\mathbf{b}^{\mathrm{tru}}\|_1=\|\widetilde{\mathbf{b}}\|_1$) by $\Psi(\mathbf{b}^{\mathrm{tru}})<\Psi(\widetilde{\mathbf{b}})$. Now, our new goal for enhancement of TREX$_2$ is to minimize Ψ while keeping the optimality of the TREX$_2$ in the sense of (132) for $q=2$ (see Scenario 2 in Section 1). This goal is achieved by solving the hierarchical nonconvex optimization (141) for $q=2$.

In our experiments, the design matrix $\mathbf{X}\in\mathbb{R}^{N\times p}$ in (126) is given to satisfy $\mathbf{X}_{:2}=\mathbf{X}_{:3}=\mathbf{X}_{:4}$ with a sample of zero-mean Gaussian random variable followed by normalization to satisfy $\|\mathbf{X}_{:j}\|=\sqrt{N}$ ($j=1,\ldots,p$). The additive noise $\mathbf{e}\in\mathbb{R}^N$ in (126) is drawn from the unit white Gaussian distribution. We tested the performances of the estimators under (SNR) $=10\log\left(\frac{\|\mathbf{X}\mathbf{b}^{\mathrm{tru}}\|^2}{\|\sigma\mathbf{e}\|^2}\right)\in[10,1000]\cup\{+\infty\}$, where $\sigma\in\mathbb{R}$ is adjusted to obtain a



(a) Comparison of TREX$_2$ and HTREX$_2$ in the process of convergences under the noise $\mathbf{e} = 0$.

(b) Estimation accuracy achieved by TREX$_2$ and HTREX$_2$ for various SNR.

Fig. 4 Transient performances in an over-determined case; Criteria “Function Value”, “Distance” are given in (146) and (SNR) $= 10\log\left(\frac{\|\mathbf{X}\mathbf{b}^{\mathrm{tru}}\|^2}{\|\sigma\mathbf{e}\|^2}\right)$ [dB].

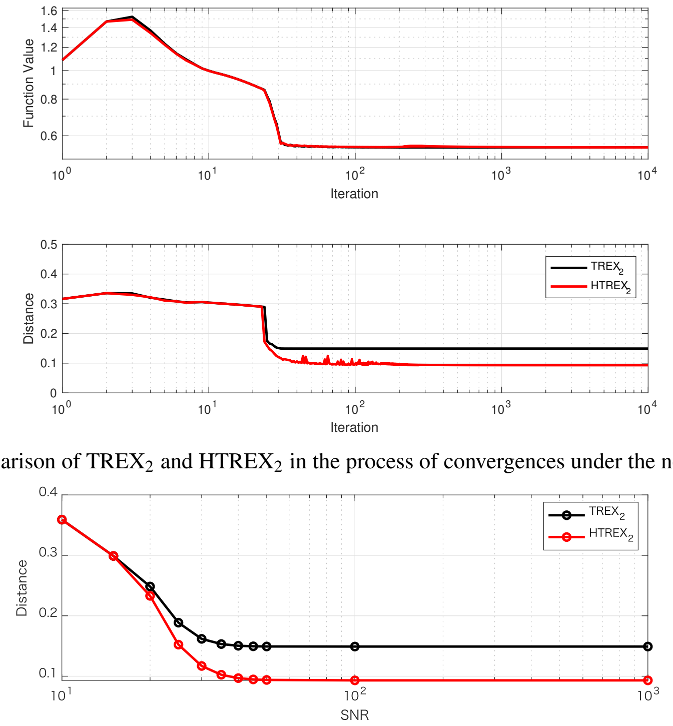
specific SNR. Note that, in this setting, $\left\{\mathbf{b} \in \mathbb{R}^p \mid \mathbf{X}\mathbf{b} = \mathbf{X}\mathbf{b}^{\mathrm{tru}} \text{ and } \|\mathbf{b}\|_1 = \|\mathbf{b}^{\mathrm{tru}}\|_1\right\}$ is apparently an infinite set containing both $\mathbf{b}^{\mathrm{tru}}$ and $\widetilde{\mathbf{b}}$.

We compared the performances of $\mathbf{b}_{\mathrm{TREX}_2}$ in (132) ($\beta = 1/2$) and $\mathbf{b}_{\mathrm{HTREX}_2}$ in (141) employing Ψ in (145). To approximate iteratively $\mathbf{b}^{(j)}_{\mathrm{TREX}_2}$ ($j = 1, 2, \ldots, 2p$) for (133) and $\mathbf{b}^{(j)}_{\mathrm{HTREX}_2}$ ($j = 1, 2, \ldots, 2p$) for (142), we used respectively TREX$_2$ (140) (Fact 26 with $\alpha_n = 1.95$ ($n \in \mathbb{N}$)) and the proposed algorithm (143) (HTREX$_2$ with $\alpha = 1.95$ and $\lambda_n = \frac{1}{n}$ for $n \in \mathbb{N}$). As performance measures, we used, in Figure 4 and Figure 5,

$$\left[\begin{array}{ll} \textbf{Function Value} \text{ (see (141))} & \min_{j=1,\ldots,2p}\left(g_{(j,2)}(\mathbf{M}_j\mathbf{b}_n) + \|\mathbf{b}_n\|_1\right), \\ \textbf{Distance} & \|\mathbf{b}_n - \mathbf{b}^{\mathrm{tru}}\|. \end{array}\right. \tag{146}$$

The experiments were performed both in an over-determined case ($N = 30$ and $p = 20$) in Figure 4 and an under-determined case ($N = 20$ and $p = 30$) in Figure 5.

Figure 4(a) and 5(a) illustrate the process of convergences of TREX$_2$ and HTREX$_2$ in the absence of noise, i.e., $\mathbf{e} = \mathbf{0} \in \mathbb{R}^N$. From these figures, we observe that (i) Function Values of TREX$_2$ and HTREX$_2$ converge to the same level, and that (ii) Distance (to $\mathbf{b}^{\mathrm{tru}}$) of HTREX$_2$ converges to a lower level than that of TREX$_2$. Figure 4(b) and 5(b) summarize the behavior of Distance (to $\mathbf{b}^{\mathrm{tru}}$), against various SNR, by TREX$_2$ and HTREX$_2$ after 10000 iterations. For all the SNR, HTREX$_2$ seems to succeed in improving the performance of TREX$_2$.



(a) Comparison of TREX$_2$ and HTREX$_2$ in the process of convergences under the noise $\mathbf{e} = 0$.

(b) Estimation accuracy achieved by TREX$_2$ and HTREX$_2$ for various SNR.

Fig. 5 Transient performances in an under-determined case; Criteria "Function Value", "Distance" are given in (146) and $(\mathrm{SNR}) = 10\log\left(\frac{\|\mathbf{X}\mathbf{b}^{\mathrm{tru}}\|^2}{\|\sigma\mathbf{e}\|^2}\right)$ [dB].

6 Concluding Remarks

In this paper, we have demonstrated how the modern proximal splitting operators can be plugged nicely into the hybrid steepest descent method (HSDM) for their applications to the hierarchical convex optimization problems which require further strategic selection of a most desirable vector from the set of all solutions of the standard convex optimization. For simplicity as well as for broad applicability, we have chosen to cast our target in the iterative approximation of a viscosity solution of the standard convex optimization problem, where the 1st stage cost function is given as a superposition of multiple nonsmooth convex functions, involving linear operators, while its viscosity solution is a minimizer of the 2nd stage cost function which is Gâteaux differentiable convex function with Lipschitzian gradient. The key ideas for the successful collaboration between the proximal splitting operators and the HSDM are not only in (i) the previously known expressions of the solution set of the standard convex optimization problem as the fixed point set of computable nonexpansive operators but also in (ii) linear relations build strategically between the solution set and the fixed point set. Fortunately, we have shown that such key ideas can be achieved by extending carefully the strategies behind the Douglas-Rachford splitting operators as well as the LAL operators defined in certain product Hilbert spaces. We have also presented applications of the proposed algorithmic strategies to certain unexplored hierarchical enhancements of the support vector machine and the Lasso estimator.

Acknowledgements Isao Yamada would like to thank Heinz H. Bauschke, Russell Luke, and Regina S. Burachik for their kind encouragement and invitation of the first author to the dream meeting: *Splitting Algorithms, Modern Operator Theory, and Applications* (September 17-22, 2017) in Oaxaca, Mexico where he had great opportunity to receive insightful deep comments by

Hédy Attouch. He would also like to thank Patrick Louis Combettes and Christian L. Müller for their invitation of the first author to a special mini-symposium *Proximal Techniques for High-Dimensional Statistics* in the SIAM conference on Optimization 2017 (May 22-25, 2017) in Vancouver. Their kind invitations and their excellent approach to the TREX problem motivated very much the authors to study the application of the proposed strategies to the hierarchical enhancement of Lasso in this paper. Isao Yamada would also like to thank Raymond Honfu Chan for his kind encouragement and invitation to *the Workshop on Optimization in Image Processing* (June 27-30, 2016) at the Harvard University. Lastly, the authors thank to Yunosuke Nakayama for his help in the numerical experiment related to the proposed hierarchical enhancement of the SVM.

Appendices

A: Proof of Proposition 9(a)

Fact 5(i)⇔(ii) in Section 2.1 yields

$$\begin{aligned}\text{(42–44)} &\Leftrightarrow (\exists \nu_\star \in \mathscr{K})\ \nu_\star \in \partial f(x_\star) \text{ and } -\nu_\star \in \partial g(x_\star)\\ &\Leftrightarrow 0 \in \partial f(x_\star) + \partial g(x_\star).\end{aligned}$$

The remaining follows from the proof in [40, Proposition 18]. □

B: Proof of Proposition 10(a)(d)

(a) From (47) and (48), there exists $(x_\star, \nu_\star) \in \mathscr{S}_{\mathrm{pLAL}} \times \mathscr{S}_{\mathrm{dLAL}}$. Fact 5(i)⇔(ii) in Section 2.1 yields the equivalence

$$\begin{aligned}&(x_\star, \nu_\star) \in \mathscr{S}_{\mathrm{pLAL}} \times \mathscr{S}_{\mathrm{dLAL}} \text{ and (49)}\\ \Leftrightarrow\ & A^*\nu_\star \in \partial f(x_\star) \text{ and } -\nu_\star \in \partial \iota_{\{0\}}(Ax_\star)\\ \Leftrightarrow\ & A^*\nu_\star \in \partial f(x_\star) \text{ and } Ax_\star = 0 \qquad \text{(A.1)}\\ \Leftrightarrow\ & x_\star = \mathrm{prox}_f(x_\star - A^*Ax_\star + A^*\nu_\star) \text{ and } \nu_\star = \nu_\star - Ax_\star\\ \Leftrightarrow\ & (x_\star, \nu_\star) \in \mathrm{Fix}(T_{\mathrm{LAL}}). \qquad \text{(A.2)}\end{aligned}$$

□

(d) Choose arbitrarily $(\bar{x}, \bar{\nu}) \in \mathrm{Fix}(T_{\mathrm{LAL}})$, i.e.,

$$(\bar{x}, \bar{\nu}) = T_{\mathrm{LAL}}(\bar{x}, \bar{\nu}) = \left(\mathrm{prox}_f(\bar{x} - A^*A\bar{x} + A^*\bar{\nu}), \bar{\nu} - A\bar{x}\right).$$

Let $(x_n, \nu_n)_{n\in\mathbb{N}} \subset \mathscr{X} \times \mathscr{K}$ be generated, with any $(x_0, \nu_0) \in \mathscr{X} \times \mathscr{K}$, by

$$(x_{n+1}, \nu_{n+1}) = T_{\mathrm{LAL}}(x_n, \nu_n) = \left(\mathrm{prox}_f(x_n - A^*Ax_n + A^*\nu_n), \nu_n - Ax_{n+1}\right). \qquad \text{(A.3)}$$

Then [150, (B.3)] yields

$$
\begin{aligned}
0 \leq & \|x_n - \bar{x}\|_{\mathscr{X}}^2 - \|x_{n+1} - \bar{x}\|_{\mathscr{X}}^2 - \|(x_{n+1} - x_n) - (\bar{x} - \bar{x})\|_{\mathscr{X}}^2 \\
& + \|Ax_{n+1} - A\bar{x} - (Ax_n - A\bar{x})\|_{\mathscr{K}}^2 - \|Ax_n - A\bar{x}\|_{\mathscr{K}}^2 \\
& - \|Ax_{n+1} - A\bar{x} + \bar{\nu} - \nu_n\|_{\mathscr{K}}^2 + \|\bar{\nu} - \nu_n\|_{\mathscr{K}}^2 \\
= & \|x_n - \bar{x}\|_{\mathscr{X}}^2 - \|x_{n+1} - \bar{x}\|_{\mathscr{X}}^2 - \|x_{n+1} - x_n\|_{\mathscr{X}}^2 \\
& + \|Ax_{n+1} - Ax_n\|_{\mathscr{K}}^2 - \|Ax_n\|_{\mathscr{K}}^2 - \|\bar{\nu} - \nu_{n+1}\|_{\mathscr{K}}^2 + \|\bar{\nu} - \nu_n\|_{\mathscr{K}}^2 \\
\leq & (\|x_n - \bar{x}\|_{\mathscr{X}}^2 + \|\nu_n - \bar{\nu}\|_{\mathscr{K}}^2) - (\|x_{n+1} - \bar{x}\|_{\mathscr{X}}^2 + \|\nu_{n+1} - \bar{\nu}\|_{\mathscr{K}}^2) \\
& + (\|A\|_{\mathrm{op}}^2 - 1)\|x_{n+1} - x_n\|_{\mathscr{X}}^2 - \|Ax_n\|_{\mathscr{K}}^2. \qquad \text{(A.4)}
\end{aligned}
$$

(A.4) and $\|A\|_{\mathrm{op}} < 1$ imply that $(\|x_n - \bar{x}\|_{\mathscr{X}}^2 + \|\nu_n - \bar{\nu}\|_{\mathscr{K}}^2)_{n\in\mathbb{N}}$ decreases monotonically, i.e., $(x_n, \nu_n)_{n\in\mathbb{N}}$ is Fejér monotone with respect to $\mathrm{Fix}(T_{\mathrm{LAL}})$, and $(\|x_n - \bar{x}\|_{\mathscr{X}}^2 + \|\nu_n - \bar{\nu}\|_{\mathscr{K}}^2)_{n\in\mathbb{N}}$ converges to some $c \geq 0$. From this observation, we have

$$
\begin{aligned}
& \sum_{n=0}^{N} \left[(1 - \|A\|_{\mathrm{op}}^2)\|x_{n+1} - x_n\|_{\mathscr{X}}^2 + \|Ax_n\|_{\mathscr{K}}^2\right] \\
\leq & \sum_{n=0}^{N} \left[(\|x_n - \bar{x}\|_{\mathscr{X}}^2 + \|\nu_n - \bar{\nu}\|_{\mathscr{K}}^2) - (\|x_{n+1} - \bar{x}\|_{\mathscr{X}}^2 + \|\nu_{n+1} - \bar{\nu}\|_{\mathscr{K}}^2)\right] \\
= & (\|x_0 - \bar{x}\|_{\mathscr{X}}^2 + \|\nu_0 - \bar{\nu}\|_{\mathscr{K}}^2) - (\|x_{N+1} - \bar{x}\|_{\mathscr{X}}^2 + \|\nu_{N+1} - \bar{\nu}\|_{\mathscr{K}}^2) \\
\to & (\|x_0 - \bar{x}\|_{\mathscr{X}}^2 + \|\nu_0 - \bar{\nu}\|_{\mathscr{K}}^2) - c < \infty \quad (N \to \infty)
\end{aligned}
$$

and thus

$$
\lim_{n\to\infty} \|x_{n+1} - x_n\|_{\mathscr{X}} = 0 \text{ and } \lim_{n\to\infty} \|Ax_n\|_{\mathscr{K}} = 0. \qquad \text{(A.5)}
$$

By [51, Theorem 9.12], the bounded sequence of $(x_n, \nu_n)_{n\in\mathbb{N}}$ has some subsequence $(x_{n_j}, \nu_{n_j})_{j\in\mathbb{N}}$ which converges weakly to a some point, say $(x_\star, \nu_\star)$, in the Hilbert space $\mathscr{X} \times \mathscr{K}$. Therefore, by applying [9, Theorem 9.1(iii)⇔(i)] to $f \in \Gamma_0(\mathscr{X})$, we have

$$
f(x_\star) \leq \liminf_{j\to\infty} f(x_{n_j}) \qquad \text{(A.6)}
$$

and, by the Cauchy-Schwarz inequality and (A.5),

$$
\begin{aligned}
\|Ax_\star\|_{\mathscr{K}}^2 &= \langle Ax_\star - Ax_{n_j}, Ax_\star\rangle_{\mathscr{K}} + \langle Ax_{n_j}, Ax_\star\rangle_{\mathscr{K}} \\
&\leq \langle x_\star - x_{n_j}, A^*Ax_\star\rangle_{\mathscr{X}} + \|Ax_{n_j}\|_{\mathscr{K}}\|Ax_\star\|_{\mathscr{K}} \to 0 \quad (j \to \infty),
\end{aligned}
$$

which implies $Ax_\star = 0$.

Meanwhile, by (A.3), we have

$$
\begin{aligned}
& x_{n_j} = \mathrm{prox}_f(x_{n_j-1} - A^*Ax_{n_j-1} + A^*\nu_n) = (\mathrm{I} + \partial f)^{-1}(x_{n_j-1} - A^*Ax_{n_j-1} + A^*\nu_{n_j-1}) \\
\Leftrightarrow\ & x_{n_j-1} - x_{n_j} - A^*Ax_{n_j-1} + A^*\nu_{n_j-1} \in \partial f(x_{n_j}) \\
\Leftrightarrow\ & (\forall x \in \mathscr{X})\ f(x_{n_j}) + \langle x_{n_j-1} - x_{n_j} - A^*Ax_{n_j-1} + A^*\nu_{n_j-1}, x - x_{n_j}\rangle_{\mathscr{X}} \leq f(x), \qquad \text{(A.7)}
\end{aligned}
$$

where the inner product therein satisfies

$$
\lim_{j\to\infty} \langle x_{n_j-1} - x_{n_j} - A^*Ax_{n_j-1} + A^*\nu_{n_j-1}, x - x_{n_j}\rangle_{\mathscr{X}} = \langle A^*\nu_\star, x - x_\star\rangle_{\mathscr{X}}, \qquad \text{(A.8)}
$$

which is verified by $Ax_\star = 0$, the triangle inequality, the Cauchy-Schwarz inequality, and (A.5), as follows:

$$
\begin{aligned}
&(\forall x \in \mathscr{X}) \\
&\quad |\langle x_{n_j-1} - x_{n_j} - A^*Ax_{n_j-1} + A^*\nu_{n_j-1}, x - x_{n_j}\rangle_{\mathscr{X}} - \langle A^*\nu_\star, x - x_\star\rangle_{\mathscr{X}}| \\
&= |\langle x_{n_j-1} - x_{n_j}, x - x_{n_j}\rangle_{\mathscr{X}} - \langle Ax_{n_j-1}, A(x - x_{n_j})\rangle_{\mathscr{K}} \\
&\qquad\qquad\qquad + \langle \nu_{n_j-1}, A(x - x_{n_j})\rangle_{\mathscr{K}} - \langle \nu_\star, Ax\rangle_{\mathscr{K}}| \\
&= |\langle x_{n_j-1} - x_{n_j}, x - x_{n_j}\rangle_{\mathscr{X}} - \langle Ax_{n_j-1}, A(x - x_{n_j})\rangle_{\mathscr{K}} \\
&\qquad + \langle \nu_{n_j-1}, -Ax_{n_j}\rangle_{\mathscr{K}} - \langle \nu_{n_j} - \nu_{n_j-1}, Ax\rangle_{\mathscr{K}} - \langle \nu_\star - \nu_{n_j}, Ax\rangle_{\mathscr{K}}| \\
&\leq (\|x_{n_j-1} - x_{n_j}\|_{\mathscr{X}} \|x - x_{n_j}\|_{\mathscr{X}} + \|Ax_{n_j-1}\|_{\mathscr{K}} \|A(x - x_{n_j})\|_{\mathscr{K}} \\
&\qquad + \|\nu_{n_j-1}\|_{\mathscr{K}} \|-Ax_{n_j}\|_{\mathscr{K}} + \|Ax_{n_j}\|_{\mathscr{K}} \|Ax\|_{\mathscr{K}} + |\langle \nu_\star - \nu_{n_j}, Ax\rangle_{\mathscr{K}}|) \\
&\to 0 \quad (j \to \infty).
\end{aligned}
$$

Now, by (A.7), (A.6) and (A.8), we have for any $x \in \mathscr{X}$

$$
\begin{aligned}
f(x) &\geq f(x_\star) + \liminf_{j\to\infty} \langle x_{n_j-1} - x_{n_j} - A^*Ax_{n_j-1} + A^*\nu_{n_j-1}, x - x_{n_j}\rangle_{\mathscr{X}} \\
&= f(x_\star) + \lim_{j\to\infty} \langle x_{n_j-1} - x_{n_j} - A^*Ax_{n_j-1} + A^*\nu_{n_j-1}, x - x_{n_j}\rangle_{\mathscr{X}} \\
&= f(x_\star) + \langle A^*\nu_\star, x - x_\star\rangle_{\mathscr{X}},
\end{aligned}
$$

which implies

$$A^*\nu_\star \in \partial f(x_\star). \tag{A.9}$$

By recalling (A.1)⇔(A.2), (A.9) and $Ax_\star = 0$ prove $(x_\star, \nu_\star) \in \mathrm{Fix}(T_{\mathrm{LAL}})$. The above discussion implies that every weak sequential cluster point (see Footnote 7 in Section 2.2) of $(x_n, \nu_n)_{n\in\mathbb{N}}$, which is Fejér monotone with respect to $\mathrm{Fix}(T_{\mathrm{LAL}})$, belongs to $\mathrm{Fix}(T_{\mathrm{LAL}})$. Therefore, [9, Theorem 5.5] guarantees that $(x_n, \nu_n)_{n\in\mathbb{N}}$ converges weakly to a point in $\mathrm{Fix}(T_{\mathrm{LAL}})$. □

C: Proof of Theorem 15

Now by recalling Proposition 9 in Section 2.3 and Remark 16 in Section 3.1, it is sufficient to prove Claim 15. Let $x_\star \in \mathscr{S}_p \neq \varnothing$. Then the Fermat's rule, Fact 4(b) (applicable due to the qualification condition (32)) in Section 2.1, $\breve{A}^* \colon \mathscr{K} \to \mathscr{X} \times \mathscr{K} \colon \nu \mapsto (A^*\nu, -\nu)$ for $\breve{A}$ in (63), the property of $\iota_{\{0\}}$ in (28), the straightforward calculations, and Fact 5(ii)⇔(i) (in Section 2.1) yield

$$\begin{aligned}
x_\star \in \mathscr{S}_p &\Leftrightarrow 0 \in \partial(f+g\circ A)(x_\star) = \partial f(x_\star) + A^*\partial g(Ax_\star) \\
&\Leftrightarrow\ y_\star = Ax_\star \text{ and } 0 \in \partial f(x_\star) + A^*\partial g(y_\star) \\
&\Leftrightarrow (\exists \nu_\star \in \mathscr{K})\ y_\star = Ax_\star \text{ and } \begin{cases} A^*\nu_\star \in \partial f(x_\star) \\ -\nu_\star \in \partial g(y_\star) \end{cases} \\
&\Leftrightarrow (\exists \nu_\star \in \mathscr{K})\ \check{A}(x_\star, y_\star) = 0 \text{ and } \check{A}^*\nu_\star \in \partial F(x_\star, y_\star) \\
&\Leftrightarrow (\exists \nu_\star \in \mathscr{K})\ -\nu_\star \in \partial\iota_{\{0\}}(\check{A}(x_\star, y_\star)) \text{ and } \check{A}^*\nu_\star \in \partial F(x_\star, y_\star) \\
&\Rightarrow (\exists \nu_\star \in \mathscr{K})\ -\check{A}^*\nu_\star \in \check{A}^*\partial\iota_{\{0\}}(\check{A}(x_\star, y_\star)) \text{ and } \check{A}^*\nu_\star \in \partial F(x_\star, y_\star) \\
&\Rightarrow (\exists \nu_\star \in \mathscr{K})\ -\check{A}^*\nu_\star \in \partial(\iota_{\{0\}}\circ\check{A})(x_\star, y_\star) \text{ and } \check{A}^*\nu_\star \in \partial F(x_\star, y_\star) \\
&\Leftrightarrow (\exists \nu_\star \in \mathscr{K})\ -\check{A}^*\nu_\star \in \partial\iota_{\mathscr{N}(\check{A})}(x_\star, y_\star) \text{ and } \check{A}^*\nu_\star \in \partial F(x_\star, y_\star) \\
&\Leftrightarrow (\exists \nu_\star \in \mathscr{K}) \\
&\qquad \begin{cases} (x_\star, y_\star) \in \operatorname{argmin}(F + \iota_{\mathscr{N}(\check{A})})(\mathscr{X}\times\mathscr{K}) \\ \check{A}^*\nu_\star \in \operatorname{argmin}(F^* + \iota^*_{\mathscr{N}(\check{A})}\circ(-\mathrm{I}))(\mathscr{X}\times\mathscr{K}) \\ \min(F + \iota_{\mathscr{N}(\check{A})})(\mathscr{X}\times\mathscr{K}) = -\min(F^* + \iota^*_{\mathscr{N}(\check{A})}\circ(-\mathrm{I}))(\mathscr{X}\times\mathscr{K}), \end{cases}
\end{aligned}$$

which confirms Claim 15. □

D: Proof of Theorem 17

Now by recalling Proposition 9 in Section 2.3 and Remark 18 in Section 3.1, it is sufficient to prove (82) by verifying Claim 17. We will use

$$A^*\circ\partial g\circ A = \sum_{i=1}^{m} A_i^*\circ\partial g_i\circ A_i = \sum_{i=1}^{m}\partial(g_i\circ A_i) \tag{A.10}$$

which is verified by $g = \bigoplus_{i=1}^{m} g_i$, Fact 4(c) (see Section 2.1), and $\mathrm{ri}(\mathrm{dom}(g_j) - \mathrm{ran}(A_j)) = \mathrm{ri}(\mathrm{dom}(g_j) - \mathbb{R}) = \mathbb{R} \ni 0$ $(j = 1,2,\ldots,m)$. Let $x_\star^{(m+1)} \in \mathscr{S}_p \neq \varnothing$. Then by using the Fermat's rule, Fact 4(b) (applicable due to (32)), (A.10), D in (81), and H in (80), we deduce the equivalence

$$
\begin{aligned}
& x_\star^{(m+1)} \in \mathscr{S}_p \\
& \Leftrightarrow 0 \in \partial(f + g \circ A)(x_\star^{(m+1)}) = \partial f(x_\star^{(m+1)}) + A^* \partial g(A x_\star^{(m+1)}) \\
& \qquad\qquad\qquad\qquad\qquad\quad = \partial f(x_\star^{(m+1)}) + \sum_{i=1}^{m} \partial(g_i \circ A_i)(x_\star^{(m+1)}) \\
& \Leftrightarrow (j = 1, \dots, m)\; x_\star^{(j)} = x_\star^{(m+1)} \text{ and } 0 \in \partial f(x_\star^{(m+1)}) + \sum_{i=1}^{m} \partial(g_i \circ A_i)(x_\star^{(i)}) \\
& \Leftrightarrow (\exists \nu^{(1)}, \dots, \nu^{(m)} \in \mathscr{X})(j = 1, \dots, m) \begin{cases} x_\star^{(j)} = x_\star^{(m+1)} \\ \nu^{(j)} \in \partial(g_j \circ A_j)(x_\star^{(j)}) \\ -\sum_{i=1}^{m} \nu^{(i)} \in \partial f(x_\star^{(m+1)}) \end{cases} \\
& \Leftrightarrow (\exists \nu^{(1)}, \dots, \nu^{(m)} \in \mathscr{X}) \\
& \quad \begin{cases} (x_\star^{(1)}, \dots, x_\star^{(m+1)}) \in D \\ \left(\nu^{(1)}, \dots, \nu^{(m)}, -\sum_{i=1}^{m} \nu^{(i)}\right) \in \left[\bigtimes_{j=1}^{m} \partial(g_j \circ A_j)(x_\star^{(j)})\right] \times \partial f(x_\star^{(m+1)}) \\ \qquad\qquad\qquad\qquad\qquad\qquad = \partial H(x_\star^{(1)}, \dots, x_\star^{(m+1)}). \end{cases}
\end{aligned}
$$

Then by $-\left(\nu^{(1)}, \dots, \nu^{(m)}, -\sum_{i=1}^{m} \nu^{(i)}\right) \in D^\perp = \partial \iota_D(x_\star^{(1)}, \dots, x_\star^{(m+1)})$ (see (27)) and by Fact 5(ii)⇔(i) in Section 2.1, we have

$$
\begin{aligned}
x_\star^{(m+1)} \in \mathscr{S}_p \;& \Leftrightarrow (\exists \nu^{(1)}, \dots, \nu^{(m)} \in \mathscr{X}) \\
& \quad \begin{cases} -\left(\nu^{(1)}, \dots, \nu^{(m)}, -\sum_{i=1}^{m} \nu^{(i)}\right) \in \partial \iota_D(x_\star^{(1)}, \dots, x_\star^{(m+1)}) \\ \left(\nu^{(1)}, \dots, \nu^{(m)}, -\sum_{i=1}^{m} \nu^{(i)}\right) \in \partial H(x_\star^{(1)}, \dots, x_\star^{(m+1)}) \end{cases} \\
& \Leftrightarrow (\exists \nu^{(1)}, \dots, \nu^{(m)} \in \mathscr{X}) \\
& \quad \begin{cases} (x_\star^{(1)}, \dots, x_\star^{(m+1)}) \in \operatorname{argmin}(H + \iota_D)(\mathscr{X}^{m+1}) \\ \left(\nu^{(1)}, \dots, \nu^{(m)}, -\sum_{i=1}^{m} \nu^{(i)}\right) \in \operatorname{argmin}(H^* + \iota_D^* \circ (-\mathrm{I}))(\mathscr{X}^{m+1}) \\ \min(H + \iota_D)(\mathscr{X}^{m+1}) = -\min(H^* + \iota_D^* \circ (-\mathrm{I}))(\mathscr{X}^{m+1}), \end{cases}
\end{aligned}
$$

which confirms Claim 17. □

E: Proof of Theorem 19

Now by recalling Proposition 10 in Section 2.3 and Remark 20 in Section 3.2, it is sufficient to prove Claim 19. Let $x_\star \in \mathscr{S}_p \neq \varnothing$. Then the Fermat's rule, Fact 4(b) (applicable due to (32)) in Section 2.1, $\check{A}^* \colon \mathscr{K} \to \mathscr{X} \times \mathscr{K} \colon \nu \mapsto (A^*\nu, -\nu)$ for $\check{A}$ in (63), the property of $\iota_{\{0\}}$ in (28), the straightforward calculations, and Fact 5(ii)⇔(i) (in Section 2.1) yield

$$
\begin{aligned}
x_\star \in \mathscr{S}_p &\Leftrightarrow 0 \in \partial(f+g\circ A)(x_\star) = \partial f(x_\star) + A^*\partial g(Ax_\star)\\
&\Leftrightarrow\ y_\star = Ax_\star \text{ and } 0 \in \partial f(x_\star) + A^*\partial g(y_\star)\\
&\Leftrightarrow (\exists \nu_\star \in \mathscr{K})\ y_\star = Ax_\star \text{ and } \begin{cases} \mathfrak{u}A^*\nu_\star \in \partial f(x_\star)\\ -\mathfrak{u}\nu_\star \in \partial g(y_\star)\end{cases}\\
&\Leftrightarrow (\exists \nu_\star \in \mathscr{K})\ (\mathfrak{u}\check{A})(x_\star,y_\star) = 0 \text{ and } (\mathfrak{u}\check{A})^*\nu_\star \in \partial F(x_\star,y_\star)\\
&\Leftrightarrow (\exists \nu_\star \in \mathscr{K})\ -\nu_\star \in \partial\iota_{\{0\}}((\mathfrak{u}\check{A})(x_\star,y_\star)) \text{ and } (\mathfrak{u}\check{A})^*\nu_\star \in \partial F(x_\star,y_\star)\\
&\Leftrightarrow (\exists \nu_\star \in \mathscr{K})\\
&\qquad \begin{cases} (x_\star,y_\star) \in \operatorname{argmin}\ (F+\iota_{\{0\}}\circ(\mathfrak{u}\check{A}))(\mathscr{X}\times\mathscr{K})\\ \nu_\star \in \operatorname{argmin}(F^*\circ(\mathfrak{u}\check{A})^*)(\mathscr{K})\\ \min\ (F+\iota_{\{0\}}\circ(\mathfrak{u}\check{A}))(\mathscr{X}\times\mathscr{K}) = -\min(F^*\circ(\mathfrak{u}\check{A})^*)(\mathscr{K}),\end{cases}
\end{aligned}
$$

which confirms Claim 19. □

F: Proof of Theorem 23

(a) We have seen in (66) that, under the assumptions of Theorem 23(a), for any vector $x_\star \in \mathscr{X}$,

$$
x_\star \in \mathscr{S}_p[\text{in (10)}] \text{ if and only if } (x_\star,y_\star) = P_{\mathscr{N}(\check{A})}(\zeta_\star) \tag{A.11}
$$

for some $y_\star \in \mathscr{X}$ and some $\zeta_\star \in \operatorname{Fix}(\mathbf{T}_{\mathrm{DRS_I}})$, where $\check{A}\colon \mathscr{X}\times\mathscr{K} \to \mathscr{K}\colon (x,y) \mapsto Ax-y$ (see (63)), $\mathscr{N}(\check{A}) = \{(x,Ax) \in \mathscr{X}\times\mathscr{K} \mid x \in \mathscr{X}\}$, and $\mathbf{T}_{\mathrm{DRS_I}} = (2\operatorname{prox}_F - \mathrm{I})\circ(2P_{\mathscr{N}(\check{A})} - \mathrm{I})$ for $F\colon \mathscr{X}\times\mathscr{K} \to (-\infty,\infty]\colon (x,y) \mapsto f(x)+g(y)$ (see (60) and (62)).

Choose $\zeta_\star := (\zeta_\star^x,\zeta_\star^y) \in \operatorname{Fix}(\mathbf{T}_{\mathrm{DRS_I}})$ arbitrarily and let $\mathbf{z}_\star := (x_\star,y_\star) := P_{\mathscr{N}(\check{A})}(\zeta_\star)$. Then we have

$$
\begin{aligned}
&\zeta_\star \in \operatorname{Fix}(\mathbf{T}_{\mathrm{DRS_I}}) \text{ and } P_{\mathscr{N}(\check{A})}(\zeta_\star) = \mathbf{z}_\star\\
&\Leftrightarrow (2\operatorname{prox}_F - \mathrm{I})\circ(2P_{\mathscr{N}(\check{A})} - \mathrm{I})(\zeta_\star) = \zeta_\star \text{ and } P_{\mathscr{N}(\check{A})}(\zeta_\star) = \mathbf{z}_\star && \text{(A.12)}\\
&\Rightarrow (2\operatorname{prox}_F - \mathrm{I})(2\mathbf{z}_\star - \zeta_\star) = \zeta_\star \ \Leftrightarrow\ \operatorname{prox}_F(2\mathbf{z}_\star - \zeta_\star) = \mathbf{z}_\star\\
&\Leftrightarrow (\mathrm{I}+\partial F)^{-1}(2\mathbf{z}_\star - \zeta_\star) = \mathbf{z}_\star \ \Leftrightarrow\ 2\mathbf{z}_\star - \zeta_\star \in \mathbf{z}_\star + \partial F(\mathbf{z}_\star)\\
&\Leftrightarrow \mathbf{z}_\star - \zeta_\star \in \partial F(\mathbf{z}_\star) = \partial f(x_\star)\times\partial g(y_\star) && \text{(A.13)}\\
&\Leftrightarrow x_\star - \zeta_\star^x \in \partial f(x_\star) \text{ and } y_\star - \zeta_\star^y \in \partial g(y_\star). && \text{(A.14)}
\end{aligned}
$$

Meanwhile, we have

$$
\begin{aligned}
\mathbf{z}_\star = P_{\mathscr{N}(\check{A})}(\zeta_\star) &\Leftrightarrow (\forall \mathbf{z} = (x,Ax) \in \mathscr{N}(\check{A}))\ \langle \zeta_\star - \mathbf{z}_\star, \mathbf{z}\rangle_{\mathscr{X}\times\mathscr{K}} = 0\\
&\Leftrightarrow (\forall x \in \mathscr{X})\ \langle \zeta_\star^x - x_\star, x\rangle_{\mathscr{X}} + \langle \zeta_\star^y - y_\star, Ax\rangle_{\mathscr{K}} = 0\\
&\Leftrightarrow (\forall x \in \mathscr{X})\ \langle (\zeta_\star^x - x_\star) + A^*(\zeta_\star^y - y_\star), x\rangle_{\mathscr{X}} = 0\\
&\Leftrightarrow A^*(\zeta_\star^y - y_\star) = -(\zeta_\star^x - x_\star). && \text{(A.15)}
\end{aligned}
$$

Equations (A.15) and (A.14) imply

$$
\begin{aligned}
&\zeta_\star \in \operatorname{Fix}(\mathbf{T}_{\mathrm{DRS_I}}) \text{ and } P_{\mathscr{N}(\check{A})}(\zeta_\star) = \mathbf{z}_\star\\
&\Rightarrow x_\star - \zeta_\star^x \in \partial f(x_\star) \text{ and } y_\star - \zeta_\star^y \in (-(A^*)^{-1}(\partial f(x_\star))) \cap \partial g(y_\star)\\
&\Rightarrow \zeta_\star = (\zeta_\star^x,\zeta_\star^y) \in (x_\star,y_\star) - \big(\partial f(x_\star)\times[(-(A^*)^{-1}(\partial f(x_\star))) \cap \partial g(y_\star)]\big). && \text{(A.16)}
\end{aligned}
$$

Moreover, by noting that (A.11) ensures $x_\star \in \mathscr{S}_p$ and $y_\star = Ax_\star$, we have from (A.16)

$$\begin{aligned}
&\zeta_\star \in \operatorname{Fix}(\mathbf{T}_{\mathrm{DRS_I}}) \text{ and } (x_\star, Ax_\star) = P_{\mathscr{N}(\check{A})}(\zeta_\star) \\
&\Rightarrow \zeta_\star \in (x_\star, Ax_\star) - \big(\partial f(x_\star) \times [(-(A^*)^{-1}(\partial f(x_\star))) \cap \partial g(Ax_\star)]\big) \\
&\Rightarrow \zeta_\star \in \bigcup_{x' \in \mathscr{S}_p} (x', Ax') - \bigcup_{x'' \in \mathscr{S}_p} \big(\partial f(x'') \times [(-(A^*)^{-1}(\partial f(x''))) \cap \partial g(Ax'')]\big)
\end{aligned}$$

Since $\zeta_\star$ is chosen arbitrarily from $\operatorname{Fix}(\mathbf{T}_{\mathrm{DRS_I}})$, we have

$$\operatorname{Fix}(\mathbf{T}_{\mathrm{DRS_I}}) \subset \bigcup_{x' \in \mathscr{S}_p} (x', Ax') - \bigcup_{x'' \in \mathscr{S}_p} \big(\partial f(x'') \times [(-(A^*)^{-1}(\partial f(x''))) \cap \partial g(Ax'')]\big),$$

from which Theorem 23(a) is confirmed.
(b) We have seen in (95) that, under the assumptions of Theorem 23(b), for any vector $x_\star \in \mathscr{X}$,

$$x_\star \in \mathscr{S}_p[\text{in (10)}] \text{ if and only if } (x_\star, y_\star, \nu_\star) \in \operatorname{Fix}(\mathbf{T}_{\mathrm{LAL}}) \tag{A.17}$$

for some $(y_\star, \nu_\star) \in \mathscr{K} \times \mathscr{K}$, where

$$\begin{aligned}
\mathbf{T}_{\mathrm{LAL}} &: \mathscr{X} \times \mathscr{K} \times \mathscr{K} \to \mathscr{X} \times \mathscr{K} \times \mathscr{K} \\
&: \begin{pmatrix} x \\ y \\ \nu \end{pmatrix} = \begin{pmatrix} \mathbf{z} \\ \nu \end{pmatrix} \mapsto \begin{pmatrix} x_T \\ y_T \\ \nu_T \end{pmatrix} = \begin{pmatrix} \mathbf{z}_T \\ \nu_T \end{pmatrix} = \begin{pmatrix} \operatorname{prox}_F(\mathbf{z} - (\mathrm{u}\check{A})^*(\mathrm{u}\check{A})\mathbf{z} + (\mathrm{u}\check{A})^*\nu) \\ \nu - \mathrm{u}\check{A}\mathbf{z}_T \end{pmatrix}
\end{aligned}$$

and $(\mathrm{u}\check{A})^* : \mathscr{K} \to \mathscr{X} \times \mathscr{K} : \nu \mapsto (\mathrm{u}A^*\nu, -\mathrm{u}\nu)$ (see (93) and (101)).

Choose $(\mathbf{z}_\star, \nu_\star) \in \operatorname{Fix}(\mathbf{T}_{\mathrm{LAL}})$ arbitrarily and denote $\mathbf{z}_\star = (x_\star, y_\star) \in \mathscr{X} \times \mathscr{K}$. By passing similar steps in (A.2)⇔(A.1), we deduce

$$\begin{aligned}
&(\mathbf{z}_\star, \nu_\star) \in \operatorname{Fix}(\mathbf{T}_{\mathrm{LAL}}) \\
&\Leftrightarrow (\mathrm{u}\check{A})^*\nu_\star \in \partial F(\mathbf{z}_\star) = \partial f(x_\star) \times \partial g(y_\star) \text{ and } \mathrm{u}\check{A}(\mathbf{z}_\star) = 0,
\end{aligned} \tag{A.18}$$

and then, from (A.18), straightforward calculations yield

$$\begin{aligned}
&(x_\star, y_\star, \nu_\star) \in \operatorname{Fix}(\mathbf{T}_{\mathrm{LAL}}) \Leftrightarrow \begin{bmatrix} \mathrm{u}A^*\nu_\star = A^*(\mathrm{u}\nu_\star) \in \partial f(x_\star) \\ -\mathrm{u}\nu_\star \in \partial g(y_\star) \\ Ax_\star = y_\star \end{bmatrix} \\
&\Rightarrow -\mathrm{u}\nu_\star \in \big[-(A^*)^{-1}(\partial f(x_\star))\big] \cap \partial g(Ax_\star) \text{ and } Ax_\star = y_\star \\
&\Leftrightarrow -\mathrm{u}(x_\star, y_\star, \nu_\star) \in \{-\mathrm{u}(x_\star, Ax_\star)\} \times [-(A^*)^{-1}(\partial f(x_\star)) \cap \partial g(Ax_\star)].
\end{aligned} \tag{A.19}$$

Moreover, from (A.19) and (A.17), we have

$$\begin{aligned}
&(x_\star, y_\star, \nu_\star) \in \operatorname{Fix}(\mathbf{T}_{\mathrm{LAL}}) \\
&\Rightarrow -\mathrm{u}(x_\star, y_\star, \nu_\star) \in \bigcup_{x \in \mathscr{S}_p} \{-\mathrm{u}(x, Ax)\} \times [-(A^*)^{-1}(\partial f(x)) \cap \partial g(Ax)].
\end{aligned}$$

Since $(x_\star, y_\star, \nu_\star)$ is chosen arbitrarily from $\operatorname{Fix}(\mathbf{T}_{\mathrm{LAL}})$, we have

$$-\mathrm{u}\operatorname{Fix}(\mathbf{T}_{\mathrm{LAL}}) \subset \bigcup_{x \in \mathscr{S}_p} \{-\mathrm{u}(x, Ax)\} \times \big[-(A^*)^{-1}(\partial f(x)) \cap \partial g(Ax)\big]$$

from which Theorem 23(b) is confirmed.
(c) We have seen in (83) that, under the assumptions of Theorem 23(c), for any vector $x_\star \in \mathscr{X}$,

$$x_\star \in \mathscr{S}_p[\text{in (10)}] \text{ if and only if } (x_\star, x_\star, \ldots, x_\star) = P_D(\mathfrak{X}_\star) \tag{A.20}$$

for some $\mathfrak{X}_\star \in \operatorname{Fix}(\mathbf{T}_{\mathrm{DRS_{II}}})$, where $D = \{(x^{(1)}, \ldots, x^{(m+1)}) \in \mathscr{X}^{m+1} \mid x^{(i)} = x^{(j)} \ (i,j = 1,2,\ldots,m+1)\}$ (see (81)), $H\colon \mathscr{X}^{m+1} \to (-\infty,\infty]\colon (x^{(1)}, \ldots, x^{(m+1)}) \mapsto \sum_{i=1}^m g_i(A_i x^{(i)}) + f(x^{(m+1)})$ (see (80)), and $\mathbf{T}_{\mathrm{DRS_{II}}} = (2\operatorname{prox}_H - \mathrm{I}) \circ (2P_D - \mathrm{I})$ (see (78)) [For the availability of prox_H and P_D as computational tools, see Remark 18(a)].

Choose $\mathfrak{X}_\star := (\zeta_\star^{(1)}, \ldots, \zeta_\star^{(m+1)}) \in \operatorname{Fix}(\mathbf{T}_{\mathrm{DRS_{II}}})$ arbitrarily, and let $\mathbf{X}_\star := (x_\star, \ldots, x_\star) = P_D(\mathfrak{X}_\star)$. Then we have

$$\begin{aligned} &\mathfrak{X}_\star \in \operatorname{Fix}(\mathbf{T}_{\mathrm{DRS_{II}}}) \text{ and } P_D(\mathfrak{X}_\star) = \mathbf{X}_\star \\ \Leftrightarrow\ &(2\operatorname{prox}_H - \mathrm{I}) \circ (2P_D - \mathrm{I})(\mathfrak{X}_\star) = \mathfrak{X}_\star \text{ and } P_D(\mathfrak{X}_\star) = \mathbf{X}_\star. \end{aligned}$$

Now, by passing similar steps for (A.12)⇒(A.13), we deduce that

$$\begin{aligned} &\mathfrak{X}_\star \in \operatorname{Fix}(\mathbf{T}_{\mathrm{DRS_{II}}}) \text{ and } P_D(\mathfrak{X}_\star) = \mathbf{X}_\star \\ \Rightarrow\ &\mathbf{X}_\star - \mathfrak{X}_\star \in \partial H(\mathbf{X}_\star) = \left[\bigtimes_{j=1}^m \partial(g_j \circ A_j)(x_\star)\right] \times \partial f(x_\star) \\ \Leftrightarrow\ &(j = 1,2,\ldots,m)\ x_\star - \zeta_\star^{(j)} \in \partial(g_j \circ A_j)(x_\star) \text{ and } x_\star - \zeta_\star^{(m+1)} \in \partial f(x_\star) \\ \Leftrightarrow\ &(j = 1,2,\ldots,m)\ x_\star - \zeta_\star^{(i)} \in A_j^* \partial g_j(A_j x_\star) \text{ and } x_\star - \zeta_\star^{(m+1)} \in \partial f(x_\star), \end{aligned} \tag{A.21}$$

where the last equivalence follows from Fact 4(c) (applicable due to $\operatorname{ri}(\operatorname{dom}(g_j) - \operatorname{ran}(A_j)) = \operatorname{ri}(\operatorname{dom}(g_j) - \mathbb{R}) = \mathbb{R} \ni 0$). Meanwhile, we have

$$\mathbf{X}_\star = P_D(\mathfrak{X}_\star) \ \Leftrightarrow\ x_\star = \frac{1}{m+1}\sum_{i=1}^{m+1} \zeta_\star^{(i)} \ \Leftrightarrow\ x_\star - \zeta_\star^{(m+1)} = -\sum_{i=1}^m (x_\star - \zeta_\star^{(i)}). \tag{A.22}$$

Equations (A.22) and (A.21) imply

$$\begin{aligned} &\mathfrak{X}_\star \in \operatorname{Fix}(\mathbf{T}_{\mathrm{DRS_{II}}}) \text{ and } P_D(\mathfrak{X}_\star) = \mathbf{X}_\star \\ \Rightarrow\ &\begin{cases} (j = 1,2,\ldots,m)\ x_\star - \zeta_\star^{(j)} \in A_j^* \partial g_j(A_j x_\star) \\ x_\star - \zeta_\star^{(m+1)} \in \partial f(x_\star) \cap [-\sum_{i=1}^m A_i^* \partial g_i(A_i x_\star)] \end{cases} \\ \Rightarrow\ &\mathbf{X}_\star - \mathfrak{X}_\star \in \left[\bigtimes_{j=1}^m A_j^* \partial g_j(A_j x_\star)\right] \times \left[\partial f(x_\star) \cap \left(-\sum_{i=1}^m A_i^* \partial g_i(A_i x_\star)\right)\right]. \end{aligned} \tag{A.23}$$

Moreover, by noting that (A.20) ensures $x_\star \in \mathscr{S}_p$, we have from (A.23)

$$\begin{aligned} &\mathfrak{X}_\star \in \operatorname{Fix}(\mathbf{T}_{\mathrm{DRS_{II}}}) \text{ and } P_D(\mathfrak{X}_\star) = \mathbf{X}_\star = (x_\star, \ldots, x_\star) \\ \Rightarrow\ &\mathfrak{X}_\star \in \mathscr{S}_p^{m+1} - \bigcup_{x \in \mathscr{S}_p} \left(\left[\bigtimes_{j=1}^m A_j^* \partial g_j(A_j x)\right] \times \left[\partial f(x) \cap \left(-\sum_{i=1}^m A_i^* \partial g_i(A_i x)\right)\right]\right). \end{aligned}$$

Since $\mathfrak{X}_\star$ is chosen arbitrarily from $\operatorname{Fix}(\mathbf{T}_{\mathrm{DRS_{II}}})$, we have

$$\operatorname{Fix}(\mathbf{T}_{\mathrm{DRS_{II}}}) \subset \mathscr{S}_p^{m+1} - \bigcup_{x \in \mathscr{S}_p} \left(\left[\bigtimes_{j=1}^{m} A_j^* \partial g_j(A_j x) \right] \times \left[\partial f(x) \cap \left(-\sum_{i=1}^{m} A_i^* \partial g_i(A_i x) \right) \right] \right),$$

from which Theorem 23(c) is confirmed. □

G: Proof of Lemma 27

Obviously, we have from (134)

$$(j = 1,2,\ldots,2p) \quad \operatorname{dom}(g_{(j,q)}) \supset \{\eta \in \mathbb{R} \mid \eta > \mathbf{x}_j^\top \mathbf{z}\} \times \mathbb{R}^N. \tag{A.24}$$

By recalling $0 \neq \mathbf{x}_j \in \mathbb{R}^N$ in (129) and $\mathbf{M}_j \in \mathbb{R}^{(N+1)\times p}$ in (135), we have

$$(j = 1,2,\ldots,2p) \begin{bmatrix} \|\mathbf{X}^\top \mathbf{x}_j\| \geq \|\mathbf{x}_j\|^2 > 0 \\ t(\|\mathbf{X}^\top \mathbf{x}_j\|)^{-2} \mathbf{M}_j \mathbf{X}^\top \mathbf{x}_j = \begin{pmatrix} t \\ t\,\frac{\mathbf{X}\mathbf{X}^\top \mathbf{x}_j}{\|\mathbf{X}^\top \mathbf{x}_j\|^2} \end{pmatrix} \quad (\forall t \in \mathbb{R}), \end{bmatrix}$$

and therefore

$$(j = 1,2,\ldots,2p) \quad \mathbf{M}_j \operatorname{dom}(\|\cdot\|_1) = \mathbf{M}_j(\mathbb{R}^p) \supset \operatorname{span} \begin{pmatrix} 1 \\ \frac{\mathbf{X}\mathbf{X}^\top \mathbf{x}_j}{\|\mathbf{X}^\top \mathbf{x}_j\|^2} \end{pmatrix}. \tag{A.25}$$

To prove $\operatorname{dom}(g_{(j,q)}) - \mathbf{M}_j \operatorname{dom}(\|\cdot\|_1) = \mathbb{R} \times \mathbb{R}^N$, choose arbitrarily $(\eta, \mathbf{y}) \in \mathbb{R} \times \mathbb{R}^N$. Then (A.24) and (A.25) guarantee

$$\begin{aligned} \begin{pmatrix} \eta \\ \mathbf{y} \end{pmatrix} &= \begin{pmatrix} \mathbf{x}_j^\top \mathbf{z} + 1 \\ \mathbf{y} + (\mathbf{x}_j^\top \mathbf{z} + 1 - \eta)\frac{\mathbf{X}\mathbf{X}^\top \mathbf{x}_j}{\|\mathbf{X}^\top \mathbf{x}_j\|^2} \end{pmatrix} - \begin{pmatrix} \mathbf{x}_j^\top \mathbf{z} + 1 - \eta \\ (\mathbf{x}_j^\top \mathbf{z} + 1 - \eta)\frac{\mathbf{X}\mathbf{X}^\top \mathbf{x}_j}{\|\mathbf{X}^\top \mathbf{x}_j\|^2} \end{pmatrix} \\ &\in \{\tilde{\eta} \in \mathbb{R} \mid \tilde{\eta} > \mathbf{x}_j^\top \mathbf{z}\} \times \mathbb{R}^N - \operatorname{span} \begin{pmatrix} 1 \\ \frac{\mathbf{X}\mathbf{X}^\top \mathbf{x}_j}{\|\mathbf{X}^\top \mathbf{x}_j\|^2} \end{pmatrix} \\ &\subset \operatorname{dom}(g_{(j,q)}) - \mathbf{M}_j \operatorname{dom}(\|\cdot\|_1), \end{aligned}$$

implying thus

$$\operatorname{ri}(\operatorname{dom}(g_{(j,q)}) - \mathbf{M}_j \operatorname{dom}(\|\cdot\|_1)) = \operatorname{ri}(\mathbb{R} \times \mathbb{R}^N) = \mathbb{R} \times \mathbb{R}^N \ni 0. \tag{A.26}$$

□

H: Proof of Theorem 28

By recalling Remark 29 in Section 5.2, it is sufficient to prove Claim 28, for which we use the following inequality: for each $j = 1,2,\ldots,2p$,

$$(\forall (\eta, \mathbf{y}) \in \mathbb{R} \times \mathbb{R}^N) \quad \left\| \mathbf{M}_j^\top \begin{pmatrix} \eta \\ \mathbf{y} \end{pmatrix} \right\| \geq \left| \eta \|\mathbf{x}_j\|^2 + \langle \mathbf{x}_j, \mathbf{y} \rangle \right|, \tag{A.27}$$

where $\mathbf{x}_j \in \mathbb{R}^N$ in (129) and $\mathbf{M}_j \in \mathbb{R}^{(N+1)\times p}$ in (135). Equation (A.27) is confirmed by

$$(j=1,2,\ldots,2p)(\forall(\eta,\mathbf{y})\in\mathbb{R}\times\mathbb{R}^N) \quad \mathbf{M}_j^\top \begin{pmatrix}\eta\\ \mathbf{y}\end{pmatrix} = \eta\mathbf{X}^\top\mathbf{x}_j+\mathbf{X}^\top\mathbf{y}$$

and

$$\begin{cases} \left[\eta\mathbf{X}^\top\mathbf{x}_j+\mathbf{X}^\top\mathbf{y}\right]_j = \eta\|\mathbf{x}_j\|^2+\langle\mathbf{x}_j,\mathbf{y}\rangle & \text{if } j\in\{1,2,\ldots,p\}\\ \left[\eta\mathbf{X}^\top\mathbf{x}_j+\mathbf{X}^\top\mathbf{y}\right]_{j-p} = -\eta\|\mathbf{x}_j\|^2-\langle\mathbf{x}_j,\mathbf{y}\rangle & \text{if } j\in\{p+1,p+2,\ldots,2p\}. \end{cases}$$

Let $U_S := \sup\{\|\mathbf{b}\| \mid \mathbf{b}\in S\}(<\infty)$. By supercoercivity of φ and Example 3, the subdifferential of its perspective $\widetilde{\varphi}$ at each $(\eta,\mathbf{y})\in\mathbb{R}\times\mathbb{R}^N$ can be expressed as (25), and thus, to prove Claim 28, it is sufficient to show

(i) $(\mathbf{M}_j^\top)^{-1}(S)\cap\partial\widetilde{\varphi}(\mathbb{R}_{++}\times\mathbb{R}^N)$ is bounded;

(ii) $(\mathbf{M}_j^\top)^{-1}(S)\cap\partial\widetilde{\varphi}(0,0)$ is bounded.

Proof of (i) Choose $(\eta,\mathbf{y})\in\mathbb{R}_{++}\times\mathbb{R}^N$ arbitrarily. Then, from (25), every $\mathbf{c}_{(\eta,\mathbf{y})}\in(\mathbf{M}_j^\top)^{-1}(S)\cap\partial\widetilde{\varphi}(\eta,\mathbf{y})\in\mathbb{R}\times\mathbb{R}^N$ can be expressed with some $\mathbf{u}\in\partial\varphi(\mathbf{y}/\eta)$ as

$$\mathbf{c}_{(\eta,\mathbf{y})} = (\varphi(\mathbf{y}/\eta)-\langle\mathbf{y}/\eta,\mathbf{u}\rangle,\mathbf{u}) = (-\varphi^*(\mathbf{u}),\mathbf{u}), \tag{A.28}$$

where the last equality follows from $\varphi(\mathbf{y}/\eta)+\varphi^*(\mathbf{u})=\langle\mathbf{y}/\eta,\mathbf{u}\rangle$ due to the Fenchel-Young identity (23). By $\mathbf{M}_j^\top\mathbf{c}_{(\eta,\mathbf{y})}\in S$ and by applying the inequality (A.27) to (A.28), we have

$$\begin{aligned} U_S \geq \|\mathbf{M}_j^\top\mathbf{c}_{(\eta,\mathbf{y})}\| = \left\|\mathbf{M}_j^\top\begin{pmatrix}-\varphi^*(\mathbf{u})\\ \mathbf{u}\end{pmatrix}\right\| &\geq \left|(-\varphi^*(\mathbf{u}))\|\mathbf{x}_j\|^2+\langle\mathbf{x}_j,\mathbf{u}\rangle\right|\\ &= |\Upsilon(\mathbf{u})| \geq \Upsilon_+(\mathbf{u}), \end{aligned} \tag{A.29}$$

where $\Upsilon\colon\mathbb{R}^N\to\mathbb{R}\colon\mathbf{v}\mapsto\varphi^*(\mathbf{v})\|\mathbf{x}_j\|^2-\langle\mathbf{x}_j,\mathbf{v}\rangle$ and $\Upsilon_+\colon\mathbb{R}^N\to\mathbb{R}\colon\mathbf{v}\mapsto\max\{\Upsilon(\mathbf{v}),0\}$ are coercive convex functions (see Section 2.1) and independent from the choice of $(\eta,\mathbf{y})$. The coercivity of Υ_+ ensures the existence of an open ball $B(0,\hat{U}_{(\mathrm{i})})$ of radius $\hat{U}_{(\mathrm{i})}>0$ such that $\operatorname{lev}_{\leq U_S}\Upsilon_+ := \{\mathbf{v}\in\mathbb{R}^N \mid \Upsilon_+(\mathbf{v})\leq U_S\}\subset B(0,\hat{U}_{(\mathrm{i})})$, and thus (A.29) implies

$$\|\mathbf{u}\|\leq\hat{U}_{(\mathrm{i})}. \tag{A.30}$$

Moreover, by $\mathbf{x}_j\neq 0$, the triangle inequality, the Cauchy-Schwarz inequality, (A.29), and (A.30), we have

$$\begin{aligned} |\varphi^*(\mathbf{u})| &= \left|\frac{\Upsilon(\mathbf{u})}{\|\mathbf{x}_j\|^2}+\frac{\langle\mathbf{x}_j,\mathbf{u}\rangle}{\|\mathbf{x}_j\|^2}\right| \leq \left|\frac{\Upsilon(\mathbf{u})}{\|\mathbf{x}_j\|^2}\right|+\left|\frac{\langle\mathbf{x}_j,\mathbf{u}\rangle}{\|\mathbf{x}_j\|^2}\right|\\ &\leq \left|\frac{\Upsilon(\mathbf{u})}{\|\mathbf{x}_j\|^2}\right|+\frac{\|\mathbf{u}\|}{\|\mathbf{x}_j\|} \leq \frac{U_S}{\|\mathbf{x}_j\|^2}+\frac{\hat{U}_{(\mathrm{i})}}{\|\mathbf{x}_j\|} =: U_{(\mathrm{i})}, \end{aligned}$$

which yields $\mathbf{c}_{(\eta,\mathbf{y})}=(-\varphi^*(\mathbf{u}),\mathbf{u})\in[-U_{(\mathrm{i})},U_{(\mathrm{i})}]\times B(0,\hat{U}_{(\mathrm{i})})$. Since $(\eta,\mathbf{y})\in\mathbb{R}_{++}\times\mathbb{R}^N$ is chosen arbitrarily and $\mathbf{c}_{(\eta,\mathbf{y})}\in(\mathbf{M}_j^\top)^{-1}(S)\cap\partial\widetilde{\varphi}(\eta,\mathbf{y})$ is also chosen arbitrarily, we have

$$(\mathbf{M}_j^\top)^{-1}(S)\cap\partial\widetilde{\varphi}(\mathbb{R}_{++}\times\mathbb{R}^N)\subset[-U_{(\mathrm{i})},U_{(\mathrm{i})}]\times B(0,\hat{U}_{(\mathrm{i})}),$$

which confirms the statement (i).
Proof of (ii) By introducing

$$\mathfrak{B} := \left\{ \mathbf{v} \in \mathbb{R}^N \,\middle|\, \left| \left\langle \frac{2}{\|\mathbf{x}_j\|^2}\mathbf{x}_j, \mathbf{v} \right\rangle \right| > |\varphi^*(\mathbf{v})| \right\}, \tag{A.31}$$

we can decompose the set $(\mathbf{M}_j^\top)^{-1}(S) \cap \partial\widetilde{\varphi}(0,0)$ into

$$(\mathbf{M}_j^\top)^{-1}(S) \cap \partial\widetilde{\varphi}(0,0) \cap (\mathbb{R} \times \mathfrak{B}) \text{ and } (\mathbf{M}_j^\top)^{-1}(S) \cap \partial\widetilde{\varphi}(0,0) \cap (\mathbb{R} \times \mathfrak{B}^c). \tag{A.32}$$

In the following, we show the boundedness of each set in (A.32).

First, we show the boundedness of $\mathfrak{B}$ by contradiction. Suppose that $\mathfrak{B} \not\subset B(0,r)$ for all $r > 0$. Then there exists a sequence $(\mathbf{u}_k)_{k\in\mathbb{N}} \subset \mathbb{R}^N$ such that

$$(\forall k \in \mathbb{N})\ \frac{2}{\|\mathbf{x}_j\|} \geq \left| \left\langle \frac{2}{\|\mathbf{x}_j\|^2}\mathbf{x}_j, \frac{\mathbf{u}_k}{\|\mathbf{u}_k\|} \right\rangle \right| > \frac{|\varphi^*(\mathbf{u}_k)|}{\|\mathbf{u}_k\|} \text{ and } \|\mathbf{u}_k\| \geq k,$$

which contradicts the supercoercivity of φ^*, implying thus the existence of $r_* > 0$ such that $\mathfrak{B} \subset B(0,r_*)$.

Next, we show the boundedness of the former set in (A.32). Choose arbitrarily

$$(\mu, \mathbf{u}) \in (\mathbf{M}_j^\top)^{-1}(S) \cap \partial\widetilde{\varphi}(0,0) \cap (\mathbb{R} \times \mathfrak{B}).$$

By $\mathbf{x}_j \neq 0$, $\mathbf{M}_j^\top(\mu, \mathbf{u}^\top)^\top \in S \subset B(0,U_S)$, the inequality (A.27), the triangle inequality, the Cauchy-Schwarz inequality, and $\mathbf{u} \in \mathfrak{B} \subset B(0,r_*)$, we have

$$\begin{aligned} \frac{U_S}{\|\mathbf{x}_j\|^2} &\geq \frac{1}{\|\mathbf{x}_j\|^2} \left\| \mathbf{M}_j^\top \begin{pmatrix} \mu \\ \mathbf{u} \end{pmatrix} \right\| \geq \frac{1}{\|\mathbf{x}_j\|^2} |\mu \|\mathbf{x}_j\|^2 + \langle \mathbf{x}_j, \mathbf{u} \rangle| \\ &\geq |\mu| - \left| \left\langle \frac{\mathbf{x}_j}{\|\mathbf{x}_j\|^2}, \mathbf{u} \right\rangle \right| \geq |\mu| - \frac{\|\mathbf{u}\|}{\|\mathbf{x}_j\|} \geq |\mu| - \frac{r_*}{\|\mathbf{x}_j\|} \end{aligned}$$

which yields

$$\hat{U}_{(\mathrm{iia})} := \frac{U_S}{\|\mathbf{x}_j\|^2} + \frac{r_*}{\|\mathbf{x}_j\|} \geq |\mu|.$$

Therefore, we have $(\mu, \mathbf{u}) \in [-\hat{U}_{(\mathrm{iia})}, \hat{U}_{(\mathrm{iia})}] \times B(0, r_\star)$. Since $(\mu, \mathbf{u}) \in (\mathbf{M}_j^\top)^{-1}(S) \cap \partial\widetilde{\varphi}(0,0) \cap (\mathbb{R} \times \mathfrak{B})$ is chosen arbitrarily, we have

$$(\mathbf{M}_j^\top)^{-1}(S) \cap \partial\widetilde{\varphi}(0,0) \cap (\mathbb{R} \times \mathfrak{B}) \subset [-\hat{U}_{(\mathrm{iia})}, \hat{U}_{(\mathrm{iia})}] \times B(0, r_\star). \tag{A.33}$$

Finally, we show the boundedness of the latter set in (A.32). Let

$$(\mu, \mathbf{u}) \in (\mathbf{M}_j^\top)^{-1}(S) \cap \partial\widetilde{\varphi}(0,0) \cap (\mathbb{R} \times \mathfrak{B}^c). \tag{A.34}$$

From (25), we have

$$\partial\widetilde{\varphi}(0,0) = \{(\mu', \mathbf{u}') \in \mathbb{R} \times \mathbb{R}^N \mid \mu' + \varphi^*(\mathbf{u}') \leq 0\}. \tag{A.35}$$

Note that coercivity of φ^* ($\Rightarrow \exists \min \varphi^*(\mathbb{R}^N) \in \mathbb{R}$, see Fact 2) and (A.35) yield $\varphi^*(\mathbf{u}) \in [\min \varphi^*(\mathbb{R}^N), -\mu]$ and thus

$$|\varphi^*(\mathbf{u})| \leq \max\{|\min \varphi^*(\mathbb{R}^N)|, |\mu|\} \leq |\min \varphi^*(\mathbb{R}^N)| + |\mu|. \tag{A.36}$$

By $\mathbf{x}_j \neq 0$, $\mathbf{M}_j^\top(\mu, \mathbf{u}^\top)^\top \in S \subset B(0,U_S)$ (see (A.34)), the inequality (A.27), the triangle inequality, $\mathbf{u} \in \mathfrak{B}^c$ (see (A.34) and (A.31)), and (A.36), we have

$$\begin{aligned}\frac{2}{\|\mathbf{x}_j\|^2}U_S &\geq \frac{2}{\|\mathbf{x}_j\|^2}\left\|\mathbf{M}_j^\top\begin{pmatrix}\mu\\ \mathbf{u}\end{pmatrix}\right\| \geq \frac{2}{\|\mathbf{x}_j\|^2}|\mu\|\mathbf{x}_j\|^2+\langle\mathbf{x}_j,\mathbf{u}\rangle|\\ &\geq 2|\mu|-\left|\left\langle\frac{2}{\|\mathbf{x}_j\|^2}\mathbf{x}_j,\mathbf{u}\right\rangle\right| \geq 2|\mu|-|\varphi^*(\mathbf{u})|\\ &\geq 2|\mu|-|\min\varphi^*(\mathbb{R}^N)|-|\mu| = |\mu|-|\min\varphi^*(\mathbb{R}^N)|\end{aligned}$$

and thus, with (A.36),

$$\hat{U}_{(\mathrm{iib})} := \frac{2}{\|\mathbf{x}_j\|^2}U_S+2|\min\varphi^*(\mathbb{R}^N)| \geq |\mu|+|\min\varphi^*(\mathbb{R}^N)| \geq |\varphi^*(\mathbf{u})| \geq \varphi^*(\mathbf{u}).$$

Hence, we have

$$(\mu,\mathbf{u}) \in [-\hat{U}_{(\mathrm{iib})},\hat{U}_{(\mathrm{iib})}]\times\mathrm{lev}_{\leq\hat{U}_{(\mathrm{iib})}}(\varphi^*).$$

Since $(\mu,\mathbf{u}) \in (\mathbf{M}_j^\top)^{-1}(S)\cap\partial\widetilde{\varphi}(0,0)\cap(\mathbb{R}\times\mathfrak{B}^c)$ is chosen arbitrarily, we have

$$(\mathbf{M}_j^\top)^{-1}(S)\cap\partial\widetilde{\varphi}(0,0)\cap(\mathbb{R}\times\mathfrak{B}^c) \subset [-\hat{U}_{(\mathrm{iib})},\hat{U}_{(\mathrm{iib})}]\times\mathrm{lev}_{\leq\hat{U}_{(\mathrm{iib})}}(\varphi^*). \tag{A.37}$$

Consequently, by using (A.33) and (A.37) and by letting $U_{(\mathrm{ii})} := \max\{\hat{U}_{(\mathrm{iia})},\hat{U}_{(\mathrm{iib})}\}$, we have

$$(\mathbf{M}_j^\top)^{-1}(S)\cap\partial\widetilde{\varphi}(0,0) \subset [-U_{(\mathrm{ii})},U_{(\mathrm{ii})}]\times[\mathrm{lev}_{\leq U_{(\mathrm{ii})}}(\varphi^*)\cup B(0,r_\star)],$$

which guarantees the boundedness of $(\mathbf{M}_j^\top)^{-1}(S)\cap\partial\widetilde{\varphi}(0,0)$, due to the coercivity of φ^*, implying thus finally the statement (ii).

□

References

[1] Argyriou, A., Baldassarre, L., Micchelli, C.A., Pontil, M.: On sparsity inducing regularization methods for machine learning. In: B. Schölkopf, Z. Luo, V. Vovk (eds.) Empirical Inference, pp. 205–216. Springer Berlin, Heidelberg (2013)

[2] Aronszajn, N.: Theory of reproducing kernels. Trans. Amer. Math. Soc. **68**, 337–404 (1950)

[3] Attouch, H.: Viscosity solutions of minimization problems. SIAM J. Optim. **6**, 769–806 (1996)

[4] Attouch, H., Cabot, A., Chbani, Z., Riahi, H.: Accelerated forward-backward algorithms with perturbations. Application to Tikhonov regularization. (preprint)

[5] Baillon, J.-B., Bruck, R.E., Reich, S.: On the asymptotic behavior of nonexpansive mappings and semigroups in Banach spaces. Houst. J. Math. **4**, 1–9 (1978)

[6] Bauschke, H.H.: The approximation of fixed points of compositions of nonexpansive mappings in Hilbert space. J. Math. Anal. Appl. **202**, 150–159 (1996)

[7] Bauschke, H.H., Borwein, J.M.: On projection algorithms for solving convex feasibility problems. SIAM Rev. **38**, 367–426 (1996)

[8] Bauschke, H.H., Combettes, P.L.: A weak-to-strong convergence principle for Fejér monotone methods in Hilbert space. Math. Oper. Res. **26**, 248–264 (2001)

[9] Bauschke, H.H., Combettes, P.L.: Convex Analysis and Monotone Operator Theory in Hilbert Space, 2nd edn. Springer (2017)

[10] Bauschke, H.H., Moursi, M.: On the Douglas-Rachford algorithm. Math. Program. **164**, 263–284 (2017)

[11] Beck, A., Teboulle, M.: Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems. IEEE Trans. Image Process. **18**, 2419–2434 (2009)

[12] Ben-Israel, A., Greville, T.N.E.: Generalized Inverses: Theory and Applications, 2nd edn. Springer-Verlag (2003)

[13] Bien, J., Gaynanova, I., Lederer, J., Müller, C.L.: Non-convex global minimization and false discovery rate control for the TREX. J. Comput. Graph. Stat. **27**, 23–33 (2018)

[14] Bishop, C.M.: Machine Learning and Pattern Recognition. Information Science and Statistics. Springer, Heidelberg (2006)

[15] Blum, A., Rivest, R.L.: Training a 3-node neural network is NP-complete. Neural Networks **5**, 117–127 (1992)

[16] Boser, B.E., Guyon, I.M., Vapnik, V.N.: A training algorithm for optimal margin classifiers. In: Proc. the 5th Annual ACM Workshop on Computational Learning Theory (COLT), pp. 144–152 (1992)

[17] Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Found. Trends® Mach. Learn. **3**, 1–122 (2011)

[18] Cabot, A.: Proximal point algorithm controlled by a slowly vanishing term: Applications to hierarchical minimization. SIAM J. Optim. **15**, 555–572 (2005)

[19] Candler, W., Norton, R.: Multilevel programming. Technical Report 20, World Bank Development Research Center, Washington D.C., USA (1977)

[20] Cegielski, A.: Iterative Methods for Fixed Point Problems in Hilbert Spaces. Springer (2012)

[21] Censor, Y., Davidi, R., Herman, G.T.: Perturbation resilience and superiorization of iterative algorithms. Inverse Probl. **26**, 065008 (2010)

[22] Censor, Y., Zenios, S.A.: Parallel Optimization: Theory, Algorithm, and Optimization. Oxford University Press (1997)

[23] Chaari, L., Ciuciu, P., Mériaux, S., Pesquet, J.C.: Spatio-temporal wavelet regularization for parallel MRI reconstruction: Application to functional MRI. Magn. Reson. Mater. Phys. Biol. Med. **27**, 509–529 (2014)

[24] Chambolle, A., Dossal, C.: On the convergence of the iterates of the "fast iterative shrinkage/thresholding algorithm". J. Optim. Theory Appl. **166**, 968–982 (2015)

[25] Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Trans. Intell. Syst. Technol. **2**, 1–27 (2011)

[26] Chaux, C., Pesquet, J.C., Pustelnik, N.: Nested iterative algorithms for convex constrained image recovery problems. SIAM J. Imaging Sci. **2**, 730–762 (2009)

[27] Chidume, C.: Geometric Properties of Banach Spaces and Nonlinear Iterations (Chapter 7: Hybrid steepest descent method for variational inequalities). vol. 1965 of Lecture Notes in Mathematics. Springer (2009)

[28] Chipman, J.S.: Linear restrictions, rank reduction, and biased estimation in linear regression. Linear Algebra Appl. **289**, 55–74 (1999)

[29] Chipman, J.S., Rao, M.M.: The treatment of linear restrictions in regression analysis. Econometrics **32**, 198–204 (1964)

[30] Coloson, B., Marcotte, P., Savard, G.: An overview of bilevel optimization. Ann. Oper. Res. **153**, 235–256 (2007)

[31] Combettes, P.L.: The foundations of set theoretic estimation. Proc. IEEE **81**, 182–208 (1993)

[32] Combettes, P.L.: Inconsistent signal feasibility problems: Least squares solutions in a product space. IEEE Trans. Signal Process. **42**, 2955–2966 (1994)

[33] Combettes, P.L.: Strong convergence of block-iterative outer approximation methods for convex optimization. SIAM J. Control Optim. **38**, 538–565 (2000)

[34] Combettes, P.L.: Iterative construction of the resolvent of a sum of maximal monotone operators. J. Convex Anal. **16**, 727–748 (2009)

[35] Combettes, P.L.: Perspective functions: Properties, constructions, and examples. Set-Valued Var. Anal. **26**, 247–264 (2017)

[36] Combettes, P.L., Bondon, P.: Hard-constrained inconsistent signal feasibility problems. IEEE Trans. Signal Process. **47**, 2460–2468 (1999)

[37] Combettes, P.L., Hirstoaga, S.A.: Approximating curves for nonexpansive and monotone operators. J. Convex Anal. **13**, 633–646 (2006)

[38] Combettes, P.L., Müller, C.L.: Perspective functions: Proximal calculus and applications in high-dimensional statistics. J. Math. Anal. Appl. **457**, 1283–1306 (2018)

[39] Combettes, P.L., Pesquet, J.-C.: Image restoration subject to a total variation constraint. IEEE Trans. Image Process. **13**, 1213–1222 (2004)

[40] Combettes, P.L., Pesquet, J.-C.: A Douglas-Rachford splitting approach to nonsmooth convex variational signal recovery. IEEE J. Sel. Top. Signal Process. **1**, 564–574 (2007)

[41] Combettes, P.L., Pesquet, J.-C.: A proximal decomposition method for solving convex variational inverse problems. Inverse Probl. **24**, 065014 (2008)

[42] Combettes, P.L., Pesquet, J.-C.: Proximal splitting methods in signal processing. In: H.H. Bauschke, R. Burachik, P. Combettes, V. Elser, D. Luke, H. Wolkowicz (eds.) Fixed-Point Algorithms for Inverse Problems in Science and Engineering, pp. 185–212. Springer-Verlag (2011)

[43] Combettes, P.L., Pesquet, J.-C.: Stochastic quasi-Fejér block-coordinate fixed point iterations with random sweeping. SIAM J. Optim. **25**, 1221–1248 (2015)

[44] Combettes, P.L., Wajs, V.R.: Signal recovery by proximal forward-backward splitting. SIAM Multiscale Model. Simul. **4**, 1168–1200 (2005)

[45] Combettes, P.L., Yamada, I.: Compositions and convex combinations of averaged nonexpansive operators. J. Math. Anal. Appl. **425**, 55–70 (2015)

[46] Cominetti, R., Courdurier, M.: Coupling general penalty schemes for convex programming with the steepest descent and the proximal point algorithm. SIAM J. Optim. **13**, 745–765 (2002)

[47] Condat, L.: A primal-dual splitting method for convex optimization involving lipschitzian, proximable and linear composite terms. J. Optim. Theory Appl. **158**, 460–479 (2013)

[48] Cortes, C., Vapnik, V.N.: Support-vector networks. Mach. Learn. **20**, 273–297 (1995)

[49] Cover, T.M.: Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. IEEE Trans. Electron. Comput. **14**, 326–334 (1965)

[50] Dalalyan, A.S., Hebiri, M., Lederer, J.: On the prediction performance of the Lasso. Bernoulli **23**, 552–581 (2017)

[51] Deutsch, F.: Best Approximation in Inner Product Spaces. New York: Springer-Verlag (2001)

[52] Deutsch, F., Yamada, I.: Minimizing certain convex functions over the intersection of the fixed point sets of nonexpansive mappings. Numer. Funct. Anal. Optim. **19**, 33–56 (1998)

[53] Donoho, D.L.: De-noising by soft-thresholding. IEEE Trans. Inf. Theory **41**, 613–627 (1995)

[54] Donoho, D.L., Johnstone, I.M.: Ideal spatial adaptation via wavelet shrinkage. Biometrika **81**, 425–455 (1994)

[55] Dontchev, A.L., Zolezzi, T.: Well-posed optimization problems. vol. 1543 of Lecture Notes in Mathematics. Springer-Verlag (1993)

[56] Dotson Jr., W.G.: On the Mann iterative process. Trans. Amer. Math. Soc. **149**, 65–73 (1970)

[57] Douglas, J., Rachford, H.H.: On the numerical solution of heat conduction problems in two or three space variables. Trans. Amer. Math. Soc. **82**, 421–439 (1956)

[58] Dupé, F.X., Fadili, M.J., Starck, J.-L.: A proximal iteration for deconvolving Poisson noisy images using sparse representations. IEEE Trans. Image Process. **18**, 310–321 (2009)

[59] Dupé, F.X., Fadili, M.J., Starck, J.-L.: Deconvolution under Poisson noise using exact data fidelity and synthesis or analysis sparsity priors. Stat. Methodol. **9**, 4–18 (2012)

[60] Durand, S., Fadili, M.J., Nikolova, M.: Multiplicative noise removal using L1 fidelity on frame coefficients. J. Math. Imaging Vision **36**, 201–226 (2010)

[61] Eckstein, J., Bertsekas, D.P.: On the Douglas-Rachford splitting method and proximal point algorithm for maximal monotone operators. Math. Program. **55**, 293–318 (1992)

[62] Eckstein, J., Yao, W.: Understanding the convergence of the alternating direction method of multipliers: Theoretical and computational perspectives. Pac. J. Optim. **11**, 619–644 (2015)

[63] Eicke, B.: Iteration methods for convexly constrained ill-posed problems in Hilbert space. Numer. Funct. Anal. Optim. **13**, 413–429 (1992)

[64] Ekeland, I., Themam, R.: Convex Analysis and Variational Problems. Classics in Applied Mathematics 28. SIAM (1999)

[65] Fisher, A.R.: The use of multiple measurements in taxonomic problems. Ann. Hum. Genet. **7**, 179–188 (1936)

[66] Gabay, D.: Applications of the method of multipliers to variational inequalities. In: M. Fortin, R. Glowinski (eds.) Augmented Lagrangian Methods: Applications to the solution of boundary value problems. North-Holland, Amsterdam (1983)

[67] Gandy, S., Recht, B., Yamada, I.: Tensor completion and low-n-rank tensor recovery via convex optimization. Inverse Probl. **27**, 025010 (2011)

[68] Gandy, S., Yamada, I.: Convex optimization techniques for the efficient recovery of a sparsely corrupted low-rank matrix. J. Math-For-Industry **2**, 147–156 (2010)

[69] van de Geer, S., Lederer, J.: The Lasso, correlated design, and improved oracle inequalities. IMS Collections **9**, 303–316 (2013)

[70] Goebel, K., Reich, S.: Uniform Convexity, Hyperbolic Geometry, and Nonexpansive Mappings. Marcel Dekker, New York (1984)

[71] Groetsch, C. W.: A note on segmenting Mann iterates. J. Math. Anal. Appl. **40**, 369–372 (1972)

[72] Halpern, B.: Fixed points of nonexpanding maps. Bull. Amer. Math. Soc. **73**, 957–961 (1967)

[73] Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning, 2nd edn. Springer Series in Statistics (2009)

[74] Hastie, T., Tibshirani, R., Wainwright, M.: Statistical Learning with Sparsity: The Lasso and Generalizations. CRC press (2015)

[75] Haugazeau, Y.: Sur les inéquations variationnelles et la minimisation de fonctionnelles convexes. Thèse, Universite de Paris (1968)

[76] He, B., Yuan, X.: On the $O(1/n)$ convergence rate of the Douglas-Rachford alternating direction method. SIAM J. Numer. Anal. **50**, 700–709 (2012)

[77] Hebiri, M., Lederer, J.: How correlations influence Lasso prediction. IEEE Trans. Inf. Theory **59**, 1846–1854 (2013)

[78] Helou, E.S., De Pierro, A.R.: On perturbed steepest descent methods with inexact line search for bilevel convex optimization. Optimization **60**, 991–1008 (2011)

[79] Helou, E.S., Simões, L.E.A.: ε-subgradient algorithms for bilevel convex optimization. Inverse Probl. **33**, 055020 (2017)

[80] Herman, G.T., Garduño, E., Davidi, R., Censor, Y.: Superiorization: An optimization heuristic for medical physics. Med. Phys. **39**, 5532–5546 (2012)

[81] Hestenes, M.R.: Multiplier and gradient methods. J. Optim. Theory Appl. **4**, 303–320 (1969)

[82] Hiriart-Urruty, J.-B., Lemaréchal, C.: Convex Analysis and Minimization Algorithms. Springer (1993)

[83] Iemoto, S., Takahashi, W.: Strong convergence theorems by a hybrid steepest descent method for countable nonexpansive mappings in Hilbert spaces. Sci. Math. Jpn. **69**, 227–240 (2009)

[84] Judd, J.S.: Learning in networks is hard. In: Proc. 1st Int. Conf. Neural Networks, pp. 685–692 (1987)

[85] Kailath, T., Sayed, A.H., Hassibi, B.: Linear Estimation. Prentice-Hall (2000)

[86] Kitahara, D., Yamada, I.: Algebraic phase unwrapping based on two-dimensional spline smoothing over triangles. IEEE Trans. Signal Process. **64**, 2103–2118 (2016)

[87] Koltchinskii, V., Lounici, K., Tsybakov, A.: Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. Ann. Statist. **39**, 2302–2329 (2011)

[88] Krasnosel'skiĭ, M.A.: Two remarks on the method of successive approximations. Uspekhi Mat. Nauk **10**, 123–127 (1955)

[89] Lederer, J., Müller, C.L.: Don't fall for tuning parameters: Tuning-free variable selection in high dimensions with the TREX. In: Proc. Twenty-Ninth AAAI Conf. Artif. Intell., pp. 2729–2735 (2015)

[90] Lions, P.L.: Approximation de points fixes de contractions. C. R. Acad. Sci. Paris Sèrie A-B **284**, 1357–1359 (1977)

[91] Lions, P.L., Mercier, B.: Splitting algorithms for the sum of two nonlinear operators. SIAM J. Numer. Anal. **16**, 964–979 (1979)

[92] Lobo, M.S., Vandenberghe, L., Boyd, S., Lebret, H.: Applications of second-order cone programming. Linear Algebra Appl. **284**, 193–228 (1998)

[93] Luenberger, D.G.: Optimization by Vector Space Methods. Wiley (1969)

[94] Mainge, P.E.: Extension of the hybrid steepest descent method to a class of variational inequalities and fixed point problems with nonself-mappings. Numer. Funct. Anal. Optim. **29**, 820–834 (2008)

[95] Mangasarian, O.L.: Iterative solution of linear programs. SIAM J. Numer. Amal. **18**, 606–614 (1981)

[96] Mann, W.: Mean value methods in iteration. Proc. Amer. Math. Soc. **4**, 506–510 (1953)

[97] Marquardt, D.W.: Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. Technometrics **12**, 591–612 (1970)

[98] Martinet, B.: Régularisation d'inéquations variationnelles par approximations successives. Rev. Française Informat. Recherche Opérationnelle **4**, 154–159 (1970)

[99] Martinet, B.: Détermination approchée d'un point fixe d'une application pseudo-contractante. C. R. Acad. Sci. Paris Ser. A-B **274**, 163–165 (1972)

[100] Moore, E.H.: On the reciprocal of the general algebraic matrix. Bull. Amer. Math. Soc. **26**, 394–395 (1920)

[101] Moreau, J.J.: Fonctions convexes duales et points proximaux dans un espace hilbertien. C. R. Acad. Sci. Paris Ser. A Math. **255**, 2897–2899 (1962)

[102] Moreau, J.J.: Proximité et dualité dans un espace hilbertien. Bull. Soc. Math. France **93**, 273–299 (1965)

[103] Nesterov, Y.: A method of solving a convex programming problem with convergence rate $O(1/k^2)$. Soviet Math. Dokl. **27**, 372–376 (1983)

[104] Nikazad, T., Davidi, R., Herman, G.T.: Accelerated perturbation-resilient block-iterative projection methods with application to image reconstruction. Inverse Probl. **28**, 035005 (2012)

[105] Ogura, N., Yamada, I.: Non-strictly convex minimization over the fixed point set of the asymptotically shrinking nonexpansive mapping. Numer. Funct. Anal. Optim. **23**, 113–137 (2002)

[106] Ogura, N., Yamada, I.: Non-strictly convex minimization over the bounded fixed point set of nonexpansive mapping. Numer. Funct. Anal. Optim. **24**, 129–135 (2003)

[107] Ono, S., Yamada, I.: Hierarchical convex optimization with primal-dual splitting. IEEE Trans. Signal Process. **63**, 373–388 (2014)

[108] Ono, S., Yamada, I.: Signal recovery with certain involved convex data-fidelity constraints. IEEE Trans. Signal Process. **63**, 6149–6163 (2015)

[109] Passty, G.B.: Ergodic convergence to a zero of the sum of monotone operators in Hilbert space. J. Math. Anal. Appl. **72**, 383–390 (1979)

[110] Penfold, S.N., Schulte, R.W., Censor, Y., Rosenfeld, A.B.: Total variation superiorization schemes in proton computed tomography image reconstruction. Med. Phys. **37**, 5887–5895 (2010)

[111] Penrose, R.: A generalized inverse for matrices. Proc. Cambridge Philos. Soc. **51**, 406–413 (1955)

[112] Piotrowski, T., Cavalcante, R., Yamada, I.: Stochastic MV-PURE estimator? Robust reduced-rank estimator for stochastic linear model. IEEE Trans. Signal Process. **57**, 1293–1303 (2009)

[113] Piotrowski, T., Yamada, I.: MV-PURE estimator: Minimum-variance pseudo-unbiased reduced-rank estimator for linearly constrained ill-conditioned inverse problems. IEEE Trans. Signal Process. **56**, 3408–3423 (2008)

[114] Polyak, B.T.: Sharp minimum. International Workshop on Augmented Lagrangians (1979)

[115] Potter, L.C., Arun, K.S.: A dual approach to linear inverse problems with convex constraints. SIAM J. Control Optim. **31**, 1080–1092 (1993)

[116] Powell, M.J.D.: A method for nonlinear constraints in minimization problems. In: R. Fretcher (ed.) Optimization, pp. 283–298. Academic Press (1969)

[117] Pustelnik, N., Chaux, C., Pesquet, J.-C.: Parallel proximal algorithm for image restoration using hybrid regularization. IEEE Trans. Image Process. **20**, 2450–2462 (2011)

[118] Rao, C.R., Mitra, S.K.: Generalized Inverse of Matrices and Its Applications. John Wiley & Sons (1971)

[119] Reich, S.: Weak convergence theorems for nonexpansive mappings in Banach spaces. J. Math. Anal. Appl. **67**, 274–276 (1979)

[120] Rigollet, P., Tsybakov, A.: Exponential screening and optimal rates of sparse estimation. Ann. Statist. **39**, 731–771 (2011)

[121] Rockafellar, R.T.: Monotone operators and proximal point algorithm. SIAM J. Control Optim. **14**, 877–898 (1976)

[122] Rockafellar, R.T., Wets, R.J.-B.: Variational Analysis, 1st edn. Springer (1998)

[123] Sabharwal, A., Potter, L.C.: Convexly constrained linear inverse problems: Iterative least-squares and regularization. IEEE Trans. Signal Process. **46**, 2345–2352 (1998)

[124] Saitoh, S.: Theory of Reproducing Kernels and Its Applications. Longman Scientific & Technical, Harlow (1988)

[125] Schölkopf, B., Luo, Z., Vovk, V.: Empirical Inference. Springer-Verlag (2013)

[126] Schölkopf, B., Smola, A.J.: Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT press (2002)

[127] Solodov, M.: An explicit descent method for bilevel convex optimization. J. Convex Anal. **14**, 227–237 (2007)

[128] Solodov, M.: A bundle method for a class of bilevel nonsmooth convex minimization problems. SIAM J. Optim. **18**, 242–259 (2008)

[129] Takahashi, N., Yamada, I.: Parallel algorithms for variational inequalities over the cartesian product of the intersections of the fixed point sets of nonexpansive mappings. J. Approx. Theory **153**, 139–160 (2008)

[130] Takahashi, W.: Nonlinear Functional Analysis— Fixed Point Theory and its Applications. Yokohama Publishers (2000)

[131] Theodoridis, S.: Machine Learning: Bayesian and Optimization Perspective. Academic Press (2015)

[132] Tibshirani, R.: Regression shrinkage and selection via the Lasso. J. Roy. Statist. Soc. Ser. B **58**, 267–288 (1996)

[133] Tikhonov, A.N.: Solution of incorrectly formulated problems and the regularization method. Soviet Math. Dokl. **4**, 1035–1038 (1963)

[134] Tseng, P.: Applications of a splitting algorithm to decomposition in convex programming and variational inequalities. SIAM J. Control Optim. **29**, 119–138 (1991)

[135] Vapnik, V.N.: Statistical Learning Theory. John Wiley & Sons (1998)

[136] Vapnik, V.N., Lerner, A.: Pattern recognition using generalized portrait method. Automat. Rem. Contr. **24**, 774–780 (1963)

[137] Varga, R.S.: Matrix Iterative Analysis, 2nd edn. Springer, New York (2000)

[138] Vicente, L.N., Calamai, P.H.: Bilevel and multilevel programming: A bibliography review. J. Global Optim. **5**, 291–306 (1994)

[139] Vu, B.C.: A splitting algorithm for dual monotone inclusions involving cocoercive operators. Adv. Comput. Math. **38**, 667–681 (2013)

[140] Xu, H.K., Kim, T.H.: Convergence of hybrid steepest descent methods for variational inequalities. J. Optim. Theory Appl. **119**, 185–201 (2003)

[141] Yamada, I.: Approximation of convexly constrained pseudoinverse by hybrid steepest descent method. In: Proc. IEEE ISCAS (1999)

[142] Yamada, I.: The hybrid steepest descent method for the variational inequality problem over the intersection of fixed point sets of nonexpansive mappings. In: D. Butnariu, Y. Censor, S. Reich (eds.) Inherently Parallel Algorithms in Feasibility and Optimization and Their Applications, pp. 473–504. Elsevier (2001)

[143] Yamada, I.: Kougaku no Tameno Kansu Kaiseki (Functional Analysis for Engineering). Suurikougaku-Sha/Saiensu-Sha, Tokyo (2009)

[144] Yamada, I., Elbadraoui, J.: Minimum-variance pseudo-unbiased low-rank estimator for ill-conditioned inverse problems. In: Proc. IEEE ICASSP, III, pp. 325–328 (2006)

[145] Yamada, I., Ogura, N.: Hybrid steepest descent method for variational inequality problem over the fixed point set of certain quasi-nonexpansive mappings. Numer. Funct. Anal. Optim. **25**, 619–655 (2004)

[146] Yamada, I., Ogura, N., Shirakawa, N.: A numerically robust hybrid steepest descent method for the convexly constrained generalized inverse problems. In: Z. Nashed, O. Scherzer (eds.) Inverse Problems, Image Analysis, and Medical Imaging, *Contemporary Mathematics*, vol. 313, pp. 269–305. AMS (2002)

[147] Yamada, I., Ogura, N., Yamashita, Y., Sakaniwa, K.: An extension of optimal fixed point theorem for nonexpansive operator and its application to set theoretic signal estimation. Technical Report of IEICE, DSP96-106, pp. 63–70 (1996)

[148] Yamada, I., Ogura, N., Yamashita, Y., Sakaniwa, K.: Quadratic optimization of fixed points of nonexpansive mappings in Hilbert space. Numer. Funct. Anal. Optim. **19**, 165–190 (1998)

[149] Yamada, I., Yukawa, M., Yamagishi, M.: Minimizing the Moreau envelope of nonsmooth convex functions over the fixed point set of certain quasi-nonexpansive mappings. In: H.H. Bauschke, R. Burachik, P. Combettes, V. Elser, D. Luke, H. Wolkowicz (eds.) Fixed-Point Algorithms for Inverse Problems in Science and Engineering, pp. 345–390. Springer (2011)

[150] Yamagishi, M., Yamada, I.: Nonexpansiveness of a linearized augmented Lagrangian operator for hierarchical convex optimization. Inverse Probl. **33**, 044003 (2017)

[151] Yang, J., Yuan, X.: Linearized augmented Lagrangian and alternating direction methods for nuclear norm minimization. Math. Comp. **82**, 301–329 (2013)

[152] Zălinescu, C.: Convex Analysis in General Vector Spaces. World Scientific (2002)

[153] Zeidler, E.: Nonlinear Functional Analysis and its Applications, III - Variational Methods and Optimization. Springer (1985)